# HOW CAN ETHICS MAKE BETTER AI PRODUCTS?

WHITEPAPER | APRIL 2020

**Jonathan Rotner**



MITRE | SOLVING PROBLEMS FOR A SAFER WORLD™

## Executive Summary

The growth of artificial intelligence (AI) technology has rapidly expanded the types of goods and services available, from personal convenience to professional assistance to defensive and security capabilities. Under bounded circumstances, AI is tremendously powerful at analyzing patterns, working through large amounts of data, and quickly responding to inputs. Yet as AI continues to permeate and integrate into users' lives, it can lead to significant ethical concerns. Those concerns can range from people finding creepy the automated email replies that mimic individual personalities,[1] to alarm over the national security implications caused by deepfakes,[2] decrying the mishandling of private data that drives AI platforms,[3] fear of losing jobs to AI,[4] protest over how mass surveillance so significantly impacts minority groups,[5,6,7] and fear of losing one's life to AI.[8] These examples show that there continue to be fielded systems that result in real harm, despite opportunity for employing better practices and lessons learned.

Speed is one significant motivator for maintaining the status quo: speed for companies that develop AI to be the first to market, and speed for United States (US) national security representatives to win an AI arms race against China. These market and international geopolitical pressures promote quick solutions, and as a result, complex problems are reduced to purely technical approaches, and products are deployed without adequate evaluation and oversight. When moving faster, there is less time to test, understand, and act on risk and impact assessments, security and privacy concerns, and opportunities for properly calibrating trust in the AI system. Including all of these elements would produce more responsible and ethical products. Put simply, decisions to act quickly or to act responsibly live in tension.

An essential element of any solution is to demonstrate that ethical AI products are better AI products. Public and private policies can shape AI's development and deployment to result in ethical AI that simultaneously boosts economic and national security outcomes. Then, the US can take advantage of existing, international demand for superior AI products. Therefore, the recommendations in this essay are directed toward two groups: the organizations that develop and deploy AI, and policymakers that can enact change.

If the US enacts these practical and impactful steps, new AI products and governance can reflect the socio-technical complexity of the problems they are trying to solve, and work to empower those using and affected by the AI. New AI products and governance can respond to a growing consumer base increasingly aware of the drawbacks of unfettered AI. New AI approaches can expand the AI workforce and contribute to a stronger economy (which also bolsters domestic security). And new AI approaches can help the US maintain international leadership and security by establishing norms that favor the promotion of Western, democratic principles.

## Modern AI Has an Ethics Problem

The spread of modern AI technology was fueled by Silicon Valley and startup companies' belief in the beneficial disruptive power behind their products. Uber and Lyft allow for consumer choice, access to transportation in remote areas, and flexible work hours for drivers; Amazon delivers packages to customers' doors at lower prices and more quickly than competitors; Google and Facebook connect people all over the world with similar interests. Add in artificial intelligence, and the efficiency and reach of these products multiply. AI technologies can even surpass some human abilities in very specific conditions – AI can recognize common images and objects better than human beings, AI can sift through large amounts of data faster than human beings, and AI can master more languages than human beings.[9]

But these capabilities do not exist in a technological vacuum, and therefore outcomes should not be measured purely by accuracy and efficiency. These technologies have social, cultural, privacy, civil liberties, cybersecurity, national security, and ethical impacts. In the past decade, adoption of AI-enabled systems has reached the domains of healthcare, justice, education, and finance. As the trust in and acceptance of these AI-enabled systems grew, and as the stakes of the decisions from these technologies increased, the world started to notice that these systems sometimes led to real harm, like the microtargeting of individuals with falsified information to sway their election choices under false pretenses,[10] or incorrectly deporting[11] or firing individuals[12] because humans were removed from decisions, or mass surveillance leading to imprisonment and suppression of populations,[13,14,15] or self-driving cars causing deaths.[16] Over the past several years, members of academia, industry, and government have increasingly asked how these systems work and have demanded greater accountability.

This discussion of accountability extends beyond AI; after all, not all of the above applications are fully AI-dependent. But these applications and their parent companies are increasingly using AI and more advanced forms of automation. The problems present in previous generations of automated technology are now exacerbated by the scope and scale of AI.[17,18] How?

The problems present in previous generations of automated technology are now exacerbated by the scope and scale of AI

1. ***An AI system replicates the social values of its developers and also embeds them into systems. This can encode those values as the new standard.*** AI developers choose what data is used in a training set, set the model's objectives, and make assumptions about how users and the environment will interact with the AI. The values and assumptions behind these choices are (intentionally or not) encoded into the AI. All too often those values represent how young, white, technically-oriented, Western men interact with the world, which does not reflect the full spectrum of priorities and considerations of all who interact with the AI.
2. ***An AI system's reach can centralize power in the hands of a few*.** If one person makes a decision or influences another single person's behavior, the effects are limited. But AI systems allow a single individual to accumulate and amplify their influence over many people's behaviors.
3. ***People can be influenced to trust an AI system more than they should.*** Under certain conditions, people place more trust in an AI system than is warranted, by assuming it is more impartial and infallible than they are. Individuals also have cognitive biases that lead them to treat connections and correlations as conclusions. Because AI systems can connect exponentially more information than a small group can on its own, AI systems can magnify false or misleading conclusions, making them seem like facts.

4. ***It is unclear who is accountable for an AI system's decisions.*** As of today, legal responsibility for the consequences of an AI system has not been established. With no one liable if something goes wrong, and no one made responsible for fixing it, the consequences of mistakes and misuse can easily lead to abuse of privacy and civil rights.

This essay looks to address these challenges. The essay first defines AI, then explores the market and global pressures that have shaped current AI practices, examines the issues behind the status quo, and concludes by recommending public and private policies that can improve future outcomes.

## A General Interpretation of Narrow AI

AI has been around for 60 years,[19] but experts and laypeople characterize AI a little differently.

Everyone agrees AI is everywhere. It fills in the text of internet searches, it customizes social media news feeds, it recommends products to buy or movies to stream, it powers voice recognition on phones, it does some of the flying during air travel, and it verifies credit when people apply for loans. Each of these examples represents AI that has been built to perform specific, bounded tasks. An AI that recommends a movie for the greater public will not work as well if it includes experimental short films made by drama students; an AI that is trained to recognize American voices will have trouble with Scottish accents. These limitations lead some experts to refer to modern AI as "Artificial Narrow Intelligence" (ANI).

Artificial General Intelligence (AGI), on the other hand, is closer to science fiction. These systems can think and act like humans, are almost fully self-reliant, and can handle environments and problems they haven't faced before. A layperson might think of Rosie in *The Jetsons*, HAL in *2001: A Space Odyssey*, or KITT in *Knight Rider*. Abstract thinking, a skill only humans have today, would only be possible with AGI.

Where does really advanced modern technologies, such as self-driving cars, fit in? They represent attempts to expand "narrow" AI. When an environment changes or the clarity of the task is muddied, it gets harder to develop robust and dependable AI. Self-driving cars use lots of sensors, computational power, hours on the road, and simulated scenarios to "bound" different possibilities into situations that are more recognizable – they try to make the unknown a little more familiar, rather than reason abstractly like AGI.[20]

This essay focuses on ANI, which will be referred to here as "AI." Some of the lessons learned could apply to simple automation (following basic rules) as well as very advanced learning systems.

## Existing Commercial, Ethical AI Approaches Are a Good Step, But Fall Short of Meaningful Change

Public and private entities have developed and deployed AI with good intentions. But more and more examples are showing that despite meaning well, current methods of deploying automated technologies are inducing significant harm. In 2018 and 2019, journalists, academics, and individuals within AI-developing organizations called for change. AI developers and deployers are aware of the negative consequences to operationalized algorithms. But right now, market forces reward "moving fast and breaking things."[21] Companies want to maintain market advantage by keeping models proprietary and by releasing products more quickly than competitors do. These pressures push evaluation

and assessment to a point in time after the product is released in the wild (if at all), which tests the unforeseen impacts on potential unwilling participants. Market dominance and profits are penalizing ethics and responsibility.[22,23,24]

In response to calls for change, organizations that develop and deploy AI have embraced three kinds of ethical approaches:

1) Declarations of ethical, responsible practices
2) Creation of ethical frameworks that promote responsible AI development
3) Introduction of toolkits that assess and visualize statistical biases in datasets

Over 80 institutions[25] have published their own technology and AI mission statements and ethical guidelines (the best and often-cited ones, in the author's opinion, are in the endnotes[26,27,28,29]). These statements vary in detail and specificity, but almost all declare principles of transparency, non-discrimination, accountability, and safety. Some declarations include that AI and technology must aid and benefit society and protect human rights.

