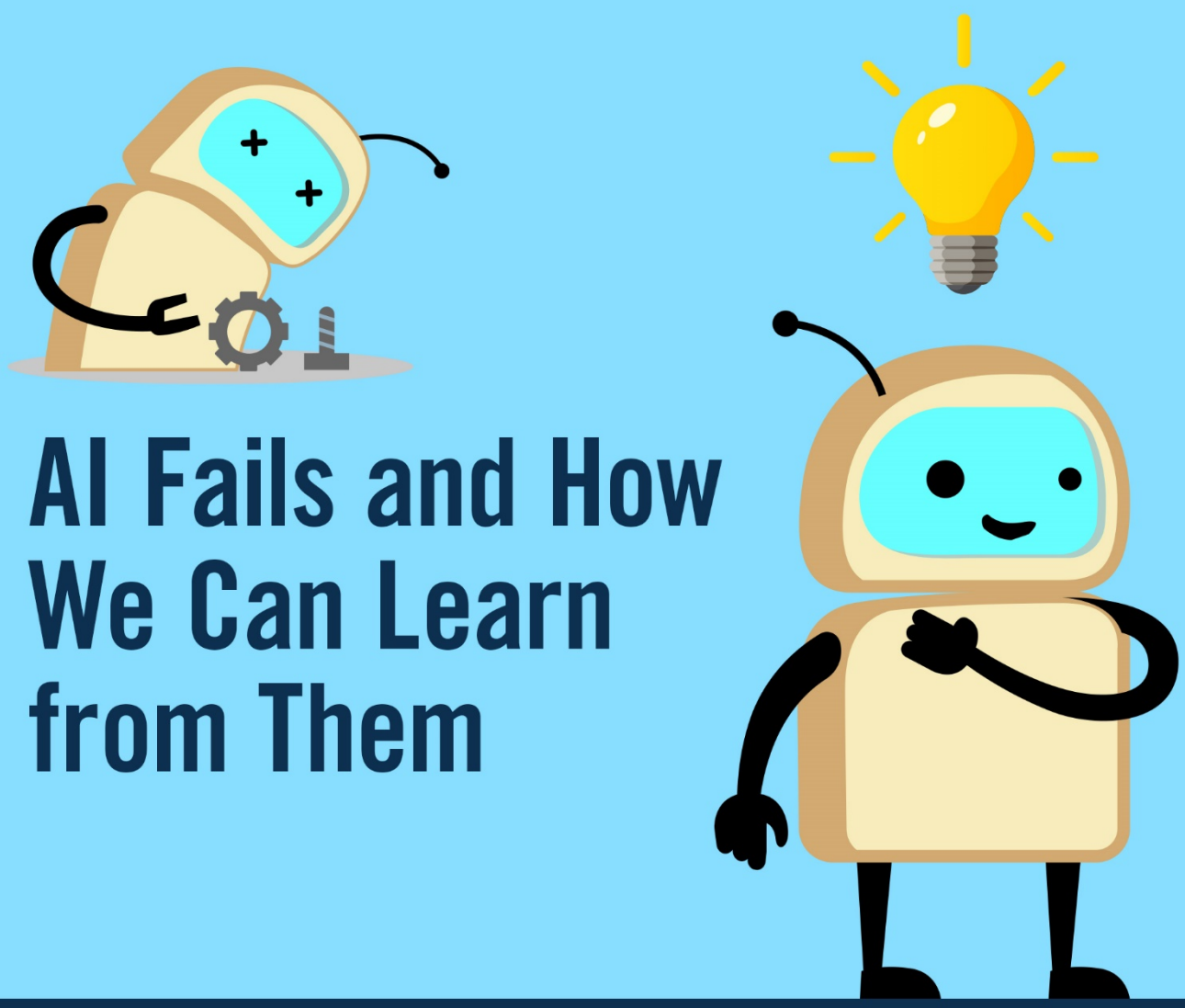


AI Fails and How We Can Learn from Them

Mr. Jonathan Rotner, Mr. Ron Hodge, and Dr. Lura Danley

July 2020

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®

OUR SUCCESS WITH ARTIFICIAL INTELLIGENCE HINGES ON HOW WE LEARN FROM FAILURES

Benjamin Franklin once said, “If you don't plan to fail, you fail to plan.”[1] Yet the notion of failure can make people uncomfortable, which pushes them to avoid talking about fails, instead of seeing failure as an opportunity. This paper makes the case that understanding and sharing information about artificial intelligence (AI) failures can provide lessons for better preventing, anticipating, or mitigating future fails.

These lessons derive from a more holistic view of automated technologies. Such technologies are more than independent widgets; they are part of a complex ecosystem that interacts with and influences human behavior, decision making, preferences, strategies, and ways of life in beneficial, and sometimes less beneficial, ways.

"AI Fails" proposes a shift in perspective: we should measure the success of an AI system by its impact on human beings, rather than prioritizing its mathematical or economic properties (e.g., accuracy, false alarm rate, or efficiency). Such a shift has the potential to empower the development and deployment of amazing as well as responsible AI.

AI’s Balancing Act: Amazing Possibilities and Potential Harm

The most advanced of these technologies – AI – is not just emerging everywhere, it is being rapidly integrated into people’s lives. The 2018 Department of Defense AI Strategy provides a great way to think about AI: simply as “the ability of machines to perform tasks that normally require human intelligence.”[2]

AI has tremendously valuable applications, for instance when it promises to translate a person’s conversation into another language in real time,[3] more accurately diagnose patients and propose treatments,[4] or take care of the elderly.[5] In these cases, everyone can enthusiastically accept AI. However, when it is reported that individuals can be microtargeted with falsified information to sway their election choices,[6] that mass surveillance leads to imprisonment and suppression of populations,[7,8,9] or that self-driving cars have caused deaths,[10] people realize that AI can lead to real harm. In these cases, the belief in AI’s inevitability can elicit terror. As AI developers and deployers, we experience and observe both extremes of this continuum, and everything in between.

Embracing and Learning from AI’s Deep History

This paper draws heavily on decades of research and expertise, particularly in domains where the cost of failure is high enough (e.g., the military or aviation) that human factors and human-machine teaming have been thoroughly analyzed and the findings well integrated into system development. Though many of these fails and lessons apply to more than AI, collectively they represent the systemic challenges faced by AI developers and practitioners.

In addition, AI is fundamentally different from other technologies in several ways, notably that 1) decisions aren't static, since data and model versions are updated all the time, and 2) models don't always come with explanations, which means that even designers may not know what factors affect or even drive decisions.

AI is also fundamentally different in the way it interacts with humans, since 1) the technology is new enough to most people that they can be (and have been) influenced to trust an AI system more than they should, and 2) its reach is vast enough that a single AI with a single programmed objective can scale to affect human decisions at a global level.

> *In this paper the term "AI" encompasses capabilities ranging from previous and often simpler versions of automated technology whose lessons are still applicable, through more sophisticated AI approaches, some of whose lessons are relatively new and unresolved.*

Intended Audience

This paper is intended for:

- AI experts who already knows about data, models, and development. But we can't build AI in a vacuum. Especially because AI systems are increasingly affecting human behavior and livelihoods, we must take steps to better understand how the system will interact with its environment, and how to help non-experts become better informed, engaged, and empowered as they interact with the technology.
- Everyone else, including AI users, policymakers, and those affected by AI. Because AI applications are steadily being integrated into daily life, these readers need to understand enough about how a particular application works, its intended uses, and its limitations in order to use it appropriately and beneficially.

Studying previous generations of automated technologies can help us to identify stepping-stones for developing AI, introduce AI to new audiences, and provide context for understanding today's challenges. This paper aims to serve as a tool for the many AI experts, engineers, students, decision makers, and others who will be required to develop, deliver, and use AI as part of their roles in the modern workforce or simply as citizens.

Key Lessons

This first half of this paper presents examples of AI fails, along with research- and evidence-based discussions of how we might view these fails from a human-centric perspective. The second half of the paper offers recommendations on practical steps that can be taken, right now, to apply these insights.

The key lessons from a human-centric mindset regarding AI are:

1. *Developing AI is a multidisciplinary problem.* AI challenges and products can be technical or based on human behavior, and often are a blend of the two. By including multidisciplinary perspectives, we can more clearly articulate design tradeoffs between different priorities and outcomes. Then the broader team can work towards having the human and technical sides of AI reinforce, rather than interfere with, each other.

2. *An AI application affects more than just end-users*. Input from stakeholders is essential to helping us structure the AI's objectives to increase adoption and reduce potential undesired consequences. We need to involve end-users, domain experts, and the communities affected by AI, early and repeatedly. These stakeholders can also provide societal and political contexts of the domain where the AI will operate, and can share information about how previous attempts to address their issues fared. Adopting the mindset that all stakeholders are our customers will help us design with all their goals in mind and to create resources that give them the context and tools they need to work with the AI successfully.

3. *Our assumptions shape AI*. There is no such thing as a neutral, impartial, or unbiased AI. Our underlying assumptions about the data, model, user behaviors, and environment affect the AI's objectives and outcomes. We should remember that those assumptions stem from our own, often subconscious, social values, and that an AI system can unintentionally replicate and encode those values into practice when the AI is deployed. Given the current composition of the AI development workforce, all too often those values represent how young, white, technically oriented, Western men interact with the world, and no homogeneous group, regardless of its characteristics, can reflect the full spectrum of priorities and considerations of all possible system users. To address this concern, we should strive for diversity in teammates' experiences and backgrounds, be responsive when teammates or stakeholders raise issues, and provide documentation about the assumptions that went into the AI system.

4. *Documentation can be a key tool in reducing future failures.* When we make a good product, end-users and consumers will want to use it, and other AI developers may want to repurpose it for their own domains. To do so appropriately and safely, they will need to know what uses of the AI we did and did not intend, the design tradeoffs we considered and acted on, and the risks we identified and the mitigations we put in place. Therefore, the original developers need to capture their assumptions and tradeoff decisions, and organizations have to develop processes that facilitate proactive and ongoing outreach.

5. *Accountability must be tied to an AI's impact.* When using the data or AI could cause financial, psychological, physical, or other harm, we must consider if AI offers the best solution to a given problem. In addition to our good intentions and commitment to ethical values, the oversight, accountability, and enforcement mechanisms in place can facilitate ethical outcomes. These mechanisms shouldn't equate to excessive standardization or policies that stymie technological development. Instead, they should encourage proactive approaches to implementing the previous lessons. The more the AI application could influence people's behavior and livelihoods, the more careful considerations and governance are needed.

Reach Out to Us

This document is intended to be a community resource and would benefit from the addition of your input. To submit an example of AI success, failure, or specific solution, send an email to one of the authors: Jonathan Rotner, jrotner@mitre.org

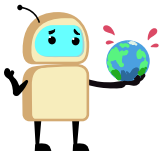
An online version of this paper is hosted at https://sites.mitre.org/aifails

TABLE OF CONTENTS

HOW TO NAVIGATE THIS PAPER

The paper describes 20 overall fails, which are sorted into the 6 categories shown here:

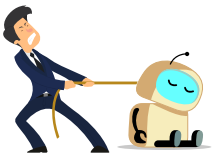
The Cult of AI
Perceiving AI to Be More Mature Than It Is



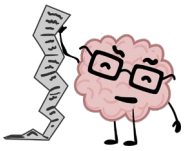
You Call This "Intelligence"?
AI Meets the Real World

Turning Lemons into Reflux
When AI Makes Things Worse

We're Not Done Yet
After Developing the AI



Failure to Launch
How People Can React to AI

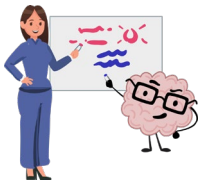


AI Registry
The Things We'll Need That Support AI

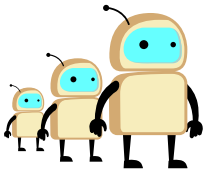
Each *category* includes a brief introduction, which is followed by the three or four fails relevant to that category.

Each *fail* demonstrates the results of AI misapplications and presents ways to learn from what went wrong. Every fail starts with a description, a discussion of why it's a fail and what happens as a result of the fail, and several real-world examples related to the theme of the fail. At the end of each fail is a list of lessons learned that could be applicable. The items in that list serve as hyperlinks that will take you to the specific recommendations and practical considerations behind each lesson learned.

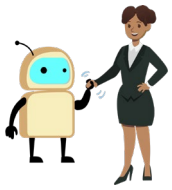
Finally, all 16 *lessons learned* (in 4 categories) are listed in the second half of the paper, as shown below:



Expand Early Project Considerations



Build Resiliency into the AI and the Organization



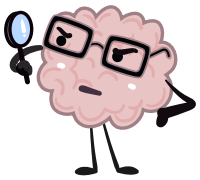
Calibrate Our Trust in the AI and the Data

Broaden the Ways to Assess AI's Impacts

The best way to navigate is to dive in and explore. So, go out of order, jump around to different sections, or follow what's most interesting to you! You can start with any of these icons, they're all hyperlinks too.

FAILS

The Cult of AI: Perceiving AI to Be More Mature Than It Is



AI is all about boundaries: the AI works well if we as developers and deployers define the task and genuinely understand the environment in which the AI will be used. New AI applications are exciting in part because they exceed previous technical boundaries − like AI winning at chess, then *Jeopardy*, then Go, then StarCraft. But what happens when we assume that AI is ready to break those barriers before the technology or the environment is truly ready? This section presents examples where AIs exceeded either technical or environmental limits – whether because AI was put in roles it wasn't suited for, user expectations didn't align with its abilities, or because the world was assumed to be simpler than it really is.

Fail #1. No Human Needed: the AI's Got This

Fail: We often intend to design AIs to assist their human partners, but what we create can end up *replacing* some human partners. When the AI isn't ready to completely perform the task without the help of humans, this could lead to significant problems.