In addition to declarations of responsible practices, many universities, companies, and international government organizations have created ethical frameworks to help developers and deployers consider more responsible AI options (the best and often-cited ones, in the author's opinion, are in the endnotes;[30,31,32,33,34,35] each resource has a complementary perspective). Some toolkits are akin to risk management and assessment frameworks, others are questions designed to inspire critical thinking, and others are checklists meant to encourage data scientists and program managers to think outside of their domains and consider legal, social, cybersecurity, and humanistic implications to their work.

Finally, AI corporations, legislatures, and the public are becoming more aware, even more than a few years ago, of how datasets can affect the accuracy and "bias" of an AI's results. Homogenous datasets cannot be extended to more diverse environments and inputs or, to give an example, facial recognition algorithms that train on only white individuals fare poorly when used by darker-skinned individuals.[36,37] To combat overly narrow datasets, mathematicians and programmers created tools that deploy and visualize traditional, statistical sampling methods.[38] Specific toolkits have also been developed for datasets that contain associations that are human-influenced[39,40,41,42,43] (for example, the term "female" is more closely associated with "homemaker" than with "computer programmer" in a search of Google News articles). Finally, researchers have developed model-intervention methods that impose constraints in order to limit how extreme or homogenous a machine learning algorithm's output can become.[44,45,46]

> The challenge with ethical declarations and frameworks is that they are almost universally voluntary commitments

These current approaches are important steps, but evidence shows that they are not enough. The challenge with ethical declarations and frameworks (approaches 1 and 2) is that they are almost universally voluntary commitments. Few have recommendations, specifics, or use cases for how to make the principles actionable and implementable,[47] and pledges to uphold ethical principles do not guarantee ethical behavior.[48]

The challenge with toolkits that examine statistical interpretations of bias (approach 3) is that they are oversimplified approaches to complex, not strictly mathematical problems. Bias can be statistical (like polling only young people instead of multiple generations) as well as human-influenced (like stereotyping, craving certainty, confirmation bias, and many other types of bias[49]). Bias is nuanced and can enter at many different stages of machine learning development, so it is hard to detect.[50,51,52] When it is detected, solutions are limited – mathematical solutions do not

fix existing systemic inequalities or human-influenced bias that might be patterned in the data. And if these biases aren't caught, then they are both encoded into the algorithm and amplified as the reach of the algorithm scales.[53,54,55] Finally, the allure of a purely technical, seemingly objective solution takes resources and attention away from the educational and sociopolitical approaches that are necessary to address the root causes of issues.[56,57,58]

For examples of each of these challenges, see the story-box (for story-box sources, see notes in the paragraph above).

| EXAMPLES OF OVERSIMPLIFIED APPROACHES TO COMPLEX, SOCIO-TECHNICAL PROBLEMS |
|---|
| ***Statistical and Human-influenced Forms of Bias*** |
| Statistical bias can result from unrepresentative training sets, like in the case of webcam algorithms that could not track faces of darker-skinned individuals because all the training data (and most of the developers) were of white-skinned individuals.<br><br>Human-influenced bias can result from relying on data and processes that historically have their own undesired outcomes. Search queries for "beautiful" and "ugly" women more often associate black, Asian, and older women with images of unattractiveness, while photos of young, white women appear more frequently as examples of beauty. |
| ***Encoded and Amplified Inequalities*** |
| Algorithms embed social values into systems, when human developers choose what data and goals go into the system. Those outcomes, even if they are grounded in relevant data, don't always produce wanted outcomes. In the case of Amazon using an AI product to hire top talent, because the AI was trained on previous hires' background, it preferred male candidates to female ones and actively penalized resumes that had words more closely associated with women.<br><br>Amplification of inequalities occurs because an AI application centralizes power in the hands of the few to affect the lives of many. For example, YouTube's algorithms are designed to engage an audience for as long as possible; consequently, the recommendation engine pushes videos with more and more extreme content, since that's what keeps users' attention. |
| ***Spending Money on Technical Solutions at the Expense of Educational and Sociopolitical Ones*** |
| As new technology is incorporated for internal use, training and education does not always accompany it. For example, after the Boeing 737 MAX aircraft crashes, pilots were furious that they had not been told the aircraft had new software and that it was omitted from the manual. Reports showed that "all the risk [is put] on the pilot, who would be expected to know what to do within seconds if a system he didn't know existed … forced the plane downward."<br><br>And as new technology displaces people's jobs and becomes more deeply integrated into people's lives, the concerns of those individuals may be overlooked, and sociopolitical factors may not be addressed, for the sake of technical progress. For example, when Waymo began testing self-driving cars in Arizona, local citizens aired their frustrations over not being consulted and their fears of losing jobs by slashing tires and throwing rocks at the Waymo vehicles. |

Skeptics might see declarations, frameworks, and toolkits as virtue signaling,[59] resulting in words without action. But pragmatists can try to hold an organization accountable to its own pledges and commitments.[60] And optimists can envision that companies that follow through on their ethical pledges will create better products that more accurately reflect the complexities of the challenges they are trying to solve.

Companies that follow through on their ethical pledges will create better products that more accurately reflect the complexities of the challenges they are trying to solve

One solution to such conflicts between principle and practice may be to use those same market pressures to demonstrate that ethically designed products lead to *better overall solutions*. But before turning to recommendations, the next section discusses another force that encourages the same type of speedy and risk-discounting outcomes – threats to national security.

## The AI Arms Race Is Pushing the Government to Adopt and Deploy AI Before the Government Is Ready

For the National Security Sector (NSS) of the US government, global pressures do not take the form of profits – they are about an AI "arms race." Part of China's appetite to rival and surpass the US as a world power is to establish itself as the global leader in AI, and China is spending significant investment to do so.[61] In order to keep up with China, the US NSS feels compelled to pursue AI, perhaps with faster-than-desired timetables. The risks inherent from an accelerated deployment schedule could be mitigated by incorporating decades' worth of legal, ethical, and accountability structures into a new technology. However, the NSS is creating AI-specific governance structures while simultaneously learning about, acquiring, and fielding AI, all without a unified strategy. Adding to that challenge, the NSS feels it must act quickly in order to maintain its desire to be a leader in international legal and humanitarian standards. The prize in the AI arms race is military dominance and global, technical influence, and the US is participating as much to win as it is out of fear of losing to China.[62]

Being the "best" at AI brings about an enormous strategic advantage.[63,64] This essay defines *military dominance* as AI improvements to military capabilities: whoever has the best or fastest AI-enabled military can deter or harm their adversaries. AI-enabled military dominance can enable a nation to:

- Conduct ongoing and dynamic cyber attacks;[65]
- Operate inexpensive drones, satellites, and other sensors (especially if combined with AI) that increase the ability to surveil and reconnoiter an adversary; and
- Instill fears of an AI-enabled quick-strike, which erode the barriers working against preemptive action.[66]

These are among the factors that led a Department of Defense (DoD) and Joint Chiefs of Staff strategic assessment to say, "We must not be caught by surprise."[67]

A complementary outcome to AI dominance is *global, technical influence,* which in this essay refers to AI norms: whoever has contributed the most prevalent or relied upon AI is in the strongest position to shape international standards, sectoral best practices, and expectations of use.

In a feedback loop of enormous significance, AI-enabled global, technical influence can significantly affect international governance and standards. As the nonpartisan think tank New America writes, "part of the [Chinese]

plan's approach is to devote considerable effort to writing guidelines not only for key technologies and interoperability, but also for the ethical and security issues that arise across an AI-enabled ecosystem, from algorithmic transparency to liability, bias, and privacy." New America believes that the Chinese government places a lot of importance in being an international AI leader, "both for economic reasons and because of the national prestige."[68]

The Chinese have demonstrated their commitment to AI – sources estimate their government will spend $70 billion on AI in 2020,[69] which is anywhere from fourteen[70] to seventy times[71] more than what the US government is estimated to spend. (It is important to note that one recent report concluded that the Chinese figures are significant overestimates, and that the amount is closer to that of the US.[72] It is also important to note that the US calculation does not take into account investments in AI by the US private sector.) The Chinese are also investing in Silicon Valley startups, so they can influence AI development both from within and beyond their borders.[73] In addition to development, the Chinese are affecting AI use through international partnership. The Chinese have exported their facial recognition capabilities (and norms) to authoritarian Latin American and African governments, and China is shaping regulatory standards for facial recognition use in the United Nations' International Telecommunications Union.[74]

> The US cannot allow China to set the terms and tempo of an AI arms race. Instead, the US must create a strategy that plays to its own strengths.

On their own or through partners, domestically and internationally, the Chinese have a comprehensive plan to shape AI norms according to their own objectives.[75] And journalists have uncovered what those objectives entail: trading technology for the personal data (and faces) of citizens in foreign countries,[76] tracking the activities of government protesters,[77] controlling domestic populations by assigning credit scores,[78] and subjecting the Uighur Muslim population to torture and death.[79]

Therefore, for their own benefit and for their goal of spreading democracy, the US NSS must invest heavily and speedily in AI. The NSS believes that America's long history as a democracy operating under the rule of law, governing authorities, and a code of ethics to guide its use of powerful technologies can reduce some of the risks of deploying a technology too quickly.[80]

With conventional weapons, military commanders work side by side with legal advisors to either approve their use when they comply with international humanitarian law or advise against their use when it could result in a violation of that law.[81] The Defense Innovation Board (DIB), an independent federal advisory committee that provides advice and recommendations to DoD senior leaders, lays out the DoD's history and code of conduct, and how AI might be interpreted under existing law:[82]

> Evidence for [how the DoD makes and executes decisions] is reflected through various statements, policy documents, and existing legal obligations. Formal accords include the Law of War and existing international treaties, while numerous DoD-wide memoranda from Secretaries of Defense highlight the importance of ethical behavior across the armed services. In isolation and taken together, this body of evidence shows that DoD's ethical framework reflects the values and principles of the American people and the U.S. Constitution. ...
>
> Existing Law of War rules can apply when new technologies are used in armed conflict. … The fundamental principles of the Law of War provide a general guide for conduct during war, where no more specific rule applies, and thus provide a framework to consider novel legal and ethical issues posed by emerging

> technologies, like AI. For example, if AI was added to weapons, such weapons would be reviewed to ensure consistency with existing legal requirements, such as the requirement that the weapon not be calculated to cause unnecessary suffering or be inherently indiscriminate. Additionally, under the Law of War, commanders and other decision-makers must make decisions in good faith and based on the information available to them and the circumstances ruling them at the time. The use of AI to support command decision-making is consistent with Law of War obligations, including the duty to take feasible precautions to reduce the risk of harm to the civilian population and other protected persons and objects.[83]

In addition, the Intelligence Community (IC) has its own set of governing laws and policy, authorities, and codes of conduct (called the Augmenting Intelligence Using Machines, or AIM initiative). Its strategy document outlines: "how the IC will incorporate AIM capabilities in a manner that resolves key IC legal, policy, cultural, technical, and structural challenges while producing optimally effective analytic and operational contributions to the intelligence mission. ... The AIM initiative is about much more than technology. Implementing the strategy will entail addressing workforce challenges and understanding and shaping the policies and authorities governing how the IC deploys and uses AI."[84]

The introduction of AI technology will be no different than other technologies when the NSS considers its legal and ethical accountabilities. What is different, however, is that the government still has to figure out how and if this specific technology – because it is probabilistic and therefore more unpredictable, because it can act too quickly for operators to understand the potential consequences of its decisions, and because users don't always understand why a decision is made – will create new governance practices and norms.

The US National Security Sector is faced with a big challenge. On the one hand, international forces are pressuring the US government to deploy AI more quickly and accept more risks than it might prefer, and speed engenders unknown consequences, leading to potential harm. On the other hand, winning the AI race is the best way for the US to shape international norms and restrictions on the use and the export of AI-enabled technologies, rather than allowing legal and moral terms of AI use to be decided by despotic regimes.[85]

The US cannot allow China to set the terms and tempo of an AI arms race. Instead, the US must create a strategy that plays to its own strengths: by taking advantage of capitalist, global market forces and the increasing demand for these technologies, while at the same time reflecting the values of democracy. Remarkably, that path can lead to the most ethical outcomes as well.

## Ethically Designed AI Leads to Better Overall Solutions

The United States can significantly increase its AI market share and global influence if the technology that the US presents is simply better than the alternatives. Chinese models boast high degrees of accuracy,[86] but users and purchasers of AI systems need to move away from measuring value only in terms of accuracy and efficiency. If American products successfully reflect the complexity of the problems AI is trying to solve, they will benefit and empower customers and those affected by the technology, in a manner that establishes them as preferred solutions.

The US can do this by realizing that big challenges are multidimensional, nuanced, and encompassing. When the human and technical sides to AI approaches reinforce each other, and do not interfere with each other, this can:

- Save time and resources in product development – sometimes a fully automated solution is harder to implement and less efficient than a human-automated partnership[87]
- Better identify when bias, reward hacking (a clever cheat that goes against the spirit of the challenge), or adversarial approaches (fooling or spoofing an AI) lead to unwanted or inaccurate outcomes
- Treat the communities affected by the AI as customers, in order to foster conditions that lead to better acceptance and adoption of the technology[88]

A multidimensional approach not only will distinguish American from Chinese approaches but will lead to more welcomed, sustained, and ethical AI products.

Different companies will decide for themselves whether this economic argument is germane. But the government also has a role to play: it has historically enacted regulations and limitations on industry behavior to promote overall public health and mobility,[89] even when doing so is initially detrimental to industry profit and growth (one example is increased fuel efficiency standards for cars). Most likely, regulation will induce more expense at first,[90] but if the government sets national standards, the most innovative companies can benefit and attract new customers.[91]

There are steps the US can take that are practical, have impact, and increase accountability. These recommendations are focused on two groups: the organizations that develop and deploy AI, and policymakers that can enact change. Recommendations are listed in order of increasing levels of complexity and effort.

### 1) Empower the AI user with understanding and choice

There is no such thing as “neutral” AI. Developers make conscious and unconscious assumptions about the goals and priorities of the AI, and the important factors that the AI learns from. Often the AI and the users’ incentives are aligned, so this works out fairly well: users drive to their desired destination; users enjoy the AI-recommended movie. But when the goals don’t align (again, intentionally or unintentionally), an uneducated public or an untrained user is not made aware that they are potentially acting against their interests. Users can perceive that they are making rational and objective decisions, given the implied authority and objectivity of the AI.[92]

For example, Facebook’s newsfeed algorithm is designed to engage an audience for as long as possible, thus the recommendation engine suggests articles with more extreme content, because that is what keeps people’s attention, even if they are looking for diverse or opposing perspectives.[93,94]

There are many socio-cultural approaches to improve these outcomes, which will be visited in the subsequent recommendations. But there are technical and procedural changes to products that can empower the user with

understanding and choice. Full explainability and transparency is really hard.[95] Instead, *there are ways to create more awareness and control for users in order to help them more accurately calibrate trust in the AI-enabled system*. (These steps are not possible for all types of AI, but a full conversation of the tradeoffs and challenges is beyond the scope of this essay.)

Take the AI out of the system for a moment, and think about agreeing on a definition for a word, like "fairness." In the simple example of a bank granting a loan, is it fair for men and women to get the same number of approvals overall (100 of 200 loans accepted for men, and 100 of 150 accepted for women) or at the same percentage rate (100 of 200 accepted for men, and 75 of 150 accepted for women)? Because a single algorithm cannot do both.[96] So what are some options for developers and users to consider?

A powerful idea, when possible to implement, would be to add a dial that the user can employ to switch between different algorithm objectives[97] – for example, different versions of fairness. When the dial is accompanied by an explanation of what the algorithm is optimizing, the user, instead of the developer, can decide which outcome is appropriate for which situation. Another approach is to try to overcome some of the trust challenges when working with a "black box" system (a black box system does not allow users to see the inside of an algorithm and understand how it arrives at a decision) by including more information about how the model makes decisions, and about the developers' choices in the design process. The algorithm can provide text and visual examples of what training data was most helpful and most misleading for arriving at the correct solution (for example, "this tumor is classified as malignant because to the model it looks most like these other tumors, and it looks least like these benign conditions"). Developers can include confidence scores and descriptions of how these scores are generated. Especially helpful for policymakers, documentation about data[98] and models[99] can include where the AI developers intended or did not intend the AI to be used.[100]

Conveying design choices can be fundamentally transformative to the user's assessment of model appropriateness and trustworthiness. If models offer more information about how human and model decisions were made, adopters can have fewer surprises and can more accurately weigh the risks of integrating the technology into their processes. In addition, when algorithms are designed to offer evidence and counterevidence, they can elicit more diverse ideas and open dialogue – principles that are foundational to the health of democracies.[101]

### 2) Provide protection and resources for those who advocate for more responsible AI outcomes

From employee walkouts[102] and advocacy[103] to community engagement,[104] people are fighting back against overbearing and overwhelming AI deployment. When AI-enabled systems are rolled out without mechanisms for accountability, governance, and a way to contest decisions, the result is "black box organizations." And these organizations are encountering resistance.