Why is this a fail? Perception about what AI is suited for may not always align with the research. Deciding which tasks are better suited for humans or for machines can be traced back to Fitts's 'machines are better at' (MABA) list from 1951.[13] A modern-day interpretation of that list might allocate tasks that involve judgment, creativity, and intuition to humans, and tasks that involve responding quickly or storing and sifting through large amounts of data to the AI.[14,15] More advanced AI applications can be designed to blur those lines, but even in those cases the AI will likely need to interact with humans in some capacity.

Like any technology, AI may not work as intended or may have undesirable consequences. Consequently, if the AI is intended to work by itself, any design considerations meant to foster partnership will be overlooked, which will impose additional burdens on the human partners when they are called upon.[16,17]

Examples:

Microsoft released Tay, an AI chatbot designed "to engage and entertain" and learn from the communication patterns of the 18-to-24-year-olds with whom it interacted. Within hours, Tay started repeating some users' sexist, anti-Semitic, racist, and other inflammatory statements. Although the chatbot met its learning objective, the way it did so required individuals within Microsoft to modify the AI and address the public fallout from the experiment.[11]

Because Amazon employs so many warehouse workers, the company has used a heavily automated process that tracks employee productivity and is authorized to fire people without the intervention of a human supervisor. As a result, some employees have said they avoid using the bathroom for fear of being fired on the spot. Implementing this system has led to legal and public relations challenges, even if it did reduce the workload for the company's human resources employees or remaining supervisors.[12]

Did You Know?

A General Interpretation of Narrow AI

AI isn't new – it's been around for over 60 years.[18] But experts and laypeople characterize AI a little differently, and those misunderstandings can distort expectations.

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.[19] 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 won't meet user expectations equally 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).

The concept of Artificial General Intelligence (AGI), on the other hand, is closer to science fiction. These hypothetical systems could think and act like humans, would be almost fully self-reliant, and could 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, an ability only humans have today, would only be possible with AGI.

Where do really advanced modern technologies, such as self-driving cars, fit in? These technologies represent attempts to expand "narrow" AI. When an environment changes or the clarity of the task becomes muddied, it gets harder to develop a 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 recognizable to their drivers – they try to make the unknown a little more familiar and predictable, rather than reason abstractly as an AGI system would be expected to do.[20]

What happens when things fail? Semi-autonomous cars provide a great example of how the same burdens that have been studied and addressed over decades in the aviation industry are re-emerging in a new technology and marketplace.

Lost context – As more inputs and decisions are automated, human partners risk losing the context they often rely on to make informed decisions. Further, sometimes they can be surprised by decisions their AI partner makes because they fail to fully understand how that decision was made,[21] since information that they would usually rely on to make a decision is often obscured from them by AI processes. For example, when a semi-autonomous car passes control back to the human driver, the driver may have to make quick decisions about

what to do without knowing why the AI transferred control of the car to him or her, which increases the likelihood of making errors.

Cognitive drain – As AIs get better at conducting tasks that humans find dull and routine, humans can be left with only the hardest and most cognitively demanding tasks. For example, traveling in a semi-autonomous car might require human drivers to monitor both the vehicle to see if it's acting reliably, and the road to see if conditions require human intervention. Because the humans are then more engaged in more cognitively demanding work, they are at a higher risk of the negative effects of cognitive overload, such as decreased vigilance or increased likelihood of making errors.

Human error traded for new kinds of error – Human-AI coordination can lead to new sets of challenges and learning curves. For example, researchers have documented that drivers believe they will be able to respond to rare events more quickly and effectively than they actually can.[22] If this mistaken belief is unintentionally included in the AI's programming, it could create a dangerously false sense of security for both developers and drivers.

Reduced human skills or abilities – If the AI becomes responsible for doing everything, humans will have less opportunity to practice the skills that were often important in the development of their knowledge and expertise on the topic (i.e., experiences that enable them to perform more complex or nuanced activities). Driving studies have indicated that human attentiveness and monitoring of traffic and road conditions decrease as automation increases. Thus, at moments when experience and attention are needed most, they might potentially have atrophied due to humans' reliance on AI.

Lessons Learned from This AI Fail:

Hold AI to a Higher Standard • It's OK to Say No to Automation • AI Challenges Are Multidisciplinary, so They Require a Multidisciplinary Team • Make Our Assumptions Explicit • Try Human-AI Couples Counseling • Offer the User Choices • Promote Better Adoption through Gameplay

Return to Table of Contents

Fail #2. AI Perfectionists and AI "Pixie Dusters"

Fail: There is a temptation to overestimate the range and scale of problems that can be solved by technology. This can contribute to two mindsets: "perfectionists" who expect performance beyond what the AI can achieve, and "pixie dusters" who believe AI to be more broadly applicable than it is. Both groups could then reject current or future technical solutions (AI or not) that are more appropriate to a particular task.

Why is this a fail? Non-AI experts can have inflated expectations of AI's abilities. When AI is presented as having superhuman abilities based on proven mathematical principles, it is tremendously compelling to want to try it out.

Turn on the radio, ride the bus, watch a TV ad, and someone is talking about AI. AI hype has never been higher,[25] which means more people and organizations are asking, 'How can I have AI solve my problems?'

AI becomes even more appealing because of the belief that algorithms are "objective and true and scientific," since they are based on math. In reality, as mathematician and author Cathy O'Neil puts it, "algorithms are opinions embedded in code," and some vendors ask buyers to "put blind faith in big data."[26] Even AI experts can fall victim to this mentality, convinced that complex problems can be solved by purely technical solutions if the algorithm and its developer are brilliant enough.[27]

What can result is a false hope in a seemingly magical technology. As a result, people can want to apply it to everything, regardless of whether it's appropriate.

Examples:

In 2015, Amazon used an AI to find the top talent from stacks of resumes. One person involved with the trial run said, "Everyone wanted this holy grail... give[n] 100 resumes, it will spit out the top five, and we'll hire those." But because the AI was trained on data from previous hires, its selections reflected those existing patterns and strongly preferred male candidates to female ones.[23] Even after adjusting the AI and its hiring process, Amazon abandoned the project in 2017. The original holy grail expectation may have diverted the firm from designing a more balanced hiring process.

The 2012 Defense Science Board Study titled "The Role of Autonomy in DoD Systems" concluded that "Most [Defense Department] deployments of unmanned systems were motivated by the pressing needs of conflict, so systems were rushed to theater with inadequate support, resources, training and concepts of operation." This push to deploy first and understand later likely had an impact on warfighters' general opinions and future adoption of autonomous systems.[24]

What happens when things fail? Misaligned expectations can contribute to the rejection of relevant technical solutions. Two mentalities that emerge – "perfectionists" and "pixie dusters" (as in "AI is a magical bit of pixie dust that can be used to solve anything") – can both lead to disappointment and skepticism once expectations must confront reality.

Perfectionist deployers and users may expect perfect autonomy and a perfect understanding of autonomy, which could (rightly or wrongly) delay the adoption of AI until it meets those impossible standards. Perfectionists may prevent technologies from being explored and tested even in carefully monitored target environments, because they set too high a bar for acceptability.

In contrast, AI pixie-dusters may want to employ AI as soon and as widely as possible, even if an AI solution isn't appropriate to the problem. One common manifestation of this belief occurs when people want to take an excellent AI model and replicate it for a different problem. This technique is

In the end, it's about balance. AI has its limits and intended and appropriate uses. We have to identify the individual applications and environments for which AI is well suited, and better align non-experts' expectations to the way the AI will actually perform.

referred to as “transfer learning,” where “a model developed for one task is reused as the starting point for a model on a second task.”[28] While this approach can expedite the operationalization of a second AI model, problems arise when people are overly eager to attempt it. The new application must have the right data, equipment, environment, governance structures, and training in place for transfer learning to be successful.

Perhaps counterintuitively, an eagerness to adopt autonomy too early can backfire if the immature system behaves in unexpected, unpredictable, or dangerous ways. When pixie dusters have overinflated expectations of AI outcomes and the AI fails to meet those expectations, they can be dissuaded from trying other, even appropriate and helpful, AI-applications (as happened in the “AI Winter” in the 1980s[29]).[30]

In the end, it’s about balance. AI has its limits and intended and appropriate uses. We have to identify the individual applications and environments for which AI is well suited, and better align non-experts’ expectations to the way the AI will actually perform.

Lessons Learned from This AI Fail:

Hold AI to a Higher Standard • It’s OK to Say No to Automation • AI Challenges Are Multidisciplinary, so They Require a Multidisciplinary Team • Incorporate Privacy, Civil Liberties, and Security from the Beginning • Plan to Fail • Make Our Assumptions Explicit • Try Human-AI Couples Counseling • Offer the User Choices • Promote Better Adoption through Gameplay • Entrust Sector-specific Agencies to Establish AI Standards for Their Domains

Return to Table of Contents

Fail #3. AI Developers Are Wizards and Operators Are Muggles

Fail: When AI developers think we know how to solve a problem, we may overlook including input from the users of that AI, or the communities the AI will affect. Without consulting these groups, we may develop something that doesn’t match, or even conflicts with, what they want.

> *“Muggle” is a term used in the Harry Potter books to derogatorily refer to an individual who has no magical abilities yet lives in a magical world.*

Why is this a fail? It’s a natural inclination to assume that end-users will act the same way we do or will want the same results we want. Unless we include in the design and testing process the individuals who will use the AI, or communities affected by it, we’re unintentionally limiting the AI’s success and its adoption, as well as diminishing the value of other perspectives that would improve AI’s effectiveness.

Despite our long-standing recognition of how important it is to include those affected by what we're designing, we don't always follow through. Even if we do consult users, a single interview is not enough to discover how user behaviors and goals change in different environments or in response to different levels of pressure or emotional states, or how those goals and behaviors might shift over time.

Examples:

After one of the Boeing 737 MAX aircraft crashes, pilots were furious that they had not been told that the aircraft had new software, the software would override pilot commands in some rare but dangerous situations, and the pilot manual did not include mention of the software.[31,32]

Uber's self-driving car was not programmed to recognize jaywalking, only pedestrians crossing in or near a crosswalk,[33] which would work in some areas of the country but runs counter to the norms in others, putting those pedestrians in danger.

What happens when things fail? At best, working in a vacuum results in irritating system behavior – like a driver's seat that vibrates every time it wants to get the driver's attention.[34] Sometimes users may respond to misaligned goals by working around the AI, turning it off, or not adopting it at all. At worst, the objectives of the solution don't match users' goals, or it does the opposite of what users want. But with AI's scope and scale, the stakes can get higher.

Let's look at a relevant yet controversial AI topic to see how a different design perspective can result in drastically different outcomes. All over the country, federal, state, and local law enforcement agencies want to use facial recognition AI systems to identify criminals. As AI developers, we may want to make the technology as accurate or with as few false positives as possible, in order to correctly identify criminals. However, the communities that have been heavily policed understand the deep historical patterns of abuse and profiling that result, regardless of technology. As Betty Medsger, investigative reporter, writes, "being Black was enough [to justify surveillance]."[35] So if accuracy and false positives are the only consideration, we create an adoption challenge if communities push back against the technology, maybe leading to its not being deployed at all, even if it would be beneficial in certain situations. If we bridge this gap by involving these communities, we may learn about their tolerances for the technology and identify appropriate use cases for it.

If we start thinking about the 'customer' not only as the purchaser or user of the technology, but also as the community the deployed technology will affect, our perspective changes.

If we start thinking about the 'customer' not only as the purchaser or user of the technology, but also as the community the deployed technology will affect, our perspective changes.[36]

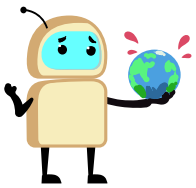
Lessons Learned from This AI Fail:

It's OK to Say No to Automation • AI Challenges Are Multidisciplinary, so They Require a Multidisciplinary Team • Involve the Communities Affected by the AI • Plan to Fail • Make Our Assumptions Explicit • Try Human-AI Couples Counseling • Offer the User Choices • Promote Better Adoption through Gameplay

• Envision Safeguards for AI Advocates • Entrust Sector-specific Agencies to Establish AI Standards for Their Domains

Return to Table of Contents

You Call This "Intelligence"? AI Meets the Real World



AI systems can perform specific, defined tasks so well that their capability can appear superhuman. For instance, 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.[37] However, it is important to remember that an AI's success is task specific and AI's ability to complete a task – such as recognizing images – is contingent on the data it receives and the environment it operates in. Because of this, sometimes AI applications are fooled in ways that humans never would be, particularly if these systems encounter situations beyond their abilities. The examples below describe situations where environmental factors exceeded AI's "superhuman" capabilities and invalidated any contingency planning that developers or deployers introduced.

Fail #4. Sensing Is Believing

Fail: When sensors are faulty, or the code that interprets the data is faulty, the result can be extremely damaging.

Why is this a fail? Humans use sight, smell, hearing, taste, and touch to perceive and make sense of the world. These senses work in tandem and can serve as backups for one another; for instance, you might smell smoke before you see it. But human senses and processing aren't perfect; they can be influenced, be confused, or degrade.

What happens when things fail? Similar to humans, some automated systems rely on sensors to get data about their operating environments and rely on code to process and act on that data. And like human senses, these

Examples:

When the battery died on early versions of "smart" thermostats, houses got really, really cold.[38] Later versions had appropriate protections built into the code to ensure this wouldn't happen.

A preliminary analysis of the Boeing 737 MAX airline crashes found that a faulty sensor "erroneously reported that the airplane was stalling… which triggered an automated system… to point the aircraft's nose down," when the aircraft was not actually stalling.[39] Boeing subsequently included the safety features that would have alerted pilots to the disagreement between working sensors and the failed sensor to all models.