The individuals who work at technology companies are coming forward to air their concerns. They are representing and responding to the ethical declarations that their organizations proclaim. In the government, there is whistleblower protection, but all AI organizations need to protect workers' rights to not only whistle-blow, but share ethical concerns with management and maintain the ability to work on applications they deem ethical, all without repercussion.[105]

Ethical accountability lies with the organizations that develop and deploy AI. *Because it is the data science and engineering companies that are approached to apply seemingly objective and relatively speedier fixes to nuanced, ingrained, and expensive problems, it is their responsibility to bring in other voices.*[106] And because the organizations that acquire AI solutions know their own domains best and know the historic and systemic challenges that have

prevented easy solutions, it is their responsibility to ensure that the right stakeholders are included. One example illustrates an opportunity to learn from undesired consequences: when a hospital purchased an algorithm that weighed healthcare costs against where care could have the best outcomes, the result was significantly less access to care for black patients than for white patients. Data scientists trained the algorithm on existing data patterns, where less money is traditionally spent on black patients than on white patients with the same level of need. Therefore, the algorithm falsely concluded that black patients are healthier than equally sick white patients.[107] Including doctors, hospital administrators, nurses, and patients early in the development process, and giving them a vote in the design,[108] could have prevented this unfortunate outcome.

This anecdote illustrates a broader truth: more resources and better protection are also needed for the communities that are affected by AI-enabled systems. Too often, the populations with the least means are most affected by AI-enabled systems, whether through ads for housing that perpetuate discrimination,[109] being targeted by mass surveillance,[110] being subject to judicial oversight,[111] or getting relatively low access to healthcare.[112] As the AI Now Institute at New York University (a research institute dedicated to understanding the social implications of AI technologies) puts it, "More funding and support are needed for litigation, labor organizing, and community participation on AI accountability issues. … This includes supporting public advocates who represent those cut off from social services due to algorithmic decision making, civil society organizations and labor organizers that support groups that are at risk of job loss and exploitation, and community-based infrastructures that enable public participation."[113] This support also includes assigning responsible parties and processes to administer changes at the deploying organization, and making clear how those affected by the AI can alert those parties.[114,115]

If ethical outcomes are part of an organization's values, it needs to devote resources and establish accountability to ensure those values are upheld.

### 3) Require objective, third-party verification and validation

*Because algorithms are making decisions that affect the livelihoods, finances, health, and the civil liberties of entire communities, the government has to protect the public, even if doing so may be initially detrimental to industry profit and growth*. By incentivizing participation, the government could offset initial increased costs for AI in order to help promote the emergence of a new marketplace that responds to a demand signal for ethical AI.

Objective, third-party verification and validation (O3VV) would allow independent parties to scrutinize an algorithm's outcomes, both technically and in ways that incorporate the social and historical norms established in the relevant domain. For meaningful oversight, O3VV needs to understand the entire lifecycle of the AI-enabled system: from evaluating the relevance of the training datasets, to analyzing the model's goals and how it measures success, to documenting the intended and unintended deployment environments, to considering how other people and algorithms use and depend on the system after each update.[116]

Think of O3VV like an Energy Star seal – the voluntary program established by the Environmental Protection Agency that allows consumers to choose products that prioritize energy efficiency.[117] Or think of "green energy" companies that respond to consumer preference for sustainable businesses and products, and enjoy more profits at the same time.[118] Both models center on a recognized, consensual set of criteria, as well as an (ideally, independent) evaluative body that confirms compliance with the standard.

Following these examples, O3VV should reflect the public sentiment asking for change. Evaluators should come from multiple academic backgrounds and represent all the communities affected by the AI. O3VVs could take on consumer

protection roles, placing emphasis on how the decisions affect real people's lives. [119,120] O3VV agencies could take the form of a government auditing program, Federally Funded Research and Development Centers (FFRDCs), certified private companies, and a consensually developed "seal" program.

In order for O3VV to become established practice, the government needs to incentivize participation. Currently, there are no standards for using AI that have been certified by O3VV, nor are there incentives for companies to go through a certification process, or for professionals and academics to contribute to the process.[121] One approach calls for a licensing program for O3VV professionals, and another calls for increasing monetary incentives for deploying certified systems.[122] Another idea is to allow FFRDCs, which by law are not allowed to compete with industry and which work only in the public interest, access to proprietary AI datasets and model information in order to perform independent verification and validation. Especially if the government is a consumer, it can require that vendors adhere to these steps before the government will purchase their products.[123,124]

### 4) Entrust sector-specific agencies to establish ethical AI standards for their domains

New technologies will more broadly adopted if they follow established practices, expectations, and authorities in a given domain. The following two examples can illustrate how.

First, a children's hospital in Philadelphia deployed a black box AI that looks for a rare but serious infection (sepsis). The AI used patients' electronic health records and vital-sign readings to predict which fevers could lead to an infection. The AI identified significantly more life-threatening cases than did doctors alone (albeit with many false alarms), but what made the story so compelling and the application so successful was that doctors could examine the identified patients as well as initiate their own assessments without alerts from the AI. Doctors could use the AI's queues while still employing their own judgment, decision making, and authority, to achieve improved outcomes.[125,126]

Second, state and local jurisdictions in the US have deployed COMPAS, a black box risk-assessment tool that assesses inmate recidivism (repeating or returning to criminal behavior). COMPAS uses a combination of personal and demographic factors to predict the likelihood an inmate would commit another crime. COMPAS produced controversial results: the number of white inmates with a certain score re-offended at the same rates as black inmates with that score, but among defendants who did not re-offend, black inmates were twice as likely as white inmates to be classified as medium or high risk. As in the hospital example, judges could ignore COMPAS's input or refer to it, but final assessment and responsibility lay with the judge.[127,128,129]

In each of these cases, the expert could discount or act on the AI's recommendation. But the difference between these two examples lies in the historical and cultural norms, rules, and expectations that exist in the two domains. The public might be less at ease with using AI in the judicial context for any number of domain-specific reasons: because judges rule in "case of first impression" when a higher court has not ruled on a similar case before,[130] or because the court uses twelve jurors rather than a single judge, a practice established as representative of a good cross-section of perspectives.[131] In contrast, the public might be more at ease with AI offering predictions on medical diagnoses because doctors routinely use "evidence-based medicine"[132] to integrate their own clinical experience with the latest research, established guidelines, and other clinicians' perspectives, of which the algorithm could be considered a part. Doctors also take the Hippocratic oath, pledging to work for the benefit of the sick,[133] whereas judges must weigh both individual and collective good in their decisions.

In short, different sectors have different expectations; therefore, institutional expertise should be central to determining the benefits and risks of incorporating each type of AI system.

*Sector-specific agencies already have the historical and legislative perspectives needed to understand how technology affects the domain under their responsibility; now, each of those agencies should be empowered to expand its oversight and auditing powers to a new technology***.** The White House recently called for the same process in its draft principles for guiding federal regulatory and non-regulatory approaches to AI: "Sector-specific policy guidance or frameworks. Agencies should consider using any existing statutory authority to issue non-regulatory policy statements, guidance, or testing and deployment frameworks, as a means of encouraging AI innovation in that sector."[134] It is incumbent on individual agencies to permit, regulate, temper, and even ban[135] AI-enabled systems as determined by the experts and established practices in each domain.

At the core of tech tragedies[136,137,138] related to AI in 2018 and 2019 are questions of accountability: who is responsible when AI systems harm someone? Where are the points of intervention, and what additional research and regulation is needed to ensure those interventions are effective? Currently there are few answers to these questions. Government legislation on AI ethical standards means enacting a legal framework that ensures that AI-powered technologies are well researched, the AI's impacts are tested and understood, and the AI is developed with the goal of helping humanity.[139] It is possible to develop government legislation that demands chains of accountability while allowing industry the flexibility to enact accountability structures and enforcement mechanisms that work for that organization.

### 5) Expand and integrate the US AI talent pool

The AI workforce gap is often cited as the largest barrier to AI adoption:[140] the demand for talent is not met by the number of qualified workers.[141] At the same time, AI is increasingly being deployed in domains outside of traditional science, technology, engineering, and mathematics (STEM) fields, where users are not as familiar with the details of and limitations to the technology. Expanding AI education and training to non-STEM fields[142] and providing opportunities for STEM and non-STEM individuals to practice AI together could not only reduce the AI workforce gap, but also create products that better reflect a diverse user base.

Organizations that plan to employ AI have an opportunity to improve the products they develop and deploy by including individuals from non-STEM fields. In the status quo, STEM individuals are not normally trained to research social perspectives. For example, in an experiment designed to "engage and entertain," the chatbot Tay was released into the wild. Designed to learn from the communication patterns of 18- to 24-year-olds, it instead took on sexist, anti-Semitic, racist, and other inflammatory statements.[143] Designers and deployers of the algorithm are not and will never be fully trained ethicists and domain experts. But because they did not include those who better understood the relevant social realities, they unintentionally demonstrated the need for multidisciplinary perspectives. Multidisciplinary teams better represent the values of a diverse consumer base, leading to better aligned products.[144] Multidisciplinary teams allow for more innovation and creativity, resulting in more profitable products.[145,146] And multidisciplinary teams can better prevent, moderate, and recover from unintended consequences.

Therefore, the US should foster more multidisciplinary approaches in academic education and professional training. Cross-discipline pollination can provide opportunities for non-STEM individuals to be exposed to and become more proficient in using AI, and for STEM individuals to see the importance of vigilance and caution when deciding whether AI is an appropriate approach for a task. At minimum, exposing more people to AI will increase the sheer number of

qualified workers. But more impactfully, *including professionals outside the existing AI workforce allows ideas and expertise to check the growing imbalance of power and accountability between those who are victims of a technology and those who embed their authorities and preferences through the development and deployment of the technology.*[147] And thinking most long-term, acting on the prediction that most future jobs will require knowledge of AI[148] allows the US to bolster opportunities for all future members of the workforce through AI education.

Crossing educational borders and expanding familiarity with AI will only not only lead to better quality products, but simultaneously strengthen the US' economic position.

## Going Forward

Enacting these recommendations is an uphill battle, but it's also an opportunity. Voices in opposition to the AI status quo are gathering, indicating there is support for change. That change will be smoothest if solutions take advantage of the very same market forces and international, geopolitical pressures facing the US today. Through public and private policy change, the US can indeed show that ethical AI products are better AI products, and simultaneously create new markets and demonstrate moral leadership.

Taking action would create a tremendous amount of credibility and domestic stability. By pursuing ethical AI outcomes, the US government and private industries can learn from previous AI development and deployment mistakes; protect and empower AI employees, users, and affected communities; foster a stronger and more comprehensive AI workforce; and shape how international partners inform their own AI processes and deployments, all while pursuing better outcomes for the problems at hand.

The time to act is now.

*Jonathan Rotner is a human-centered technologist who helps program managers, algorithm developers, and operators appreciate technology's impact on human behavior. He works to increase communication and trust when working with automated processes.*

*A big thank you to Lisa Bembenick, Eric Bloedorn, and Richard Games for their support; Michael Aisenberg, Duane Blackburn, and Chuck Howell for their review; Sheila Gagen and Lidia Sabatini for their edits; and Lura Danley and Ron Hodge for their guidance.*

# References


[1] M. Anderson, "Useful or creepy? Machines suggest Gmail replies," *AP News*, Aug. 30, 2018. [Online]. Available: https://apnews.com/bcc384298fe944e89367e42e20d43f05

[2] "House Intelligence Committee hearing on 'Deepfake' videos," *C-SPAN*, June 13, 2019. [Online]. Available: https://www.c-span.org/video/?461679-1/house-intelligence-committee-hearing-deepfake-videos

[3] [Online]. Available: https://informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/. Best viewed in a browser other than Internet Explorer

[4] C. Forrest, "Fear of losing job to AI is the no. 1 cause of stress at work," *TechRepublic*, June 6, 2017. [Online]. Available: https://www.techrepublic.com/article/report-fear-of-losing-job-to-ai-is-the-no-1-cause-of-stress-at-work/

[5] S. Browne, *Dark Matters: On the Surveillance of Blackness*, Durham, NC, USA: Duke University Press Books, 2015.

[6] A. M. Bedoya, "The color of surveillance: What an infamous abuse of power teaches us about the modern spy era," *Slate*, Jan. 18, 2016. [Online]. Available: https://slate.com/technology/2016/01/what-the-fbis-surveillance-of-martin-luther-king-says-about-modern-spying.html

[7] M. Cyril, "Watching the Black body," Electronic Frontier Foundation, Feb. 28, 2019. [Online]. Available: https://www.eff.org/deeplinks/2019/02/watching-black-body

[8] P. McCausland, "Self-driving Uber car that hit and killed woman did not recognize that pedestrians jaywalk," *NBC News*, Nov. 9, 2019. [Online]. Available: https://www.nbcnews.com/tech/tech-news/self-driving-uber-car-hit-killed-woman-did-not-recognize-n1079281

[9] R. Steinberg, "6 areas where artificial neural networks outperform humans," *Venture Beat*, Dec. 8, 2017. [Online]. Available: https://venturebeat.com/2017/12/08/6-areas-where-artificial-neural-networks-outperform-humans/

[10] A. Chang, "The Facebook and Cambridge Analytica scandal, explained with a simple diagram," *Vox*, May 2, 2018. [Online]. Available: https://www.vox.com/policy-and-politics/2018/3/23/17151916/facebook-cambridge-analytica-trump-diagram

[11] N. Sonnad, "A flawed algorithm led the UK to deport thousands of students," *Quartz*, May 3, 2018. [Online]. Available: https://qz.com/1268231/a-toeic-test-led-the-uk-to-deport-thousands-of-students/

[12] C. Lecher, "How Amazon automatically tracks and fires warehouse workers for 'productivity,'" *The Verge*, Apr. 25, 2019. [Online]. Available: https://www.theverge.com/2019/4/25/18516004/amazon-warehouse-fulfillment-centers-productivity-firing-terminations

[13] P. Taddonio, "How China's government is using AI on its Uighur Muslim population," *Frontline*, Nov. 21, 2019. [Online]. Available: https://www.pbs.org/wgbh/frontline/article/how-chinas-government-is-using-ai-on-its-uighur-muslim-population/

[14] D. Z. Morris, "China will block travel for those with bad 'social credit,'" *Fortune*, March 18, 2018. [Online]. Available: https://fortune.com/2018/03/18/china-travel-ban-social-credit/

[15] R. Adams, "Hong Kong protesters are worried about facial recognition technology. But there are many other ways they're being watched," *BuzzFeed News*, Aug. 17, 2019. [Online]. Available: https://www.buzzfeednews.com/article/rosalindadams/hong-kong-protests-paranoia-facial-recognition-lasers

[16] S. Gibbs, "Tesla Model S cleared by auto safety regulator after fatal Autopilot crash," *The Guardian*, Jan. 20, 2017. [Online]. Available: https://www.theguardian.com/technology/2017/jan/20/tesla-model-s-cleared-auto-safety-regulator-after-fatal-autopilot-crash

[17] "Algorithms and artificial intelligence: CNIL's report on the ethical issues," CNIL [Commission Nationale de l'Informatique et des Libertés], May 25, 2018. [Online]. Available: https://www.cnil.fr/en/algorithms-and-artificial-intelligence-cnils-report-ethical-issues

[18] M. Whittaker et al., *AI Now Report 2018*. New York, NY, USA: AI Now Institute, 2018. [Online]. Available: https://ainowinstitute.org/AI_Now_2018_Report.pdf

[19] T. Lewis, "A brief history of artificial intelligence," *Live Science*, Dec. 4, 2014. [Online]. Available: https://www.livescience.com/49007-history-of-artificial-intelligence.html

[20] T. D. Jajal, "Distinguishing between narrow AI, general AI and super AI," Medium, May 21, 2018. [Online]. Available: https://medium.com/@tjajal/distinguishing-between-narrow-ai-general-ai-and-super-ai-a4bc44172e22

[21] "What does 'move fast and break things' really mean?" *Quora*. Accessed on: Jan. 17, 2020. [Online]. Available: https://www.quora.com/What-does-move-fast-and-break-things-really-mean