A woman discovered that any person's fingerprint could unlock her phone's "vault-like security" after she had fitted the phone with a $3 screen protector. Customers were told to avoid logging in through fingerprint until the vendor could fix the code.[40]

sensors and the interpretation of their readings are imperfect and can be *influenced* by the composition or labelling of the training dataset, can get *confused* by erroneous or unexpected inputs, and can *degrade* as parts get older. AI applications tend to break if we haven't included redundancy, guardrails to control behavior, or code to gracefully deal with programming errors.

We can learn from a long history of research on sensor failure, for example in the automobile, power production, manufacturing, and aviation industries. In the latter case research findings have led to certification requirements like triple redundancy for any parts on an aircraft necessary for flight.[41]

Lessons Learned from This AI Fail:

AI Challenges Are Multidisciplinary, so They Require a Multidisciplinary Team • Plan to Fail • Ask for Help: Hire a Villain • Use Math to Reduce Bad Outcomes Caused by Math • Make Our Assumptions Explicit • Offer the User Choices • Monitor the AI's Impact and Establish Layers of Accountability • Require Objective, Third-party Verification and Validation

Return to Table of Contents

Fail #5. Insecure AI

Fail: When AI's software and information technology (IT) architecture are not hardened against cybersecurity threats, users and systems are vulnerable to accidental or malicious interference.

Why is this a fail? AI's software and IT architecture are as vulnerable to cybersecurity threats as other connected technologies – and potentially vulnerable in new ways as well. Just deploying an AI into the world introduces it as a new attack surface (i.e., something to attack).[44] Even the most secure AI can face continuous attacks that aim to expose, alter, disable, destroy, or gain unauthorized access to it. Therefore, we must design all software systems in a way that makes cyber protections and privacy considerations inherent to the design from the beginning.[45]

Examples:

Responding to what it thought were explicit commands but was actually background noise, an Amazon Echo recorded a family's private conversation and sent it to a random user.[42] This is one way in which users can unknowingly cause data spills.

HIS Group, a Japanese hotel chain, installed in-room cameras with facial recognition and speech recognition to cater to guest's needs. Hackers were able to remotely view video streams.[43] This is one way purchasers can unintentionally create situations that attract malicious behavior.

What happens when things fail? Smart devices – like internet-connected speakers, wireless door locks, and wireless implants – have increasingly been introduced into people's homes and even into their bodies, which makes the consequences of their being hacked especially terrifying.[46,47,48] Such systems are often networked because they rely on cloud resources to do some of the processing, or they communicate with other networked sensors. The market growth of these kind of products will make such devices more common.

> AI systems often have access to potentially sensitive user information

Another cybersecurity threat arises because AI systems often have access to potentially sensitive user information. For example, smart home devices have an unusual level of access, including contact lists, conversations, voice signatures, and times when someone is home. Any system that collects GPS data can recreate a detailed picture of someone's location and movement patterns.[49] Physical AI systems that provide critical capabilities, such as autonomous vehicles, could become targets for attacks that would put a person's safety at risk.[50] Finally, existing methods of de-identifying individuals from their personal data have been shown to be ineffective (although researchers are working on this challenge).[51] As organizations seek to collect more data for their algorithms, the rewards for stealing this information grow as well.

Lessons Learned from This AI Fail:

Hold AI to a Higher Standard • AI Challenges Are Multidisciplinary, so They Require a Multidisciplinary Team • Incorporate Privacy, Civil Liberties, and Security from the Beginning • Plan to Fail • Ask for Help: Hire a Villain • Use Math to Reduce Bad Outcomes Caused by Math • Make Our Assumptions Explicit • Monitor the AI's Impact and Establish Layers of Accountability • Require Objective, Third-party Verification and Validation

Return to Table of Contents

Fail #6. AI Pwned

Fail: Malicious actors can fool an AI or get it to reveal protected information.

"Pwned" is a computer-slang term that means "to own" or to completely get the better of an opponent or rival.[52]

Why is this a fail? Cyber-attacks that target AI systems are called "adversarial AI." AI may not have the defenses to prevent malicious actors from fooling the algorithm into doing what they want, or from interfering with the data on which the model trains, all without making any changes to the algorithm or gaining access to the code. At the most basic level, adversaries present lots of input to the AI and monitor what it does in response, so that they can track how the model makes very specific decisions. Adversaries can then very

slightly alter the input so that a human cannot tell the difference, but the AI has great confidence in its wrong conclusion.[55] Adversaries can also extract sensitive information about individual elements of the training sets[56] or adversaries can make assumptions about which data sources are used and then insert data to bias the learning process.[57]

What happens when things fail? The results can have serious real-world consequences. Researchers have demonstrated examples of a self-driving car not "seeing" a stop sign[58] and Google Home interpreting a greeting as a command to unlock the front door.[59] Researchers have also documented a hacker's ability to identify and decipher an individual's healthcare records from a published database of de-identified names.[60]

Examples:

Researchers created eyeglasses whose frames had a special pattern that defeats facial recognition algorithms by executing targeted (impersonation of another person) or untargeted (avoiding identification) attacks on the algorithms.[53] A human being would easily be able to identify the person correctly.

Researchers explored a commercial facial recognition system that used a picture of a face as input, searched its database, and outputted the name of the person with the closest matching face (and a confidence score in that match). Over time, the researchers discovered information about the individual faces the system had been trained on – information they should not have had access to. They then built their own AI system that, when supplied with a person's name, returned an imperfect image of the person, revealing data that had never been made public and should not have been.[54] This kind of attack illustrates that the sensitive information used for training an AI may not be as well protected as desired.

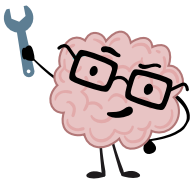
Pwning an AI is particularly powerful because 1) it is invisible to humans, so it is hard to detect; 2) it scales, so that a method to fool one AI can often trick other AIs; and 3) it works.[61]

Lessons Learned from This AI Fail:

Hold AI to a Higher Standard • AI Challenges Are Multidisciplinary, so They Require a Multidisciplinary Team • Incorporate Privacy, Civil Liberties, and Security from the Beginning • Involve the Communities Affected by the AI • Plan to Fail • Ask for Help: Hire a Villain • Use Math to Reduce Bad Outcomes Caused by Math • Make Our Assumptions Explicit • Monitor the AI's Impact and Establish Layers of Accountability • Require Objective, Third-party Verification and Validation

Return to Table of Contents

Turning Lemons into Reflux: When AI Makes Things Worse



Sometimes the biggest challenges emerge when AI does exactly what it is programmed to do! An AI doesn't recognize social contexts or constructs, and this section examines some of the unwanted impacts that can result from the divergence between technical and social outcomes. The three fails explore three components of the AI: the training data fed into the model, the objective of the AI and the metrics chosen to measure its success, and the AI's interactions with its environment.

Fail #7. Irrelevant Data, Irresponsible Outcomes

Fail: A lack of understanding about the training data, its properties, or the conditions under which the data was collected can result in flawed outcomes for the AI application.

Why is this a fail? Many AI approaches reflect the patterns in the data they are fed. Unfortunately, data can be inaccurate, incomplete, unavailable, outdated, irrelevant, or systematically problematic. Even relevant and accurate data may be unrepresentative and unsuitable for the new AI task. Since data is highly contextual, the original purposes for collecting the data may be unknown or not appropriate to the new task, and/or the data may reflect historical and societal imbalances and prejudices that are now deemed illegal or harmful to segments of society.[67]

What happens when things fail? When an AI system is trained on data with flawed patterns, the system doesn't just replicate them, it can encode and amplify them.[68] Without qualitative

Examples:

In 2008, early webcam facial tracking algorithms could not identify faces of darker skinned individuals because most of the training data (and the developers) were white skinned.[62] One particularly illuminating demonstration of this fail occurred in 2018, when Amazon's facial recognition system confused pictures of 28 members of Congress (the majority of them dark-skinned) with mugshots.[63] The ten-year persistence of these fails highlights the systemic and cultural barriers to fixing the problem, despite it being well acknowledged.

40,000 Michigan residents were wrongly accused of fraud by a state-operated computer system that had an error rate as high as 93%. Why? The system could not convert some data from legacy sources, and documentation and records were missing, meaning the system often issued a fraud determination without having access to all the information it needed. A lack of human supervision meant the problem was not addressed for over a year, but that wouldn't change the underlying problem that the data may not be usable for this application.[64]

An AI for allocating healthcare services offered more care to white patients than to equally sick black patients. Why? The AI was trained on real data patterns, where unequal access to care means less money is traditionally spent on black patients than white patients with the same level of need. Since the AI's goal was to drive down costs, it focused on the more expensive group, and therefore offered more care to white patients.[65,66] This example shows the danger of relying on existing data with a history of systemic injustice, as well as the importance of selecting between a mathematical and a human-centric measure to promote the desired outcome.

and quantitative scientific methods to understand the data and how it was collected, the quality of data and its impacts are difficult to appreciate. Even when we apply these methods, data introduces unknown nuances and patterns (which are sometimes incorrectly grouped together with human influences and jointly categorized as 'biases') that are really hard to detect, let alone fix.[69,70]

> The ten-year persistence of these fails highlights the systemic and cultural barriers to fixing the problem, despite it being well acknowledged

Statistics can help us address some of these pitfalls, but we have to be careful to collect enough, and appropriate, statistical data. The larger issue is that statistics don't capture social and political contexts and histories. We must remember that these contexts and histories have too often resulted in comparatively greater harm to minority groups (gender, sexuality, race, ethnicity, religion, etc.).[71]

Documentation about the data, including why the data was collected, the method of collection, and how it was analyzed, goes a long way toward helping us understanding the data's impact.

Lessons Learned from This AI Fail:

Hold AI to a Higher Standard • It's OK to Say No to Automation • AI Challenges Are Multidisciplinary, so They Require a Multidisciplinary Team • Involve the Communities Affected by the AI • Use Math to Reduce Bad Outcomes Caused by Math • Make Our Assumptions Explicit • Offer the User Choices • Monitor the AI's Impact and Establish Layers of Accountability • Envision Safeguards for AI Advocates • Require Objective, Third-party Verification and Validation

Return to Table of Contents

Fail #8. You Told Me to Do This

Fail: An AI will do what we program it to do. But how it does so may differ from what users want, especially if we don't consider social and contextual factors when developing the application.

Why is this a fail? Even if an AI has perfectly relevant and representative data to learn from, the way the AI tries to perform its job can lead to actions we didn't want or anticipate. We give the AI a specific task and a mathematical way to measure progress (sometimes called the "objective function" and "error function," respectively). Being human, we make assumptions about how the algorithm will perform its task, but all the algorithm does is find a mathematically valid solution, even if that solution goes against the spirit of what we intended (the literature calls this "reward hacking"). Unexpected results are more common in: complicated

systems, in applications that operate over longer periods of time, and in systems that have less human oversight.[72]

What happens when things fail? The AI doesn't recognize social context or constructs; it doesn't appreciate that some solutions go against the spirit of the rules. Therefore, the data and the algorithms aren't 'biased,' but the way the data interacts with our programmed goals can lead to biased outcomes. As designers, we set those objectives and ways of measuring success, which effectively incorporate what we value and why (consciously or unconsciously) into the AI.[77]

Examples:

An AI trained to identify cancerous skin lesions in images was successful, not because the AI learned to distinguish the shapes and colors of cancerous lesions from those of non-cancerous features, but because only the images of cancerous lesions contained rulers and the AI based its decision on the presence or absence of rulers in the photos.[73] This example shows the importance of understanding the key parameters an AI uses to make a decision, and illustrates how we may incorrectly assume that an AI makes decisions just as a human would.

An algorithm designed to win at Tetris chose to pause the game indefinitely right before the next piece would cause it to lose.[74] This example shows how an AI will mathematically satisfy its objective but fail to achieve the intended goals, and that the "spirit" of the rules is a human constraint that may not apply to the AI.

Open AI created a text-generating AI (i.e., an application that can write text all on its own) whose output was indistinguishable from text written by humans. The organization decided to withhold full details of the original model since it was so convincing that malicious actors could direct it to generate propaganda and hate speech.[75,76] This example shows how a well-performing algorithm does not inherently incorporate moral restrictions; adding that awareness would be the responsibility of the original developers or deployers.

Take the AI out of it for a moment, and just think about agreeing on a definition for a word. How would you define "fair"? (Arvind Narayanan, an associate professor of computer science at Princeton, defined "fairness" 21 different ways.)[78] For example, for college admissions that make use of SAT scores, a reasonable expectation of fairness would be that two candidates with the same score should have an equal chance of being admitted – this approach relies on "individual fairness." Yet, for a variety of socio-cultural reasons, students with more access to resources perform better on the test (in fact, the organization that creates the SAT recognized this in 2019 and began providing contextual information about the test taker's "neighborhood" and high school).[79] Therefore, another reasonable expectation of fairness would be that it takes into account demographic differences – this approach relies on "group fairness." Thus, a potential tension exists between two laudable goals: individual fairness and group fairness.

If we want algorithms to be 'fair' or 'accurate,' we have to agree on how to best scope these terms mathematically and socially. This means being aware of encoding one interpretation of the problem or preference for an outcome at the expense of the considerations of others. Therefore, we need to create frameworks and guidelines for when to apply specific AI applications, and weigh when the potential negative impacts of an AI outweigh the benefits of implementing it.