[22] M. Whittaker et al., *AI Now Report 2018*. New York, NY, USA: AI Now Institute, 2018. [Online]. Available: https://ainowinstitute.org/AI_Now_2018_Report.pdf
[23] T. Hagendorff, "The ethics of AI ethics: An evaluation of guidelines," *arXiv.org*, Oct. 11, 2019. [Online]. Available: https://arxiv.org/abs/1903.03425
[24] A. C. Uzialko, "Workplace automation is everywhere, and it's not just about robots," *Business News Daily*, Feb. 22, 2019. [Online]. Available: https://www.businessnewsdaily.com/9835-automation-tech-workforce.html
[25] "An ethics guidelines global inventory," Algorithm Watch. Accessed on: Jan. 17, 2020. [Online]. Available: https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/
[26] "Ethics in action: The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems," *IEEE.org*. Accessed on: Jan. 17, 2020. [Online]. Available: https://ethicsinaction.ieee.org/
[27] "Recommendation of the Council on Artificial Intelligence," OECD/LEGAL/0449, OECD, 2019. [Online]. Available: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449
[28] "Tenets," Partnership on AI. Accessed on: Jan. 17, 2020. [Online]. Available: https://www.partnershiponai.org/tenets/
[29] N. Diakopoulos, "Principles for accountable algorithms and a social impact statement for algorithms," Fairness, Accountability, and Transparency in Machine Learning. Accessed on: Jan. 17, 2020. [Online]. Available: https://www.fatml.org/resources/principles-for-accountable-algorithms
[30] "Ethics & Algorithms Toolkit," *ethicstoolkit.ai*. Accessed on: Jan. 17, 2010. [Online]. Available: http://ethicstoolkit.ai/
[31] "Artificial intelligence impact assessment—Netherlands," Platform for the Information Society, 2018. [Online]. Available: https://airecht.nl/blog/2018/ai-impact-assessment-netherlands
[32] M. Whittaker et al., *AI Now Report 2018*. New York, NY, USA: AI Now Institute, 2018. [Online]. Available: https://ainowinstitute.org/AI_Now_2018_Report.pdf
[33] DJ Patil, H. Mason, and M. Loukides, "Radar: Of oaths and checklists," O'Reilly Media, July 17, 2018. [Online]. Available: https://www.oreilly.com/radar/of-oaths-and-checklists/
[34] "AI & ethics: Collaborative activities for designers," *IDEO.com*. Accessed on: Jan. 17, 2020. [Online]. Available: https://www.ideo.com/post/ai-ethics-collaborative-activities-for-designers
[35] "Data ethics workbook," U.K. Department for Digital, Culture, Media & Sport, June 13, 2018. [Online]. Available: https://www.gov.uk/government/publications/data-ethics-workbook/data-ethics-workbook
[36] J. Snow, "Amazon's face recognition falsely matched 28 members of Congress with mugshots," *ACLU.org*, July 26, 2018. [Online]. Available: https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28
[37] "Face Recognition Vendor Test (FRVT) ongoing," National Institute of Standards and Technology. Accessed on: Jan. 17, 2020. [Online]. Available: https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing.
[38] "AI Fairness 360 Open Source Toolkit." Accessed on: Jan. 17, 2020. [Online]. Available: http://aif360.mybluemix.net/
[39] T. Bolukbasi, K. Chang, J. Zou, V. Saligrama, and A. Kalai, "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," *arXiv.org*, July 21, 2016. [Online]. Available: https://arxiv.org/abs/1607.06520
[40] "Aequitas," *GitHub.com*. Accessed on: Jan. 17, 2020. [Online]. Available: https://github.com/dssg/aequitas
[41] "fairlearn," *GitHub.com*. Accessed on: Jan. 17, 2020. [Online]. Available: https://github.com/Microsoft/fairlearn
[42] "What-If Tool," People + AI Research (PAIR). Accessed on: Jan. 17, 2020. [Online]. Available: https://pair-code.github.io/what-if-tool/
[43] "Facets—Know your data," People + AI Research (PAIR). Accessed on: Jan. 17, 2020. [Online]. Available: https://pair-code.github.io/facets/
[44] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K. Chang, "Men also like shopping: Reducing gender bias amplification using Corpus-level constraints," *arXiv.org*, July 29, 2017.[Online]. Available: https://arxiv.org/pdf/1707.09457.pdf
[45] A. Feng and S. Wu, "The myth of the impartial machine," *Parametric Press*, no. 1, May 1, 2019. [Online]. Available: https://parametric.press/issue-01/the-myth-of-the-impartial-machine/
[46] D. Sculley et al., "Hidden technical debt in machine learning systems." Accessed on: Jan. 17, 2020. [Online]. Available: https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

[47] "An ethics guidelines global inventory," Algorithm Watch. Accessed on: Jan. 17, 2020. [Online]. Available: https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/
[48] T. Hagendorff, "The ethics of AI ethics: An evaluation of guidelines," *arXiv.org*, Oct. 11, 2019. [Online]. Available: https://arxiv.org/abs/1903.03425
[49] S. Lebowitz and S. Lee, "20 cognitive biases that screw up your decisions," *Business Insider*, Aug. 26, 2015. [Online]. Available: https://www.businessinsider.com/cognitive-biases-that-affect-decisions-2015-8
[50] K. Hao, "This is how AI bias really happens—and why it's so hard to fix," *MIT Technology Review*, Feb. 4, 2020. [Online]. Available: https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/
[51] M. Simon, "HP looking into claim webcams can't see black people," *CNN.com*, Dec. 23, 2009. [Online]. Available: http://www.cnn.com/2009/TECH/12/22/hp.webcams/index.html
[52] E. Hayasaki, "Is AI sexist?" *Foreign Policy*, Jan. 16, 2017. [Online]. Available: https://foreignpolicy.com/2017/01/16/women-vs-the-machine/
[53] Data&Society, "Algorithmic accountability: A primer," Prepared for the Congressional Progressive Caucus Tech Algorithm Briefing: How Algorithms Perpetuate Racial Bias and Inequality, April 18, 2018. [Online]. Available: https://datasociety.net/wp-content/uploads/2018/04/Data_Society_Algorithmic_Accountability_Primer_FINAL-4.pdf
[54] J. Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," *Reuters.com*, Oct. 9, 2018. [Online]. Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G
[55] E. Lacey, "The toxic potential of YouTube's feedback loop," *Wired*, July 13, 2019. [Online]. Available: https://www.wired.com/story/the-toxic-potential-of-youtubes-feedback-loop/
[56] Z. Rogers, "Have strategists drunk the 'AI race' Kool-Aid," *War on the Rocks*, June 4, 2019. [Online]. Available: https://warontherocks.com/2019/06/have-strategists-drunk-the-ai-race-kool-aid/
[57] A. MacGillis, "The case against Boeing," *The New Yorker*, Nov. 11, 2019. [Online]. Available: https://www.newyorker.com/magazine/2019/11/18/the-case-against-boeing
[58] S. Romero, "Wielding rocks and knives, Arizonans attack self-driving cars," *The New York Times*, Dec. 31, 2018. [Online]. Available: https://www.nytimes.com/2018/12/31/us/waymo-self-driving-cars-arizona-attacks.html
[59] "Virtue signaling," *Dictionary.com*. Accessed on: Jan. 17, 2020. [Online]. Available: https://www.dictionary.com/browse/virtue-signaling
[60] "An ethics guidelines global inventory," Algorithm Watch. Accessed on: Jan. 17, 2020. [Online]. Available: https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/
[61] G. C. Allen, "Understanding China's AI strategy," Center for a New American Security, Feb. 6, 2019. [Online]. Available: https://www.cnas.org/publications/reports/understanding-chinas-ai-strategy
[62] M. C. Horowitz, "Artificial intelligence, international competition, and the balance of power," *Texas National Security Review*, vo. 1, no. 3, May 2018. [Online]. Available: https://tnsr.org/2018/05/artificial-intelligence-international-competition-and-the-balance-of-power/
[63] Z. Rogers, "Have strategists drunk the 'AI race' Kool-Aid," *War on the Rocks*, June 4, 2019. [Online]. Available: https://warontherocks.com/2019/06/have-strategists-drunk-the-ai-race-kool-aid/
[64] J. Decker, "Renewing defense innovation: Five incentives for forming Pentagon-startup partnerships," *War on the Rocks*, May 3, 2018. [Online]. Available: https://warontherocks.com/2018/05/renewing-defense-innovation-five-incentives-for-forming-pentagon-startup-partnerships/
[65] R. Ramachandran, "How artificial intelligence is changing cyber security landscape and preventing cyber attacks," *Entrepreneur India*, Sep. 14, 2019. [Online]. Available: https://www.entrepreneur.com/article/339509
[66] "How artificial intelligence could increase the risk of nuclear war," *The RAND Blog*, April 23, 2018. [Online]. Available: https://www.rand.org/blog/articles/2018/04/how-artificial-intelligence-could-increase-the-risk.html
[67] S. Ahmed et al., "AI, China, Russia, and the global order: Technological, political, global, and creative perspectives," *A Strategic Multilayer Assessment (SMA) Periodic Publication*, Dec. 2018. [Online]. Available: http://static1.1.sqspcdn.com/static/f/1399691/28061274/1547846008013/AI+China+Russia+Global+WP_FINAL.pdf
[68] J. Ding, P. Triolo, and S. Sacks, "Chinese interests take a big seat at the AI governance table," *New America blog*, June 20, 2018. [Online]. Available: https://www.newamerica.org/cybersecurity-initiative/digichina/blog/chinese-interests-take-big-seat-ai-governance-table/