Lessons Learned from This AI Fail:

Hold AI to a Higher Standard • It's OK to Say No to Automation • AI Challenges Are Multidisciplinary, so They Require a Multidisciplinary Team • Ask for Help: Hire a Villain • Use Math to Reduce Bad Outcomes Caused by Math • Make Our Assumptions Explicit • Offer the User Choices • Monitor the AI's Impact and Establish Layers of Accountability • Envision Safeguards for AI Advocates • Require Objective, Third-party Verification and Validation • Entrust Sector-specific Agencies to Establish AI Standards for Their Domains

Return to Table of Contents

Fail #9. Feeding the Feedback Loop

Fail: When an AI's prediction is geared towards assisting humans, how a user responds can influence the AI's next prediction. Those new outputs can, in turn, impact user behavior, creating a cycle that pushes towards a single end. The scale of AI magnifies the impact of this feedback loop: if an AI provides thousands of users with predictions, then all those people can be pushed toward increasingly specialized or extreme behaviors.

Why is this a fail? The scale of AI deployment can result in substantial disruption and rewiring of everyday lives. Worse, people sometimes change their perceptions and beliefs to be more in line with an algorithm, rather than the other way around.[83,84]

The enormous extent of the problem makes fixing it much harder. Even recognizing problems is harder, since the patterns are revealed through collective harms and are challenging to discover by connecting individual cases.[85]

What happens when things fail? Decisions that seem harmless and unimportant individually, when collectively scaled, can build to become at odds with public policies, financial outcomes, and even public health. Recommender systems for social media sites choose

Examples:

If you're driving in Leonia, NJ, and you don't have a yellow tag hanging from your mirror, expect a $200 fine. Why? Navigation apps have redirected cars onto quiet, residential neighborhoods, where the infrastructure is not set up to support that traffic. Because the town could not change the algorithm, it tried to fight the outcomes, one car at a time.[80]

Predictive policing AI directs officers to concentrate on certain locations. This increased scrutiny leads to more crime reports for that area. Since the AI uses the number of crime reports as a factor in its decision making, this process reinforces the AI's decisions to send more and more resources to a single location and overlook the rest.[81] This feedback loop becomes increasingly hard to break.

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 most people's attention. Widespread use of recommendation engines with similar objectives can bring fringe content – like conspiracy theories and extreme violence – into the mainstream.[82]

incendiary or fake articles for newsfeeds,[86] health insurance companies decide which normal behaviors are deemed risky based off recommendations from AI,[87] and governments allocate social services according to AIs that consider only one set of factors.[88]

Concerns over the extent of the feedback loops AI can cause have increased. One government organization has warned that this behavior has the potential to contradict the very principles of pluralism and diversity of ideas that are foundational to Western democracy and capitalism.[89]

Lessons Learned from This AI Fail:

Hold AI to a Higher Standard • It's OK to Say No to Automation • AI Challenges Are Multidisciplinary, so They Require a Multidisciplinary Team • Incorporate Privacy, Civil Liberties, and Security from the Beginning • Ask for Help: Hire a Villain • Use Math to Reduce Bad Outcomes Caused by Math • Make Our Assumptions Explicit • Offer the User Choices • Monitor the AI's Impact and Establish Layers of Accountability • Envision Safeguards for AI Advocates • Require Objective, Third-party Verification and Validation • Entrust Sector-specific Agencies to Establish AI Standards for Their Domains

Return to Table of Contents

Fail #10. A Special Case: AI Arms Race

Even in the 1950s, Hollywood imagined that computers might launch a war. While today the general population is (mostly) confident that AI won't be directly tied to the nuclear launch button, just the potential of AI in military-capable applications is escalating global tensions, without a counteracting, cautionary force.[90] The RAND Corporation, a nonprofit institution that analyzes US policy and decision making, describes the race to develop AI as sowing distrust among nuclear powers. Information about adversaries' capabilities is imperfect, and the speed at which AI-based attacks could happen means that humans have less contextual information for response and may fear losing the ability to retaliate. Since there is such an advantage to a first strike, humans, not AIs, may be more likely to launch preemptively.[91] Finally, the perception of a race may prompt the deployment of less-than-fully tested AI systems.[92]

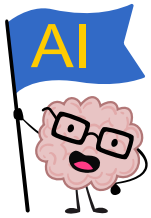
Lessons Learned from This AI Fail:

Hold AI to a Higher Standard • It's OK to Say No to Automation • AI Challenges Are Multidisciplinary, so They Require a Multidisciplinary Team • Incorporate Privacy, Civil Liberties, and Security from the Beginning • Involve the Communities Affected by the AI • Plan to Fail • Promote Better Adoption through Gameplay • Monitor the AI's Impact and Establish Layers of Accountability • Envision Safeguards for AI Advocates •

Require Objective, Third-party Verification and Validation • Entrust Sector-specific Agencies to Establish AI Standards for Their Domains

Return to Table of Contents

We're Not Done Yet: After Developing the AI



Developing AI is a dynamic, multifaceted process. Even if an AI performs optimally from a technical standpoint, other constraining factors could limit its overall performance and acceptance. Developing an AI to be safe and dependable means stakeholders must learn more about how the AI functions as the risks from its use increase. This section details factors that make that understanding challenging to achieve, and describes how proper documentation, explanations of intent, and user education can improve outcomes.

Fail #11. Testing in the Wild

Fail: Test and evaluation (T&E) teams work with algorithm developers to outline criteria for quality control, and of course they can't anticipate all algorithmic outcomes. But the consequences (and even blame) for the unexpected results are sometimes transferred onto groups who are unaware of these limitations or have not consented to being test subjects.

Why is this a fail? T&E of AI algorithms is hard. Even for AI models that aren't entirely black boxes we have only limited T&E tools[96,97] (though resources are emerging[98,99,100]). Difficulties for T&E result from:

Uncertain outcomes: Many AI models are complex, not fully explainable, and potentially non-linear (meaning they behave in unexpected ways in response to unexpected inputs), and we don't have great tools to help us understand their decisions and limitations.[101,102,103]

Model drift: Due to changes in data, the environment, or people's behavior an AI's

Examples:

Boeing initially blamed foreign pilots for the 737 MAX crashes, even though a sensor malfunction, faulty software, lack of pilot training, making a safety feature an optional purchase, and not mentioning the software in the pilot manual were all contributory causes.[93]

In 2014, UK immigration rules required some foreigners to pass an English proficiency test. A voice recognition system was used as part of the exam to detect fraud (e.g., if an applicant took the test multiple times under different names, or if a native speaker took the oral test posing as the applicant). But because the government did not understand how high the algorithm's error rate was, and each flagged recording was checked by undertrained employees, the UK cancelled thousands of visas and deported people in error.[94,95] Thus, applicants who had followed the rules suffered the consequences of the shortcomings in the algorithm.

performance will drift, or become outdated, over time.[104,105]

Unanticipated use: Because AI interacts with people who probably do not share our skills or understanding of the system, and who may not share our goals, the AI will be used in unanticipated ways.

Pressures to move quickly: There is a tension between resolving to develop and deploy automated products quickly and taking time to test, understand, and address the limitations of those products.[106]

Because all these difficulties, deployers and consumers of AI models often don't know the range or severity of consequences of the AI's application.[107]

Jonathan Zittrain, Harvard Law School professor, describes how the issues that emerge from an unpredictable system will become problematic as the number of systems increases. He introduces the concept of "intellectual debt," which applies to many fields, not only AI. For example, in medicine some drugs are approved for wide use even when "no one knows exactly how they work,"[108] but they may still have value. If the unknowns were limited to only a single AI (or drug), then causes and effects might be isolated and mitigated. But as the number of AIs and their interactions with humans grows, performing the number of tests required to uncover potential consequences becomes logistically impossible.

What happens when things fail? Users are held responsible for bad AI outcomes even if those outcomes aren't entirely (or at all) their fault. A lack of laws defining accountability and responsibility for AI means that it is too easy to blame the AI victim when something goes wrong. The default assumption in semi-autonomous vehicle crashes, as in the Boeing 737 MAX tragedies, has been that drivers are solely at fault.[109,110,111,112,113] Similarly, reports on the 737 crashes showed that "all the risk [was 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."[114] The early days of automated flying demonstrated that educating pilots about the automation capabilities and how to act as a member of a human-machine team reduced the number of crashes significantly.[115,116,117]

As a separate concern, the individuals or communities subject to an AI can become unwilling or unknowing test subjects. Pedestrians can unknowingly be injured by still-learning, semi-autonomous vehicles;[118] oncology patients can be diagnosed by an experimental IBM Watson (Watson is in a trial phase and not yet approved for clinical use);[119] Pearson can offer different messaging to different students as an experiment in gauging student engagement.[120] As the AI Now Institute at New York University (a research institute dedicated to understanding the social implications of AI technologies) puts it, "this is a repeated pattern when market dominance and profits are valued over safety, transparency, and assurance."[121]

The early days of automated flying demonstrated that educating pilots about the automation capabilities and how to act as a member of a human-machine team reduced the number of crashes significantly

Lessons Learned from This AI Fail:

AI Challenges Are Multidisciplinary, so They Require a Multidisciplinary Team • Incorporate Privacy, Civil Liberties, and Security from the Beginning • Involve the Communities Affected by the AI • Plan to Fail • Ask for Help: Hire a Villain • Use Math to Reduce Bad Outcomes Caused by Math • Make Our Assumptions Explicit • Try Human-AI Couples Counseling • Offer the User Choices • Promote Better Adoption through Gameplay • Monitor the AI's Impact and Establish Layers of Accountability • Envision Safeguards for AI Advocates • Require Objective, Third-party Verification and Validation • Entrust Sector-specific Agencies to Establish AI Standards for Their Domains

Return to Table of Contents

Fail #12. Government Dependence on Black Box Vendors

Fail: Trade secrecy and proprietary products make it challenging to verify and validate the relevance and accuracy of vendors' algorithms.

These examples demonstrate the importance of at least knowing the attributes of the data and processes for creating the AI model.

Why is this a fail? For government organizations, it's cheaper or easier to acquire algorithms from or outsource algorithm development to third-party vendors. To verify and validate the delivered technology, the government agency needs to understand the methodology that produced it: from analyzing what datasets were applied to knowing the objectives of the AI model to ensuring the operational environment was captured correctly.

Examples:

COMPAS, a tool that assesses recidivism risk of prison inmates (repeating or returning to criminal behavior), produced controversial results. In one case, because of an error in the data fed into the AI, an inmate was denied parole despite having a nearly perfect record of rehabilitation. Since COMPAS is proprietary, neither judges nor inmates know how the tool makes its decisions.[122,123]

The Houston Independent School District implemented an AI to measure teachers' performances by comparing their student's test scores to the statewide average. The teacher's union won a lawsuit, arguing that the proprietary nature of the product prevents teachers from verifying the results, thereby violating their Fourteenth Amendment rights to due process.[124]

What happens when things fail? Often the problems with the vendors' models come about because the models' proprietary nature inhibits verification and validation capabilities. For example, if the vendor modified or added to the training data that the government supplied for the algorithm, or if the government's datasets and operating environment have evolved from those provided to the vendor, then the AI won't perform as expected. Unless the contract says otherwise, the vendor keeps its training and validation processes private.

In certain cases the government agency doesn't have a mature enough understanding of AI requirements and acquisition to prevent mistakes. Sometimes a government agency doesn't buy a product, but it buys a service. For example, since government agencies usually don't have fully AI-capable workforces, an agency might provide its data to the vendor with the expectation that the vendor's experts might discover patterns in the data. In some of these instances, agencies have forgotten to keep some data to serve as a test set, since the same data cannot be used for training and testing the product.

These verification and validation challenges will become more important, yet harder to overcome, as vendors begin to pitch end-to-end AI platforms rather than specialized AI models.

Lessons Learned from This AI Fail:

Hold AI to a Higher Standard • It's OK to Say No to Automation • AI Challenges Are Multidisciplinary, so They Require a Multidisciplinary Team • Involve the Communities Affected by the AI • Plan to Fail • Ask for Help: Hire a Villain • Make Our Assumptions Explicit • Monitor the AI's Impact and Establish Layers of Accountability • Envision Safeguards for AI Advocates • Require Objective, Third-party Verification and Validation • Entrust Sector-specific Agencies to Establish AI Standards for Their Domains

Return to Table of Contents

Fail #13. Clear as Mud

Fail: The technical and operational challenges in creating a perfectly understandable model can dissuade developers from including incomplete, but still helpful, context and explanations. This omission can prevent people from using an otherwise beneficial AI.

Why is this a fail? When we introduce an AI into a new system or process, each set of stakeholders – AI developers, operators, decision makers, affected communities, and objective third-party evaluators – has different requirements for understanding, using, and trusting the AI system.[125] These requirements are also domain and situation specific.[126]

Especially as we begin to develop and adopt AI products that enhance or substitute for human judgment, it is essential that users and policymakers know more about how an AI functions and the intended and non-intended uses for the AI. Adding explanations, documentation, and context are so important because they help calibrate trust in an AI – that is, *figuring out how to trust the AI to the extent it should be trusted*. Empowering users and stakeholders with understanding can address concepts such as:

- Transparency – how does the AI work and what are its decision criteria?

- Traceability – can the AI help developers and users follow and justify its decision-making process?
- Interpretability – can developers and users understand and make sense of any provided explanations?
- Informativeness – does the AI provide information that different stakeholders find useful?
- Policy – under what conditions is the AI used and how is it incorporated into existing processes or human decision making?
- Limitations – do the stakeholders understand the limits of the AI and its intended uses?[129,130,131]

Examples:

When UPS rolled out a route-optimization AI that told drivers the best route to take, drivers initially rejected it because they felt they knew better. Once UPS updated the system to provide explanations for some of its suggestions, the program had better success.[127]

A psychiatrist realized that Facebook's 'people you may know' algorithm was recommending her patients to each other as potential 'friends,' since they were all visiting the same location.[128] Explanations to both users and developers as to why this algorithm made its recommendations could have mitigated similar breaches of privacy and removed those results from the output.

Traditionally, the conversation in the AI community has focused on transparency (AI experts refer to it as "explainability" or "explainable AI"). Approaches for generating AI explanations are very active areas of research, but coming up with useful explanations of how the model actually makes decisions remains challenging for several reasons. Technically, it can be hard because certain models are very complex. Current explainer tools can emphasize which inputs had the most influence on an answer, but not why they had that influence, which makes them valuable but incomplete. Finally, early research showed a tradeoff between accuracy and explainability, but this tradeoff may not always exist. Some of us have responded to the myth that there must be a tradeoff by overlooking more interpretable models in favor of more common but opaque ones.[132]

What happens when things fail? Cognitively, existing explanations can be misleading. Users can be tempted to impart their own associations or anthropomorphize an AI (i.e., attributing human intentions to it). Also, assuming causality when there is only correlation in an AI system will lead to incorrect conclusions.[133] If these misunderstandings can cause financial, psychological, physical, or other types of harm, then the importance of good explanations becomes even greater.[134]

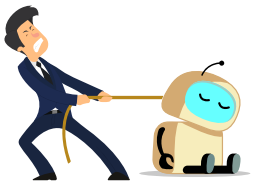
Adding explanations, documentation, and context are so important because they help calibrate trust in an AI – that is, figuring out how to trust the AI to the extent it should be trusted

The challenge lies in expanding the conversation beyond transparency and explainability to include the multitude of ways in which AI stakeholders can improve their understanding and choice. If we adopt the mindset that the users, policymakers, auditors, and others in the AI workflow are all our customers, this can help us devote more resources to providing the context that these stakeholders need.

Lessons Learned from This AI Fail:

Hold AI to a Higher Standard • AI Challenges Are Multidisciplinary, so They Require a Multidisciplinary Team • Incorporate Privacy, Civil Liberties, and Security from the Beginning • Involve the Communities Affected by the AI • Plan to Fail • Make Our Assumptions Explicit • Try Human-AI Couples Counseling • Offer the User Choices • Promote Better Adoption through Gameplay • Monitor the AI's Impact and Establish Layers of Accountability • Envision Safeguards for AI Advocates • Require Objective, Third-party Verification and Validation • Entrust Sector-specific Agencies to Establish AI Standards for Their Domains

Failure to Launch: How People Can React to AI



People often hold multiple, contradictory views at the same time. There are plenty of examples when it comes to human interaction with technology: people can be excited that Amazon or Netflix recommendations really reflect their tastes, yet worry about what that means for their privacy; they can use Siri and Google voice to help them remember things, yet lament about losing their short-term memory; they can rely on various newsfeeds to give them information, even if they know (or suspect) that the primary goal of the algorithms behind those newsfeeds is to keep their attention, not to deliver the broadest news coverage. These seeming dichotomies all revolve around trust, which involves belief and understanding, dependency and choice, perception and evidence, emotion and context. All of these elements of trust are critical to having someone accept and adopt an AI. When we as AI developers and deployers include technical, cultural, organizational, sociological, interpersonal, psychological, and neurological perspectives, we can more accurately align people's trust in the AI to the actual trustworthiness of the AI, and thereby facilitate how people adopt of the AI.

Fail #14. In AI We Overtrust

Fail: When people aren't familiar with AI, cognitive biases and external factors can prompt them to trust the AI more than they should. Even professionals can overtrust AIs deployed in their own fields. Worse, people can change their perceptions and beliefs to be more in line with an algorithm's, rather than the other way around.

Why is this a fail? When an AI is helping people do things better than they would on their own, it is easy to assume that the platform's goals mirror the user's goals. However, there is no such thing as a "neutral" AI.[135] During the design process we make conscious and unconscious assumptions about what the AI's goals and priorities should be and what data streams the AI should learn from. Lots of times, our incentives and user incentives align, so this works out wonderfully: users drive to their destinations, or they enjoy the AI-recommended movie. But when goals don't align, most users don't realize that they're potentially acting against their interests. They are convinced that they're making rational and objective decisions, because they are listening to a rational and objective AI.[136]

Furthermore, how users actually act and how they think they'll act often differs. For example, a journalist documented eight drivers in 2013 who overrode their own intuition and blindly followed their GPS, including drivers who turned onto the stairs of the entrance to a park, a driver who drove into a body of water, and another driver who ran straight into a house, all because of their interpretation of the GPS instructions.[141]

Numerous cognitive biases can contribute to overtrusting technology. Research highlights three prevalent ones:

1. Humans can have a bias to assume automation is perfect; therefore, they have high initial trust.[142] This "automation bias" leads users to trust automated and decision support systems even when it is unwarranted.

2. Similarly, people generally believe something is true if it comes from an authority or expert, even if no supporting evidence is supplied.[143] In this case, the AI is perceived as the expert.

Examples:

A research team put 42 test participants into a fire emergency scenario featuring a robot responsible for escorting them to an emergency exit. Even though the robot passed obvious exits and got lost, 37 participants continued to follow it.[137,138]

Consumers who received a digital ad said they were more interested in a product that was specifically targeted for them, and even adjusted their own preferences to align with what the ad suggested about them.[139]

In a research experiment, students were told that a robot would determine who had pushed a button and "buzzed in" first, thus winning a game. In reality, the robot tried to maximize participant engagement by evenly distributing who won. Even as the robot made noticeably inaccurate choices, the participants did not attribute the discrepancy to the robot having ulterior motives.[140]

3. Lastly, humans use mental short-cuts to make sense of complex information, which can lead to overtrusting an AI if it behaves in a way that conforms to our expectations, or if we have an unclear understanding of how the AI works. Cathy O'Neil, mathematician and author, writes that our relationship to data is similar to an ultimate belief in God: "I think it has a few hallmarks of worship – we turn off parts of our brain, we somehow feel like it's not our duty, not our right to question this."[144]

Therefore, the more an AI is associated with a supposedly flawless, data-driven authority, the more likely that humans will overtrust the AI. In these conditions, even professionals in a given field can cede their authority despite their specialized knowledge.[145,146]

Another outcome of overtrust is that the AI reinforces a tendency to align with the model's solution rather than the individual's own, pushing AI predictions to become self-fulfilling.[147] These outcomes also show that having a human supervise an AI will not necessarily work as a failsafe.

What happens when things fail? The phenomenon of overtrust in AI has contributed to two powerful and potentially frightening outcomes. First, since AIs often have a single objective and reinforce increasingly specialized ends, users aren't presented with alternative perspectives and are directed toward more individualistic, non-inclusive ways of thinking.

Second, the pseudo-authority of AI has allowed pseudosciences to re-emerge with a veneer of validity. Demonstrably invalid examples of AI have been used to look at a person's face and assess that person's tendencies toward criminality or violence,[148,149] current feelings,[150] sexual orientation,[151] and IQ or personality traits.[152] These phrenology and physiognomy products and claims are unethical, irresponsible, and dangerous.

Although these outcomes may seem extreme, overtrust has a wide range of consequences, from causing people to act against self-interest to promulgating discriminatory practices.

Lessons Learned from This AI Fail:

Hold AI to a Higher Standard • It's OK to Say No to Automation • AI Challenges Are Multidisciplinary, so They Require a Multidisciplinary Team • Incorporate Privacy, Civil Liberties, and Security from the Beginning • Involve the Communities Affected by the AI • Make Our Assumptions Explicit • Try Human-AI Couples Counseling • Offer the User Choices • Promote Better Adoption through Gameplay • Monitor the AI's Impact and Establish Layers of Accountability • Envision Safeguards for AI Advocates • Require Objective, Third-party Verification and Validation • Entrust Sector-specific Agencies to Establish AI Standards for Their Domains

Return to Table of Contents

Fail #15. Lost in Translation: Automation Surprise

Fail: End-users can be surprised by how an AI acts, or that it failed to act when expected.

Why is this a fail? When automated system behaviors cause users to ask, "What's it doing now?" or "What's it going to do next?" the literature calls this *automation surprise*.[153] These behaviors leave users unable to predict how an automated system will act, even if it is working properly. Surprise can occur when the system is too complicated to understand, when we make erroneous assumptions about the environment in which the system will be used, or when people simply expect automated systems to act the same way they do.[154] AI can exacerbate automation surprise because its decisions evolve and change over time.

What happens when things fail? The more transparent we are about what the AI can and cannot do (which isn't always possible because sometimes even we don't know), the better we can educate users of that system about how it will or will not act. Human-machine teaming (HMT) principles help us understand the importance of good communication. When an AI is designed to help the human partner understand what the automation

will do next, the human partner can anticipate those actions and act in concert with them, or override or tweak the automation if needed.[158,159,160]

Without this context and awareness, the human partner may become frustrated and stop using the AI. Alternatively, the human partner may be unprepared for the AI action and be unable to recover from a bad decision.

Lessons Learned from This AI Fail:

Hold AI to a Higher Standard • It's OK to Say No to Automation • AI Challenges Are Multidisciplinary, so They Require a Multidisciplinary Team • Involve the Communities Affected by the AI • Make Our Assumptions Explicit • Try Human-AI Couples Counseling • Offer the User Choices • Promote Better Adoption through Gameplay • Monitor the AI's Impact and Establish Layers of Accountability • Envision Safeguards for AI Advocates • Require Objective, Third-party Verification and Validation • Entrust Sector-specific Agencies to Establish AI Standards for Their Domains

Examples:

When drivers take their hands off the wheel in modern cars, they can make dangerous assumptions about the car's automated capabilities and who or what is in control of what part of the vehicle.[155] This example illustrates the importance of providing training and time for the general population to familiarize themselves with a new automated technology.

An investigation of a 2012 airplane near-crash (Tel Aviv – Airbus A320) revealed "significant issues with crew understanding of automation… and highlighted the inadequate provision by the aircraft operator of both procedures and pilot training for this type of approach."[156] This example shows how even professionals in a field need training when a new, automated system is introduced.

Facebook trained AIs through unsupervised learning (without human supervision) to learn how to negotiate. The "Bob" and "Alice" chatbots started talking to each other in their own, made-up language, which was unintelligible to humans.[157] This example shows that even AI experts can be completely surprised by an AI's outcome.

Return to Table of Contents

Fail #16. The AI Resistance

Fail: Not everyone wants AI or believes that its benefits outweigh the costs. If we dismiss the cautious as Luddites, the technology can genuinely victimize the people who use it.

"Luddite" is a term describing the 19th century English workmen who vandalized the labor-saving machinery that took their jobs. The term has since been extended to refer to one who is opposed to technological change.[161]

Why is this a fail? The reluctance to adopt AI without reservation is warranted. Just a few years ago, the AI developer community saw the increase in AI capabilities as unadulterated progress and good. Recently, we're learning that sometimes this holds true, and sometimes progress means progress only for some – that AI can have harmful impacts on users, communities, and employees of our AI companies.[164,165]

Examples:

When Waymo decided to test self-driving cars in a town in Arizona without first seeking the residents' approval, residents feared losing their jobs and their lives. Feeling they had no other options open to them, they threw rocks at the automated cars and slashed their tires as means of protest.[162]

Cambridge Analytica used AI to surreptitiously influence voters through false information that was individually targeted. Public officials, privacy specialists, and investigative journalists channeled feelings of outrage, betrayal, confusion, and distrust into increased pressure to strengthen legislative protection.[163]

What happens when things fail? Even those who are "early adopters" or an "early majority" in the technology adoption lifecycle[166] may still have reservations about fully integrating the new technology into their lives. The people who reject AI entirely may have concerns that cannot be addressed by time, education, and training. For instance, some people find the automated email replies that mimic individual personalities creepy,[167] some people are worried about the national security implications caused by deepfakes,[168] some decry the mishandling of the private data that drives AI platforms,[169] some fear losing their jobs to AI,[170] some protest the disproportionate impact of mass surveillance on minority groups,[171,172,173] and some fear losing their lives to an AI-driven vehicle.[174]

Anger, frustration, and resistance to AI are natural reactions to a society that seems to assume that technology adoption is inevitable and disruptive to their safety or way of life. The idea that the believers should just wait out the laggards and Luddites − or worse, treat them as the problem – is flawed. Therefore, we should listen to their concerns and bring in the resisters to guide the solution.

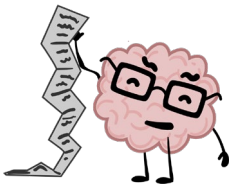
> Sometimes progress means progress only for some – that AI can have harmful impacts on users, communities, and employees of our AI companies

Lessons Learned from This AI Fail:

Hold AI to a Higher Standard • It's OK to Say No to Automation • AI Challenges Are Multidisciplinary, so They Require a Multidisciplinary Team • Incorporate Privacy, Civil Liberties, and Security from the Beginning • Involve the Communities Affected by the AI • Plan to Fail • Make Our Assumptions Explicit • Try Human-AI Couples Counseling • Offer the User Choices • Monitor the AI's Impact and Establish Layers of Accountability • Envision Safeguards for AI Advocates • Require Objective, Third-party Verification and Validation • Entrust Sector-specific Agencies to Establish AI Standards for Their Domains

Return to Table of Contents

AI Registry: The Things We'll Need That Support AI



AI isn't just about the data and algorithms. To be successful, we as developers and deployers depend on a whole line of supporting elements. This section addresses some, but not all, of those elements, including the right governing policies, the right people, the right data, and the right equipment.

Fail #17. Good (Grief!) Governance

Fail: We sometimes implement AI without a detailed strategy for how it will be governed, and there aren't any laws that ensure oversight and accountability. In that vacuum, the technology itself is redefining cultural and societal norms.

Why is this a fail? AI has reached a state of maturity where governance is a necessary, yet difficult, element. AI systems continue to be increasingly integrated into daily life, but this occurs without adequate governance, oversight, or accountability. This happens in part because:

1. AI is a probabilistic and dynamic process, meaning AI outcomes will not be fully replicable, consistent, and predictable. Therefore, new governance mechanisms must be developed.