[69] "Air Force Association with Lieutenant General VeraLinn 'Dash' Jamieson, Deputy Chief of Staff for Air Force Intelligence, Surveillance and Reconnaissance," *afa.org*, July 26, 2018. [Online]. Available: https://www.afa.org/content/dam/afa/news-images/Jamieson%20Breakfast%20Transcript.pdf
[70] C. Cornillie, "Finding the artificial intelligence money in the fiscal 2020 budget," *Bloomberg Government*, March 28, 2019. [Online]. Available: https://about.bgov.com/news/finding-artificial-intelligence-money-fiscal-2020-budget/
[71] B. Vincent, "Administration projects agencies will spend $1 billion on artificial intelligence next year," *Nextgov*, Sep. 10, 2019. [Online]. Available: https://www.nextgov.com/emerging-tech/2019/09/administration-projects-agencies-will-spend-1-billion-artificial-intelligence-next-year/159781/
[72] A. Ashwin and Z. Arnold, "Chinese public AI R&D spending: Provisional findings," Issue Brief, Center for Security and Emerging Technology, Dec. 2019. [Online]. Available: https://cset.georgetown.edu/wp-content/uploads/Chinese-Public-AI-RD-Spending-Provisional-Findings-2.pdf
[73] J. Decker, "Renewing defense innovation: Five incentives for forming Pentagon-startup partnerships," *War on the Rocks*, May 3, 2018. [Online]. Available: https://warontherocks.com/2018/05/renewing-defense-innovation-five-incentives-for-forming-pentagon-startup-partnerships/
[74] T. Stephens, "The ethics of defense technology development: An investor's perspective," *Medium*, Dec. 4, 2019. [Online]. Available: https://medium.com/@traestephens/the-ethics-of-defense-technology-development-an-investors-perspective-45c71bf6e6af
[75] A. Lowther and C. McGiffin, "America needs a 'dead hand,'" *War on the Rocks*, Aug. 16, 2019. [Online]. Available: https://warontherocks.com/2019/08/america-needs-a-dead-hand/
[76] C. White, "Chinese big tech is using Zimbabwe citizens as guinea pigs to identify and track black people," *Daily Caller*, Dec. 2, 2019. [Online]. Available: https://dailycaller.com/2019/12/02/china-surveillance-africa-facial-recognition/
[77] R. Adams, "Hong Kong protesters are worried about facial recognition technology. But there are many other ways they're being watched," *BuzzFeed News*, Aug. 17, 2019. [Online]. Available: https://www.buzzfeednews.com/article/rosalindadams/hong-kong-protests-paranoia-facial-recognition-lasers
[78] D. Z. Morris, "China will block travel for those with bad 'social credit,'" *Fortune*, March 18, 2018. [Online]. Available: https://fortune.com/2018/03/18/china-travel-ban-social-credit/
[79] R. Hughes, "China Uighurs: All you need to know on Muslim 'crackdown,'" *BBC News*, Nov. 8, 2018. [Online]. Available: https://www.bbc.com/news/world-asia-china-45474279
[80] Office of the Secretary of Defense, *Nuclear Posture Review*. Washington, DC: Department of Defense, 2018. [Online]. Available: https://media.defense.gov/2018/Feb/02/2001872886/-1/-1/1/2018-NUCLEAR-POSTURE-REVIEW-FINAL-REPORT.PDF
[81] J. Goldsmith, "Fire when ready," *Foreign Policy*, March 20, 2012. [Online]. Available: https://foreignpolicy.com/2012/03/20/fire-when-ready/
[82] Defense Information Board, "AI principles: Recommendations on the ethical use of artificial intelligence by the Department of Defense." Accessed on: Jan. 21, 2020. [Online]. Available: https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF
[83] See p.6-7 of Defense Information Board, "AI principles: Recommendations on the ethical use of artificial intelligence by the Department of Defense." Accessed on: Jan. 21, 2020. [Online]. Available: https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF ]
[84] Office of the Director of National Intelligence, "The AIM initiative: A strategy for augmenting intelligence using machines," Jan. 16, 2019. [Online]. Available: https://www.dni.gov/index.php/newsroom/reports-publications/item/1940-the-aim-initiative-a-strategy-for-augmenting-intelligence-using-machines
[85] T. Stephens, "The ethics of defense technology development: An investor's perspective," *Medium*, Dec. 4, 2019. [Online]. Available: https://medium.com/@traestephens/the-ethics-of-defense-technology-development-an-investors-perspective-45c71bf6e6af
[86] C. Middleton, "Want a facial recognition system? Buy Chinese—says US government," *Internet of Business*, June 29, 2018. [Online]. Available: https://internetofbusiness.com/want-a-facial-recognition-system-buy-chinese-says-us-government/
[87] M. Johnson, J. M. Bradshaw, R. R. Hoffman, P. J. Feltovich, and D. D. Woods, "Seven cardinal virtues of human-machine teamwork: Examples from the DARPA robotic challenge," *IEEE Intelligent Systems*, Nov./Dec. 2014.

[Online]. Available: http://www.jeffreymbradshaw.net/publications/56.%20Human-Robot%20Teamwork_IEEE%20IS-2014.pdf

[88] A. Campolo, M. Sanfilippo, M. Whittaker, and K. Crawford, "AI Now 2017 report," AI Now Institute. Accessed on: Jan. 21, 2020. [Online]. Available: https://ainowinstitute.org/AI_Now_2017_Report.html

[89] R. Meyer, "How the carmakers trumped themselves," *The Atlantic*, June 20, 2018. [Online]. Available: https://www.theatlantic.com/science/archive/2018/06/how-the-carmakers-trumped-themselves/562400/

[90] X. Dou and J. Linn, "Why and how do new vehicle fuel economy standards affect consumer vehicle purchases?" *Resources*, Sep. 4, 2018. [Online]. Available: https://www.resourcesmag.org/common-resources/why-and-how-do-new-vehicle-fuel-economy-standards-affect-consumer-vehicle-purchases/

[91] "2019 Tesla model S," *fueleconomy.gov*. Accessed on: Jan. 21, 2020. [Online]. Available: https://www.fueleconomy.gov/feg/bymodel/2019_Tesla_Model_S.shtml

[92] "Algorithms and artificial intelligence: CNIL's report on the ethical issues," CNIL [Commission Nationale de l'Informatique et des Libertés], May 25, 2018. [Online]. Available: https://www.cnil.fr/en/algorithms-and-artificial-intelligence-cnils-report-ethical-issues

[93] M. Kearns, "The ethical algorithm," Carnegie Council for Ethics in International Affairs, Nov. 6, 2019. [Online]. Available: https://www.carnegiecouncil.org/studio/multimedia/20191106-the-ethical-algorithm-michael-kearns

[94] E. Lacey, "The toxic potential of YouTube's feedback loop," *Wired*, July 13, 2019. [Online]. Available: https://www.wired.com/story/the-toxic-potential-of-youtubes-feedback-loop/

[95] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *arXiv.org*, Sep. 22, 2019. [Online]. Available: https://arxiv.org/abs/1811.10154

[96] A. Narayanan, "21 fairness definitions and their politics," presented at Conference on Fairness, Accountability, and Transparency, Feb. 23, 2018. [Online]. Available: https://fairmlbook.org/tutorial2.html

[97] M. Kearns, "The ethical algorithm," Carnegie Council for Ethics in International Affairs, Nov. 6, 2019. [Online]. Available: https://www.carnegiecouncil.org/studio/multimedia/20191106-the-ethical-algorithm-michael-kearns

[98] T. Gebru et al., "Datasheets for datasets," *arXiv.org*, Jan. 14, 2020. [Online]. Available: https://arxiv.org/abs/1803.09010

[99] M. Mitchell et al., "Model cards for model reporting," *arXiv.org*, Jan. 14, 2019. [Online]. Available: https://arxiv.org/abs/1810.03993

[100] J. Stoyanovich and B. Howe, "Follow the data! Algorithmic transparency starts with data transparency," Shorenstein Center on Media, Politics and Public Policy, Harvard Kennedy School, Nov. 27, 2018. [Online]. Available: https://ai.shorensteincenter.org/ideas/2018/11/26/follow-the-data-algorithmic-transparency-starts-with-data-transparency

[101] "Algorithms and artificial intelligence: CNIL's report on the ethical issues," CNIL [Commission Nationale de l'Informatique et des Libertés], May 25, 2018. [Online]. Available: https://www.cnil.fr/en/algorithms-and-artificial-intelligence-cnils-report-ethical-issues

[102] R. Sandler, "Amazon, Microsoft, Wayfair: Employees stage internal protests against working with ICE," *Forbes*, July 19, 2019. [Online]. Available: https://www.forbes.com/sites/rachelsandler/2019/07/19/amazon-salesforce-wayfair-employees-stage-internal-protests-for-working-with-ice/

[103] J. Bhuiyan, "How the Google walkout transformed tech workers into activists," *Los Angeles Times*, Nov. 6, 2019. [Online]. Available: https://www.latimes.com/business/technology/story/2019-11-06/google-employee-walkout-tech-industry-activism

[104] S. Romero, "Wielding rocks and knives, Arizonans attack self-driving cars," *The New York Times*, Dec. 31, 2018. [Online]. Available: https://www.nytimes.com/2018/12/31/us/waymo-self-driving-cars-arizona-attacks.html

[105] M. Whittaker et al., *AI Now Report 2018*. New York, NY, USA: AI Now Institute, 2018. [Online]. Available: https://ainowinstitute.org/AI_Now_2018_Report.pdf

[106] Z. Rogers, "Have strategists drunk the 'AI race' Kool-Aid?" *War on the Rocks*, June 4, 2019. [Online]. Available: https://warontherocks.com/2019/06/have-strategists-drunk-the-ai-race-kool-aid/

[107] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science,* vol. 366, no. 6464, pp. 447-453, Oct. 25, 2019. [Online]. Available: https://science.sciencemag.org/content/366/6464/447