2. Organizations allocate money to buy products, but often do not add funds for creating and testing internal governance policies. Therefore, those policies may not be introduced until the effects of the technology's use have had an impact on people's lives.

3. Government and private organizations sometimes keep policies that govern AI use and development hidden from the public in order to protect national security interests or trade secrets.[177]

4. There are no mature AI laws, standards or norms that apply across multiple domains, and laws within a specific domain are only now emerging. Therefore, standardizing policies or sharing best practices face additional obstacles.

Examples:

Police departments can purchase crime prediction products that estimate where crimes will occur or who will be involved. Many of the products are "black boxes," meaning it is not clear how decisions are made, and many police departments deploy them in the absence of clear or publicly available policies to guide how they should be applied.[175] Often a new technology is acquired and used first, while policy and governance for its use are developed later.

Employees of a contractor working for Google paid dark-skinned, homeless people $5 for letting the contractor take a picture of their faces in order to make its training dataset more diverse.[176] In addition, these workers may have misled the homeless about the purpose of their participation. Without comprehensive legislation about data collection and privacy infringement, ending such questionable practices becomes the responsibility of the governance policies of each company.

The result is that in the United States there are few clear governance models for industry or government to replicate, and there are limited legal authorities that specify whom to hold accountable when things go wrong.[178,179]

Did You Know?

Black Box Processes

Employing opaque AI systems or governance policies allows organizations to more easily act in hidden or non-transparent ways. Predictive policing AI systems, which suggest individuals or areas that police should focus on when fighting crime, provide a thoroughly studied example. Even though the AI tools have been in operation for several years, the public isn't given information about how the tools work, how police departments use the tools, or what the police themselves know about the technology and the policy.[180,181] Upturn, a non-profit organization promoting technology and advancing justice, wrote a report in 2016 on predictive policing and civil rights which points out *that AI prediction tools shape the roles of the police department.* Since the AI's data comprises short-term actions like generating citations and arrests, it pushes police resources towards collecting the same type of data based on short-term actions, rather than toward broader community integration and protection. The report also points out that "police hesitate to use predictive technology to analyze their own performance, even… [for] counseling and training… turning these approaches inward can be a lower stakes way to apply data driven insights in the policing context."[182]

What happens when things fail? In response to unclear legal accountabilities, organizations have embraced declarations of ethical principles and frameworks that promote responsible AI development.[183] These statements vary in detail and specificity, but almost all declare principles of transparency, non-discrimination, accountability, and safety. These current approaches represent important steps, but evidence shows that they are not enough. They are almost universally voluntary commitments, and few of the declarations include recommendations, specifics, or use cases for how to make the principles actionable and implementable (though in the largest AI companies, these are being developed).[184] Finally, researchers have shown that pledges to uphold ethical principles do not guarantee ethical behavior.[185]

Without proper governance, and legal accountability and oversight, the technology becomes the de-facto norm. Therefore, we must recognize that because we control the code, we may unintentionally become de-facto decision makers.

In parallel with private efforts, the US government is beginning to define guidance, but it is still in early stages. In January 2020, the White House published draft principles for guiding federal regulatory and non-regulatory

approaches to AI,[186] and state governments are also getting more involved in regulation.[187] However, often state laws are contradictory or lag the technology. As of January 2020, several cities in California and Massachusetts have banned the use of facial recognition technology by public entities,[188] but at the same time other US cities, as well as airports and private entities, are increasing their adoption of the same technology.[189,190] Because this field of law is so new there are limited precedents.

Absent precedent, AI applications – or more accurately we, the developers –unintentionally create new norms. The dangers that we must keep in mind are that the AI can undermine traditional figures of authority and reshape the rule of law. Without proper governance, and legal accountability and oversight, the technology becomes the de-facto norm. Therefore, we must recognize that because we control the code, we may unintentionally become de-facto decision makers.[191]

Lessons Learned from This AI Fail:

Hold AI to a Higher Standard • It's OK to Say No to Automation • AI Challenges Are Multidisciplinary, so They Require a Multidisciplinary Team • Incorporate Privacy, Civil Liberties, and Security from the Beginning • Involve the Communities Affected by the AI • Plan to Fail • Make Our Assumptions Explicit • Offer the User Choices • Monitor the AI's Impact and Establish Layers of Accountability • Envision Safeguards for AI Advocates • Require Objective, Third-party Verification and Validation • Entrust Sector-specific Agencies to Establish AI Standards for Their Domains

Return to Table of Contents

Fail #18. Just Add (Technical) People

Fail: AI skills are in ever-higher demand, but employers erroneously believe that they only need to hire technical people (with backgrounds in computer science, engineering, mathematics, or related fields), even though developing successful and beneficial AI is not purely a technical challenge.

Why is this a fail? The small size of the AI workforce is often cited as the greatest barrier to AI adoption.[192] This same problem applies in other fields; for example, healthcare and cybersecurity have similar shortages of skilled technical workers. When responding to the immediate need for AI talent, companies rightly focus on hiring and training data scientists with expertise in AI algorithms, or other specialists in the fields of computer science, engineering, mathematics, and related technical areas. While these employees are absolutely necessary to develop and implement AI at a technical level, *just as necessary* are specialists from other fields who can balance and contextualize how AI is applied in that domain.

What happens when things fail? The healthcare and cyber fields are a couple of years ahead of AI when it comes to articulating the skills and abilities necessary for a fully representative workforce. Leaders in both fields recognize that the shortage of technical skills is one challenge, while creating multidisciplinary teams is another. For example, the US government developed a National Initiative for Cybersecurity Education (NICE) framework that "describes the interdisciplinary nature of the cybersecurity workforce [and]... describes cybersecurity work and workers irrespective of where or for whom the work is performed."[196] Healthcare organizations have long realized that meeting workforce needs involves more than just hiring doctors and have acted on evidence that interdisciplinary collaboration leads to better patient outcomes.[197,198,199]

In contrast, the companies and organizations that develop and deploy AI have not yet designed or agreed on similar AI workforce guidelines, though the US government does recognize the importance of interdisciplinary and inclusive teams in several AI strategy publications.[200,201] The next step is to move from recognition to implementation.

Examples:

IBM Watson produced "unsafe and incorrect" cancer treatment recommendations, including "recommendations that conflicted with national treatment guidelines and that physicians did not find useful for treating patients." Internal IBM documents reveal that training was based on only a few hypothetical cases and a few specialists' opinions. This finding suggests that including more doctors, hospital administrators, nurses, and patients early in the development process could have led to the use of proper diagnostic guidelines and training data.[193]

A crash between a US Navy destroyer and an oil tanker resulted from a navigation system interface that was poorly designed, overly complicated, and provided limited feedback.[194] Engineers and scientists who study how poor interfaces lead to mishaps can and have helped shape better interface design and safety processes.

In 2015, Google's automated photo-tagging software mislabeled images of dark-skinned people as "gorillas." Through 2018, Google's solution was to remove "gorilla" and the names of other, similar animals from the application's list of labels.[195] Hiring employees and managers trained in diverse disciplines, and not merely technical ones, could have resulted in alternative, more inclusive, outcomes.

Lessons Learned from This AI Fail:

AI Challenges Are Multidisciplinary, so They Require a Multidisciplinary Team • Incorporate Privacy, Civil Liberties, and Security from the Beginning • Involve the Communities Affected by the AI • Try Human-AI Couples Counseling • Monitor the AI's Impact and Establish Layers of Accountability • Envision Safeguards for AI Advocates • Entrust Sector-specific Agencies to Establish AI Standards for Their Domains

Return to Table of Contents

Fail #19. Square Data, Round Problem

Fail: Having data doesn't mean we have a solution: the right data for the problem is not always easily collectable, or in formats that are ingestible or comparable. What's more, we may not be able to collect data on all the factors that a given AI application must take into account for adequately understanding the problem space.

Why is this a fail? Some AI applications require large amounts of data to be effective. Fortuitously for the AI community, we are experiencing an explosion of data being generated (2.5 quintillion bytes a day, and growing[204]). But much of this data is not ready for exploitation. The data can be full of errors, leave gaps, or not be standardized, making its practical use challenging (as seen in the United Airlines example). As a result, a surprisingly high number of businesses (79%) are basing critical decisions on data that hasn't been properly verified.[205] On the other hand, valid and useful data can be incompatible across multiple similar applications, preventing an organization from creating a fuller picture (as seen in the DoD example).

Examples:

United Airlines lost $1B in revenue in 2016 from relying on a system that drew on inaccurate and limited data. United had built a software system to forecast demand for passenger seating, but the assumptions behind the data were so flawed and out of date that two-thirds of the system's outputs were not good enough for accurate projections.[202]

The Navy, Air Force, and Army all collect different information when they investigate why an aircraft crashes or has a problem, making it difficult for the Department of Defense (DoD) to compare trends or share lessons learned.[203]

What happens when things fail? The challenge for some of us then, is to understand that more data isn't a solution to every problem. Aside from concerns over accuracy, completeness, and historical patterns, not all factors can be captured by data. Some of the problem-spaces involved have complex, interrelated factors: for example, one study on community policing found that easy-to-collect data, like the number of crime reports and citations, was used for determining how to combat crime; yet this approach overlooks factors vital to correctly addressing the issues, such as identifying community problems, housing issues, and public health patterns.[206]

> Care must be taken to ensure that the obsession for [sic] effectiveness and predictability behind the use of algorithms does not lead to us designing legal rules and categories no longer on the grounds of our ideal of justice, but so that they are more readily 'codable'

The French Data Protection Authority (the government agency responsible for the protection of personal data) warns against ignoring a complex reality for the sake of results: "care must be taken to ensure that the obsession for [sic] effectiveness and predictability behind the use of algorithms does not lead to us designing legal rules and categories no longer on the grounds of our ideal of justice, but so that they are more readily 'codable.'"[207]

Lessons Learned from This AI Fail:

Hold AI to a Higher Standard • It's OK to Say No to Automation • Incorporate Privacy, Civil Liberties, and Security from the Beginning • Involve the Communities Affected by the AI • Use Math to Reduce Bad Outcomes Caused by Math • Make Our Assumptions Explicit • Promote Better Adoption through Gameplay • Monitor the AI's Impact and Establish Layers of Accountability • Envision Safeguards for AI Advocates • Require Objective, Third-party Verification and Validation

Return to Table of Contents

Fail #20. My 8-Track Still Works So What's the Issue?

Fail: Organizations often attempt to deploy AI without considering what hardware, computational resources, and information technology (IT) systems users actually have.

Why is this a fail? The latest processors have amazing computational power, and most AI companies can pay for virtual access to the fastest and most powerful machines in the cloud. Government agencies are often an exception: short-term budget priorities, long and costly acquisition cycles, and security requirements to host their own infrastructure in-house[210,211] have pushed the government towards maintaining and sustaining existing IT, rather than modernizing the technology.[212] Another exception is established commercial institutions with vital legacy infrastructure (for instance, 92 of the top 100 banks still use mainframe computers), which have such entrenched dependencies that updating IT can have costly and potentially disruptive effects on the business.[213]

Examples:

The Department of Defense still uses 8-inch floppy disks in a system that "coordinates the operational functions of the nation's nuclear forces."[208] Implementing advanced algorithms would be impossible on this hardware.

95% of ATM transactions still use COBOL, a 58-year-old programming language (numbers as of 2017), which raises concerns about maintaining critical software over the next generation of ATMs.[209]

What happens when things fail? Any group that depends on legacy systems finds it hard to make use of the latest AI offerings, and the technology gap continues to increase over time. While an organization's current IT may not be as obsolete as the examples here, any older infrastructure has more limited libraries and software packages, and less computational power and memory, than modern systems, and therefore may not meet the requirements of heavy AI processing. So, algorithms developed elsewhere may not be compatible with existing solutions and can't simply be ported to an older generation of technology.

Lessons Learned from This AI Fail:

It's OK to Say No to Automation • Plan to Fail • Make Our Assumptions Explicit • Monitor the AI's Impact and Establish Layers of Accountability • Require Objective, Third-party Verification and Validation

Return to Table of Contents

LESSONS LEARNED

Expand Early Project Considerations

Lesson #1. Hold AI to a Higher Standard

New technology applications, and the companies that develop those technologies, 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. How?

1. *An AI system replicates the social values of its developers and also embeds them into systems.* As developers and deployers, our choices, assumptions, simplifications, and trade-offs all shape the behavior of the system, and we can (intentionally or not) encode those values as the new standard. All too often those values represent how young, white, technically oriented, Western men interact with the world. We need to improve our outreach to and understanding of a far broader set of stakeholder communities.

2. *An AI system's reach can centralize power in the hands of a few*. If one person makes a decision or influences one other person's behavior, the effects are limited. But an AI allows us to aggregate and amplify our influence over many people's behaviors. Even an entirely automated decision is never neutral – outcomes always affect people differently. Therefore, we should explore how AI changes human behavior at scale, and apply what we learn to the AI we create.[214,215]

3. *People can be influenced to trust AI more than they should.* In certain conditions, people place more trust in an AI than is warranted, because they assume it is more impartial and infallible than they are. Individuals also have cognitive biases that lead them to treat connections and correlations as conclusions and inferences. Because AI can connect exponentially more information than a small group can on its own, it can magnify the effects of false or misleading conclusions. We should do our best to ensure that the trust people place in the AI is matched by a higher degree of trustworthiness.

4. *There is a tension between global pressures to develop and deploy AI quickly, and the need to understand and mitigate an AI's impacts.* When AI systems scale, or act so fast that humans cannot respond in time, then humans must rely on the guardrails and risk mitigation practices incorporated in the system. If these protections and practices are limited because developers focused on deploying AI as rapidly as possible, the chances for unwanted outcomes increase. Therefore, we need to ensure we integrate risk assessment and mitigation protections early in the AI's development and throughout the system's lifecycle.

5. *It is unclear who is accountable for an AI system's decisions.* As of today, legal responsibility for the consequences of AI system use has not been established, and this results in a lack of accountability or in holding the wrong person accountable.[216] When no one is considered legally accountable if something goes wrong, and no one is made responsible for fixing it, the consequences of mistakes and misuse can

easily lead to abuse of privacy and civil rights. We have to exercise particular care to reach out to those who best understand the domain and risks, and be more inclusive in our design teams as a way to prevent bad outcomes to the extent possible.

We are in the best position to recognize the potential impacts of this technology. If we hold AI to a higher standard, our example has the potential to raise the standard across the board. If we establish rigorous practices for quality control and assurance within our organizations, then other AI vendors will feel pressure to match the evolved market expectations. When companies and the government set standards for workforce training, AI team composition, and governance practices, those standards become a baseline for a common lexicon, curricula in universities, and expectations across the public, private, and academic sectors.[217]

The rest of the lessons learned provide more detail on specific aspects of ensuring proper use of AI and offer actionable implementation guidance.

Lesson #2. It's OK to Say No to Automation

The first things we should ask when starting an AI project is simply, "Is this actually a problem that we need AI to address? Can AI even be effective for this purpose?" Our end goal is really to meet stakeholder needs, independent of the particular technology or approach we choose.[218]

Sometimes, automation is simply not the right choice. As a general rule, the more the outcome should depend on human judgment, the more "artificial" an AI solution is. Some more guidelines follow:

As a general rule, the more the outcome should depend on human judgment, the more "artificial" an AI solution is

- Our AI systems should incorporate more human judgment and teaming as applications and environments become more complex or dynamic.
- We should enlist human scrutiny to ensure that the data we use is relevant and representative of our purposes, and that there is no historical pattern of bias and discrimination in the data and application domain.
- If the risk of using the data or the purpose of the AI could cause financial, psychological, physical, or other types of harm, then we must ask whether we should create or deploy the AI at all.[219]

Applying AI more selectively will help stakeholders accept that those AI solutions are appropriate. Distinguishing which challenges would benefit from AI and which challenges do not lend themselves to AI, gives customers and the public more confidence that AI is deployed responsibly, justifiably, and in consideration of existing norms and public safety.

Lesson #3. AI Challenges Are Multidisciplinary, so They Require a Multidisciplinary Team

The challenges to overcome when developing or implementing AI are diverse and can be both technical and social in nature. As a result, no one person or discipline can singlehandedly "fix" AI. Those of us on the front lines of building the AI share many attributes (i.e., similar education and degrees, life experiences, and cultural backgrounds). [220] If we do not actively work to incorporate other valid perspectives into the development process, we risk having the AI reflect our assumptions about how the product will be used and by whom, instead of being based on research evidence and empirical data.

Therefore, our development teams need members with diverse demographic and professional backgrounds. Examples of members of a well-rounded team include:

- Data engineers to ensure that data is usable and relevant
- Model developers to help the AI achieve the project's objectives
- Strategic decision makers who understand the technical aspects of AI as well as broader strategic issues or business needs
- Domain specialists to supply context about how people in their field actually behave, existing business practices, and any historical biases. Domain experts can be scientific or non-scientific; they may be military personnel, teachers, doctors and patients, artists … any people who are actual experts in the area for which the AI is being designed.[221]

If we do not actively work to incorporate other valid perspectives into the development process, we risk having the AI reflect our assumptions about how the product will be used and by whom, instead of being based on research evidence and empirical data.

- Qualitative experts or social scientists to help technologists and decision makers clarify ideas, create metrics, and objectively examine factors that would affect adoption of the AI
- Human factors or cognitive engineers to help ensure that AI is not just integrated into a technology or process, but is adopted willingly and with appropriately calibrated trust
- Accident analysis experts who can draw on a long history of post-accident insights and frameworks to improve system design and anticipate areas of concern
- Legal and policy experts to oversee that data use and governance are covered by relevant authorities, to identify legal implications of the deployed AI, and to ensure that the process is following established mechanisms of oversight.
- Privacy, civil liberties, and cybersecurity experts to help evaluate and if necessary mitigate how design choices could affect concerns in their respective areas
- The users of the AI and the communities that will be affected by the AI to reinforce the importance of meeting the desired outcomes of all stakeholders
- Educators to prepare the workforce in their respective fields to overcome misperceptions about AI's capabilities, help users identify how to spot and track problems with AI, and learn from previous good and poor experiences that come from the introduction of new tools.

The most successful teams are ones in which all perspectives are voiced and considered. To that end, we must remember to not only include multidisciplinary experts on the team, but also make sure that all teammates have equal decision-making power.[222]

Lesson #4. Incorporate Privacy, Civil Liberties, and Security from the Beginning

Let's borrow and extend the "Fundamental Theorem of Security" stated by Roman Yampolskiy, a professor at the University of Louisville, to say, "Every security system will eventually fail; {every piece of data collected will be used in unanticipated ways}. If your system has not failed, just wait longer."[223] (text in curly braces represents additions to the quotation).

Many AI-enabled systems rely on growing amounts of data in order to enable more accurate and more tailored pattern recognition. As that data becomes increasingly personal and sensitive, the costs that result from those datasets being misused, stolen, or more intricately connected become much greater and more alarming. Privacy, civil liberties, and security experts are now more essential to AI development than ever, because they specialize in recognizing and mitigating against the ways in which data can be used in unforeseen ways.[224]

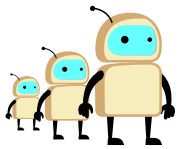
> Every security system will eventually fail; {every piece of data collected will be used in unanticipated ways}. If your system has not failed, just wait longer.

We must consider privacy-, civil liberties-, security-, and mission-related objectives at the beginning of the development project, when we can evaluate tradeoffs among the four. To aid us in understanding the risks involved and being proactive in preventing those risks, experts in these fields can help us navigate and resolve some of the following tensions:

- Collecting and using more data to achieve better quality outcomes vs. respecting individuals' privacy and ownership over their data[225]
- Making models or datasets openly available to the public for broader use and scrutiny vs. revealing more information that lets adversaries find new ways to hack the information[226]
- Meeting consumer demand for products that are becoming more integrated into their homes (and bodies) vs. mitigating the increasing consequences to their safety when those devices fail or are hacked[227]
- Balancing data and privacy protection in legislation, such as in Europe's General Data Protection Regulation (GDPR). Current policy differs across countries[228,229,230] and states.[231,232]

These considerations cannot be afterthoughts. Too often, the seductive values of cost savings and efficiencies blind commercial and government organizations to the need for addressing privacy, civil liberties, and security concerns adequately. Incorporating this expertise on our teams early offers a means for developing AI systems that can meet mission needs and simultaneously address these considerations.

Return to Table of Contents



Lesson #5. Involve the Communities Affected by the AI

When we design an application with only the end-user in mind, the application can have very different objectives and success criteria than if we design for the communities that the AI will affect. Two particularly powerful examples of one-sided implementation – facial recognition for policing, and AIs that recommend which patients receive healthcare – are described elsewhere in this paper. Those emotionally charged examples illustrate that both end-users and affected communities may be able to find common ground on desired outcomes if given the opportunity. But since the affected communities were not invited to discussions with AI developers, the developers did not design the system to reflect the communities' perspectives.

Therefore, we should be sure to include representatives from the communities that will be affected by the algorithm, in addition to the end-users. Treating these communities as customers, and even giving them a vote in choosing success criteria for the algorithm, is another step that would lead toward more human-centric outcomes.[233]

Treating these communities as customers, and even giving them a vote in choosing success criteria for the algorithm, is another step that would lead toward more human-centric outcomes.

These conversations should start early and continue past algorithm deployment. The University of Washington's Tech Policy Lab offers a step-by-step guide for facilitating inclusivity in technology policy.[234] It includes actions that can help organizations identify appropriate stakeholder groups, run group sessions, and close the loop between developers and the invited communities.

Why are these types of approaches so necessary? Education and exposure are powerful tools. They help us fill gaps in our knowledge: they help us to learn about communities' previous experiences with automation, and they give us insight regarding the level of explainability and transparency required for successful outcomes. In turn, those communities and potential users of the AI can learn how the AI works, align their expectations to the actual capabilities of the AI, and understand the risks involved in relying on the AI. Involving these communities will clarify the kinds of AI education, training, and advocacy needed to improve AI adoption and outcomes.[235,236] Then, we and the consumers of our AI products will be better able to anticipate adoption challenges, appreciate whether the risks and rewards of the systems apply evenly across individual users and communities, recognize how previous solutions (automated or not) have become successful, and protect under-represented populations.[237,238]

Lesson #6. Plan to Fail

Benjamin Franklin once said, "If you don't plan to fail, you fail to plan."[239] The uncertain and the unexpected are part of reality, but resiliency comes from having many ways to *prevent, moderate,* or *recover* from mistakes or failure.[240] Not all resilient methods have to be technical; they can rely on human participation and partnership. The overall amount of resiliency needed in an application increases as the AI's success becomes more critical for the overall outcome.

Prevent: If it's possible to reduce the criticality of the AI to the mission, we should do it. When it's not, we should follow the aircraft industry's example and eliminate single points of failure. Boeing, for example, has "three flight computers that function independently, with each computer containing three different processors manufactured by different companies."[241] Analog backups, such as old-fashioned paper and pen, can't be hacked or lose power.

If it's possible to reduce the criticality of the AI to the mission, we should do it.

Moderate: We should try to include some checks and balances. One idea might be to simply "cap" how extreme an outcome might be; as an analogy, a video-sharing platform could limit showing videos that are categorized as "too extreme."[242] Alternatively, AI projects should make use of human judgment by adding "alerts" for both us and for users; as an example, a video-sharing platform could alert viewers that a suggested video is linked to an account that has previously uploaded more extreme content.[243] These caps and alerts should correspond to the objectives and risk criteria set early in the AI development process.

Recover: We should anticipate that the AI will fail and try to envision the consequences. This means that we should consider identifying all systems that might be impacted, whether back-ups or analogs exist, if technical staff are trained to address those failures, how users are likely to respond to an AI failure, and hiring bad guys to find vulnerabilities before the technology is deployed.

We can usually improve resiliency by treating the intended users as partners. Communicating why we made particular decisions can go a long way toward reducing misunderstandings and misaligned assumptions. Also, offering a choice to the users or individuals affected by the AI allows people to decide what's best for their needs at the moment.

Lesson #7. Ask for Help: Hire a Villain

While we can leave it to bad actors or luck to identify vulnerabilities in a deployed AI, or we can proactively hire a team that's on our side to do it. Such "red teams" take the perspective of an adversary.

From the technology perspective, these surrogate villains can deploy automated software testing tools to find bugs and vulnerabilities. One interesting approach to meeting this shortfall is Netflix's "Simian Army," which intentionally introduces different types of automation failures in order to build resiliency into their architecture.[244]

One such tool is the “chaos monkey”,[245] which randomly shuts down services or elements of code to reveal where more strengthening can be beneficial.

We can also turn to professional “white-hat hackers.” White-hat hackers are experts (often formally certified) who hack for a good cause or to aid a company, organization, or government agency without causing harm.[246,247] Organizations such as Apple[248] and the Department of Defense[249] have hired white-hats or posted rewards for identifying and sharing vulnerabilities.

These surrogate villains should also go after more than just the technology. Red teams and white hats look for vulnerabilities that come from people and processes as well as the tech.[250] For example, is that entry to a building unguarded? Can a person be convinced to insert a USB stick with a virus on it into a system? Can that system be tricked into giving more access than intended? Red teams and white hats will try all that and more.

Hiring a villain reduces vulnerabilities and helps us build in more technical and procedural resiliency.

Lesson #8. Use Math to Reduce Bad Outcomes Caused by Math

First, we must accept that no data-driven solution will be perfect, and our goal shouldn’t be to achieve perfection. Instead we should try to understand and contextualize our errors.[251]

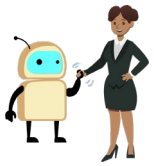
Looking at the data. We can apply existing statistical sampling mitigations to combat mathematical forms of bias that arise from sampling errors (which are distinct from bias caused by human influence). These mitigations include collecting larger samples and intentionally sampling from categorized populations (e.g., stratified random sampling).[252] In the last few years, statistical bias toolkits[253,254,255,256,257] have emerged that incorporate visualizations to help us understand our data. Specific toolkits[258] have also been developed to help us understand datasets that contain associations that are human-influenced (for example, the term “female” is more closely associated with “homemaker” than with “computer programmer” in a search of Google News articles[259]).

Looking at the algorithms. We can also offset an AI’s tendency to amplify patterns at the model level. One set of intervention methods imposes model constraints that push predictions toward a target statistical distribution[260] or uses guardrails that enforce limits to outcomes or trigger alerts for human investigation.[261] Another method helps reduce runaway feedback loops (which push behavior toward increasingly specialized and extreme ends) by restricting how outputs generated from model predictions should be fed back into the algorithm.[262] One simple diagnostic is to compare the distributions of predicted to observed outputs.[263]

Mathematical approaches can reduce the occurrence of undesired, mathematically-based outcomes. We must remember, though, that removing all mathematical error may not answer the social concerns about the AI’s impact. We must also remember that 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 more fundamental challenges behind complex issues.[264]

Return to Table of Contents

Calibrate Our Trust in the AI and the Data



Lesson #9. Make Our Assumptions Explicit

Let's start with an example: say we collect images of irises that grow in North America, and we train an AI to classify three different types of irises. The algorithm is pretty successful, and we want to share it with the world. If some potential users live in Europe and wants to use the algorithm, it's important for them to know that the accuracy would diminish for them because European irises look different, or that we only collected images in the daytime, or that we could only find a small sample for one type of iris. These users need to know the *assumptions and tradeoffs* behind the chosen training data, model parameters, and environment for that algorithm. Otherwise, they could be using the AI incorrectly or for purposes it was not intended to fulfill, but would trust in the outcomes nonetheless.