[108] M. Whittaker et al., *AI Now Report 2018*. New York, NY, USA: AI Now Institute, 2018. [Online]. Available: https://ainowinstitute.org/AI_Now_2018_Report.pdf

[109] R. Brandom, "Facebook has been charged with housing discrimination by the US government," *The Verge*, March 28, 2019. [Online]. Available: https://www.theverge.com/2019/3/28/18285178/facebook-hud-lawsuit-fair-housing-discrimination
[110] C. Haskins, "How Ring transmits fear to the American suburbs," *Vice*, Dec. 6, 2019. [Online]. Available: https://www.vice.com/en_us/article/ywaa57/how-ring-transmits-fear-to-american-suburbs
[111] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," *ProPublica*, May 23, 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[112] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science,* vol. 366, no. 6464, pp. 447-453, Oct. 25, 2019. [Online]. Available: https://science.sciencemag.org/content/366/6464/447
[113] M. Whittaker et al., *AI Now Report 2018*. New York, NY, USA: AI Now Institute, 2018. [Online]. Available: https://ainowinstitute.org/AI_Now_2018_Report.pdf
[114] A. Gonfalonieri, "Why machine learning models degrade in production," towards data science, July 25, 2019. [Online]. Available: https://towardsdatascience.com/why-machine-learning-models-degrade-in-production-d0f2108e9214
[115] "Algorithms and artificial intelligence: CNIL's report on the ethical issues," CNIL [Commission Nationale de l'Informatique et des Libertés], May 25, 2018. [Online]. Available: https://www.cnil.fr/en/algorithms-and-artificial-intelligence-cnils-report-ethical-issues
[116] M. Whittaker et al., *AI Now Report 2018*. New York, NY, USA: AI Now Institute, 2018. [Online]. Available: https://ainowinstitute.org/AI_Now_2018_Report.pdf
[117] ENERGY STAR homepage. Accessed on: Jan. 21, 2020. [Online]. Available: https://www.energystar.gov/
[118] C. Martin and M. Dent, "How Nestle, Google and other businesses make money by going green," *Los Angeles Times*, Sep. 20, 2019. [Online]. Available: https://www.latimes.com/business/story/2019-09-20/how-businesses-profit-from-environmentalism
[119] A. Campolo, M. Sanfilippo, M. Whittaker, and K. Crawford, "AI Now 2017 report," AI Now Institute. Accessed on: Jan. 21, 2020. [Online]. Available: https://ainowinstitute.org/AI_Now_2017_Report.html
[120] J. Stoyanovich and B. Howe, "Follow the data! Algorithmic transparency starts with data transparency," Shorenstein Center on Media, Politics and Public Policy, Harvard Kennedy School, Nov. 27, 2018. [Online]. Available: https://ai.shorensteincenter.org/ideas/2018/11/26/follow-the-data-algorithmic-transparency-starts-with-data-transparency
[121] Z. C. Lipton, "The doctor just won't accept that," *arXiv.org*, Nov. 24, 2017. [Online]. Available: https://arxiv.org/abs/1711.08037
[122] "Algorithms and artificial intelligence: CNIL's report on the ethical issues," CNIL [Commission Nationale de l'Informatique et des Libertés], May 25, 2018. [Online]. Available: https://www.cnil.fr/en/algorithms-and-artificial-intelligence-cnils-report-ethical-issues
[123] M. Whittaker et al., *AI Now Report 2018*. New York, NY, USA: AI Now Institute, 2018. [Online]. Available: https://ainowinstitute.org/AI_Now_2018_Report.pdf
[124] Occupational Safety and Health Administration, "OSHA's Nationally Recognized Testing Laboratory (NRTL) program," *OSHA.gov*. Accessed on: Jan. 30, 2020. [Online]. Available: https://www.osha.gov/dts/otpca/nrtl/
[125] F. Balamuth et al., "Improving recognition of pediatric severe sepsis in the emergency department: Contributions of a vital sign–based electronic alert and bedside clinician identification," *Annals of Emergency Medicine*, vol. 79, no. 6, pp. 759-768.e2, Dec. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0196064417303153
[126] G. Siddiqui, "Why doctors reject tools that make their jobs easier," *Scientific American*, Oct. 15, 2018. [Online]. Available: https://blogs.scientificamerican.com/observations/why-doctors-reject-tools-that-make-their-jobs-easier/
[127] A. M. Barry-Jester, B. Casselman, and D. Goldstein, "Should prison sentences be based on crimes that haven't been committed yet?' *FiveThirtyEight*, Aug. 4, 2015. [Online]. Available: https://fivethirtyeight.com/features/prison-reform-risk-assessment/
[128] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," *ProPublica*, May 23, 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[129] S. Corbett-Davies, E. Pierson, A. Feller, and S. Goel, "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear," *Washington Post*, Oct. 17, 2016. [Online]. Available:

https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?noredirect=on&utm_term=.a9cfb19a549d

[130] "Case of first impression," *Legal Dictionary*, March 21, 2017. [Online]. Available: https://legaldictionary.net/case-first-impression/

[131] "Fair cross section requirement," Stephen G. Rodriquez & Partners. Accessed on: Jan. 21, 2020. [Online]. Available: https://www.lacriminaldefenseattorney.com/legal-dictionary/f/fair-cross-section-requirement/

[132] I. Masic, M. Miokovic, and B. Muhamedagic, "Evidence based medicine—new approaches and challenges," *Acta Informatica Medica*, vol. 16, no. 4, pp. 219–225, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3789163/

[133] "Hippocratic Oath," *Encyclopaedia Britannica*, Dec. 4, 2019. [Online]. Available: https://www.britannica.com/topic/Hippocratic-oath

[134] R. Vought, "Guidance for regulation of artificial intelligence applications," Draft memorandum, *WhiteHouse.gov*. Accessed on: Jan. 21, 2020. [Online]. Available: https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf

[135] G. Vyse, "Three American cities have now banned the use of facial recognition technology in local government amid concerns it's inaccurate and biased," *Governing.com*, July 24, 2019. [Online]. Available: https://www.governing.com/topics/public-justice-safety/gov-cities-ban-government-use-facial-recognition.html

[136] A. MacGillis, "The case against Boeing," *The New Yorker*, Nov. 11, 2019. [Online]. Available: https://www.newyorker.com/magazine/2019/11/18/the-case-against-boeing

[137] S. Gibbs, "Tesla Model S cleared by auto safety regulator after fatal Autopilot crash," *The Guardian*, Jan. 20, 2017. [Online]. Available: https://www.theguardian.com/technology/2017/jan/20/tesla-model-s-cleared-auto-safety-regulator-after-fatal-autopilot-crash

[138] "How WhatsApp helped turn an Indian village into a lynch mob," *BBC News*, July 19, 2018. [Online]. Available: https://www.bbc.com/news/world-asia-india-44856910

[139] A. Dafoe, "AI governance: A research agenda," Future of Humanity Institute, University of Oxford, Oxford, UK, Aug. 27, 2018. [Online]. Available: https://www.fhi.ox.ac.uk/wp-content/uploads/GovAIAgenda.pdf

[140] B. Marr, "The AI skills crisis and how to close the gap," *Forbes*, June 25, 2018. [Online]. Available: https://www.forbes.com/sites/bernardmarr/2018/06/25/the-ai-skills-crisis-and-how-to-close-the-gap/#6525b57b31f3

[141] J. F. Gagne, F. Karmanov, and S. Hudson, "Global AI talent pool report 2018," *jfgagne.ai*. Accessed on: Jan. 21, 2020. [Online]. Available: https://jfgagne.ai/talent/

[142] "About," Generation AI Nexus. Accessed on: Jan. 21, 2020. [Online]. Available: https://www.ainexus.org/about

[143] E. Hunt, "Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter," *The Guardian*, March 24, 2016. [Online]. Available: https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter

[144] C. Hughes, "Multidisciplinary teamwork ensures better healthcare outcomes," Association for Talent Development, July 12, 2018. [Online]. Available: https://www.td.org/insights/multidisciplinary-teamwork-ensures-better-healthcare-outcomes

[145] V. Hunt, D. Layton, and S. Prince, "Why diversity matters," McKinsey & Company, Jan. 2015. [Online]. Available: https://www.mckinsey.com/business-functions/organization/our-insights/why-diversity-matters

[146] K. W. Phillips, "How diversity makes us smarter," *Greater Good Magazine*, Sep. 18, 2017. [Online]. Available: https://greatergood.berkeley.edu/article/item/how_diversity_makes_us_smarter

[147] M. Whittaker et al., *AI Now Report 2018*. New York, NY, USA: AI Now Institute, 2018. [Online]. Available: https://ainowinstitute.org/AI_Now_2018_Report.pdf

[148] "Workplace automation: How AI is coming for your job," *Financial Times*, Sep. 28, 2019. [Online]. Available: https://www.ft.com/content/c4bf787a-d4a0-11e9-a0bd-ab8ec6435630