Generalizing from this example, many groups of people benefit from understanding the original developers' assumptions:

- Those who acquire or want to repurpose the AI systems need to know where the data comes from and what its characteristics are in order to make sure it aligns with their purposes.
- End users and consumers need to know how to appropriately interact with the AI so that they encounter fewer surprises and can more accurately weigh the risks of integrating the technology into their processes.
- AI policymakers and legislators need to know the original intended and unintended uses for the AI in order to apply, update, monitor, and govern it in an informed way.
- Objective third parties need to assess if the data and the algorithm's outcomes are mathematically and socially representative of the historical norms established in the domain where the AI is being deployed.[265]

Once they recognize the value of conveying these assumptions, organizations can take two steps to promote this practice.

1. Have the developers fill out standardized templates that capture assumptions and decisions. No one knows better about the intended and unintended uses for their data and tools than the original developers. Two sets of researchers from industry and academia have created templates that help draw out the developers' intents, assumptions, and discussions. The first, *datasheets for datasets,* documents the dataset's "[purpose], composition, collection process, recommended uses," decisions, and justifications.[266] Data choice and relevance are particularly critical to reduce bias and avoid placing miscalibrated trust in AIs.[267]

Serving as a complementary process, *model cards for model reporting* "clarify the intended use cases of machine learning models… provide benchmarked evaluation in a variety of conditions… and disclose the context in which models are intended to be used."[268] Understanding the intended context and use for the models is crucial to avoiding unwelcome surprises once the AI is deployed (in this case for machine learning,

one type of AI). Importantly, these two documents highlight both what was intended and specifically not intended.

Adopting the two templates as standard practice will go a long way toward helping us achieve transparency, explainability, and accountability in the AI we develop. The Partnership on AI – an organization of AI experts and stakeholders looking to formulate best practices on AI[269] – posted several examples as part of their ABOUT ML (Annotation and Benchmarking on Understanding and Transparency of Machine learning Lifecycles) initiative, aimed at experimenting with and scaling machine learning documentation efforts.[270]

The documentation process should... prompt us to bring in end users and affected communities to ensure they have the information they need, and... have the opportunity to offer suggestions. At the same time, the process should prompt analysts or decision makers... to capture how the input from an algorithm affected their overall assessment of a problem.

2. Structure documentation processes in a way that facilitates proactive and ongoing outreach. We as developers are best positioned to articulate the strengths and weaknesses of our systems, but other perspectives are needed to highlight the risks and design tradeoffs that we may not have considered. For example, end-users, lawyers, and policymakers (among others) may all have different questions that help us make informed decisions about the AI's appropriate uses, and they offer different considerations for mitigating potential risks. Even then, there are limitations to that group's collective knowledge. They might not catch all the biases or shortcomings in a first go-around, but the next user group would benefit greatly from the lessons learned in previous versions. Knowing what's been considered earlier helps new development teams integrate the different perspectives that were already offered and avoid repeating the same mistakes. Therefore, the documentation process should require recurring conversations with a diverse team.

Another key aspect of the documentation process is that we should be proactive in communicating bias and other limitations of systems to potential users, and not wait for a periodic review. Conveying design choices can be fundamentally transformative to the user's assessments of model appropriateness and trustworthiness. The documentation process should involve asking questions that prompt us to bring in end users and affected communities to ensure they have the information they need, and have the opportunity to offer suggestions early enough that we can incorporate their input in the product. At the same time, the process should prompt analysts or decision makers (if internal to the organization) to capture how the input from an algorithm affected their overall assessment of a problem. Making informed decisions is a joint responsibility. How each organization will implement these processes may differ. It might be easiest to expand on existing steps like user interviews, requirements generation, project management checks, quality control reviews, and other steps in a product's lifecycle. The organization may need to develop new processes specifically for documentation and explanations. Either way, thinking about these goals in advance means that we can make transparency part of the development process from the beginning of a project, and are therefore more likely to ensure it is done well.

The documentation process should require recurring conversations with a diverse team.

Lesson #10. Try Human-AI Couples Counseling

No AI, robot, or person works completely alone. If you've ever become frustrated with automation, you aren't alone – senior researchers from the Florida Institute for Human and Machine Cognition describe the feeling: "there's nothing worse than a so-called smart machine that can't tell you what it's doing, why it's doing something, or when it will finish. Even more frustrating—or dangerous— is a machine that's incapable of responding to human direction when something (inevitably) goes wrong."[271] And although an AI may not get frustrated, it can require the same things its human partners do: explanations from and an ability to influence its partners.

Partnership is not simply a game of tag – passing a task off and saying "Good luck." Human-AI partnership means two things: communicating what each party needs or expects from all its partners (whether human or AI), and designing a system that reinforces collaboration.

The first step is talking it out. Better AI-to-human (AI → H) communication gives humans a chance to calibrate their confidence and trust in the AI. This allows humans to trust that the AI can complete a task independently, and to understand why the system made its decisions and what the outcomes were. On the other hand, better H → AI communication gives the AI a better understanding of the users' priorities and needs, so it can adjust to those preferences. Overall, improved H ↔ AI communication makes it clearer when tradeoffs will occur, who (or what) is responsible for which part of the task, how humans and AI can best contribute on interdependent tasks, and how behaviors and preferences change over time.[272,273]

The solution is not for us to build systems that people trust completely, or for users only to accept systems that never err. Instead, lessons point to the importance of forming good partnerships based on evidence and perception.

The second step is thinking "combine and succeed" rather than "divide and conquer."[274] Each teammate, whether a human or an AI, must be able to *observe*, *predict,* and *direct* the state and actions of others on the team.[275,276]

In other words, both the human and the algorithmic partners have to maintain common ground, act in expected ways, and change behavior based on the partner's input. This result can manifest itself in the forms of explanations, signals, requests for attention, and declarations of current action.[277,278]

AI adopters often ask about ways to increase trust in the AI. The solution is not for us to build systems that people trust completely, or for users only to accept systems that never err. Instead, lessons point to the importance of forming good partnerships based on evidence and perception**.** Good partnerships help humans understand the AI's abilities and intents, believe that the AI will work as anticipated, and rely on the AI to the appropriate degree. Then stakeholders can calibrate their trust and weigh the potential consequences of the AI's decisions before granting appropriate authorities to the AI.[279]

Lesson #11. Offer the User Choices

During the design process, we make dozens of choices, assumptions, simplifications, and trade-offs (CAST) that affect the outcome of the AI system. In order to better understand the application domain, we invite stakeholders to share their preferences, desired outcomes, and how they would use the system. But at the end of the day, the CASTs remain with us since we're the technical experts. One way to reduce this knowledge gap is for us to document our decisions. But are there situations where user experience, or evolving user goals or behavior, make it more appropriate for the user to make decisions? What might it look like if, after deployment, we "extended" users' involvement by empowering them to weigh in on some of the choices, when its appropriate to do so?

One idea for giving the user more appropriate agency is to present the user with options that juxtapose how specific developer decisions influence the AI's objectives. For example, when debating between different instantiations of fairness, instead of leaving that decision to the developer, we could add a "dial" that would let the user switch between definitions. In this way they could select the approach that better aligns to their principles, or they can view a range of outcomes and form a more complete picture of the solutions space. When the dial is accompanied by explanations that include context around the developer's CASTs (perhaps an overview of what the algorithm is optimizing, properties of the data, and how the algorithm defines success), this implementation could improve outcomes by appropriately shifting decisions to the stakeholder that knows the situation or environment best.[280]

Another approach consists of providing different degrees of explanations, depending on the user need. Explanations can contain different levels of detail: users may accept an AI's decision at face value, want confidence scores of those decisions, want confidence scores and descriptions of how those scores are generated,[281] or may even want examples of how the algorithm reached a decision. Certain algorithms 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").[282] With this approach the users can select how much they need to know about the AI in order to make an informed decision about applying or not applying its outcomes.

More research is needed into how empowering users with choice would affect the accuracy and desirability of outcomes, and more research is needed into how to best capture and present the developer's CASTs in such a way that is meaningful for the user. On the one hand, the AI developers comprehend the complexities of AI design and the ramification of design decisions. Giving users seeming control over aspects they don't understand has the potential to give the illusions of clarity and informed control, cause additional automation bias, or simply allow the user to select an option that gives them the answer they want.

Yet, the decisions of the developers should not substitute for the range of outcomes and intents that the user might want. More research could suggest ways to give users agency relative to their technical understanding of an AI, and appropriate to how the AI is applied in their domain. At best, this approach can reemphasize the value of algorithms offering competing perspectives, or evidence and counterevidence, which can elicit more diverse ideas and open dialogue – thus reinforcing principles that are foundational to the health of democracies.[283]

Lesson #12. Promote Better Adoption through Gameplay

There's a big difference between imagining how an AI works and interacting with it in the real world. As a way to bridge that gap, we could invite different users to play with the AI in a more controlled environment. Gameplay lets different stakeholders explore how a technology may affect their lives, their work, or their attention. It allows everyone to move from "knowing" to "feeling" and forming mental models of how the AI works.[284] Gameplay is especially important for stakeholders to better understand AI technologies, which learn and adapt the more they interact.[285]

Gameplay is vital for bringing to light some of the differences between our assumptions and the behavior of stakeholders. These differences can manifest themselves in several ways:[286]

Discovering misalignment early is better than waiting until after deployment, when the AI may have had an adverse impact

- Various groups may interpret outcomes, definitions, and behaviors differently. For example, some cultures view increased personalization as a global good, while other cultures focus on communal outcomes.
- Various groups value and endorse different outcomes. For example, more data leads to better quality outcomes, but often comes at the cost of individual privacy and autonomy.
- Individuals change the relative value of particular outcomes depending on the context. In some contexts (e.g., AI medical diagnoses) user groups prefer accuracy over explanations, but prefer the reverse for AI-enabled job recruiting.[287]

If the technology is mature enough for us to create a working prototype, gameplay can take the form of user evaluations, table-top exercises (TTXs), or experiments. One example is the Defense Advanced Research Projects Agency's (DARPA) engagement with Marines while developing the Squad X program. DARPA paired AI-enabled ground and air vehicles with a squad of Marines, then gave the teams realistic operational tasks. Through gameplay, the AI-enabled vehicles progressed from providing reconnaissance – a traditional role for unmanned vehicles – to becoming valued members of the squad, protected by and enabling the Marines to achieve their objectives more efficiently.[288,289]

If the technology is still in a conceptual phase – perhaps just a "what if" – we can try simulation techniques or traditional wargaming. Simulation helps to demonstrate and develop how individuals will use the technology and informs what design changes will make the product better. Alternatively, traditional wargaming plays out how conceptual technologies can be integrated into tactics, decision making, and future training.[290,291]

Exploring the discrepancies between expectations and actual AI behavior as well as the differences in how stakeholders interact with the AI, is a powerful way to reach technical, social, and policy resolutions in specific situations. Discovering misalignment early is better than waiting until after deployment, when the AI may have had an adverse impact.[292]

Return to Table of Contents

Lesson #13. Monitor the AI's Impact and Establish Layers of Accountability

Modern-day engineers who design AI systems have the best of intentions. While we want our systems to benefit users, communities, and society in general, the reality is that after we deploy an AI, something – the data, the environment, how users interact with the AI– will change, and the algorithm will work in unexpected ways. When weighing all these potential outcomes, it is the impact of the AI on people's lives that matters most. Therefore, we need a strategy for monitoring the AI and assigning parties to implement changes to the AI based on that impact. When individual and organizational accountability is tied to that strategy, we get more responsible outcomes.

Approaches will require continuous monitoring and ongoing investments. To act quickly against unanticipated outcomes, organizations should take the following actions:

> It is the impact of the AI on people's lives that matters most

1. *Calculate baseline criteria for performance and risk.* At the beginning of the project, we should establish baseline performance criteria for acceptable functioning of the AI. As one AI writer/practitioner described, just like a driving a new car off the lot, "the moment you put a model in production, it starts degrading."[293] If the AI "drifts" enough from its baseline, we may have to retrain or even scrap the model. Baseline performance criteria should be both mathematical and contextual, and criteria should include the perspectives of all affected stakeholders.

   In parallel with performance criteria, risk assessment criteria should guide decisions about the AI's suitability to a given application domain or intended use. Prior to deploying the system, we should determine the threshold of clarity that different stakeholders require, and how well the AI meets those requirements. Organizational guidance should be clear for higher stakes cases, when legality, ethics, or potential impact areas of concern.

2. *Regularly monitor the AI's impact and require prompt fixes.* As part of a good project management plan, we should set up continuous, automated monitoring as well as a regular schedule for human review of a model's behavior. We should check that the algorithm's outputs are meeting the baseline criteria.[294] This will not only help refine the model, but also help us act promptly as harms or biases emerge.

   Because changes will have to be made to the model, the original development team should remain involved in the project after the AI is deployed.[295,296] As the number of AI projects increases, that original development can train new maintainers.

3. *Create a team that handles feedback from people impacted by the AI, including users.* Bias, discrimination, and exclusion can occur without our even knowing it. Therefore, we should make clear and publicize how

those affected by the AI can alert this feedback team. The organization can also create guidelines on how and when to act on this feedback.

In addition, this feedback team can be proactive. Some AI relies heavily on data; this team should broadcast how an individual's data is used and implement processes for discarding old data.[297] With its GPT-2 algorithm, Google set up an email address and guided other researchers looking to build off Google's work[298] – a particularly important step given the potential harmful outcomes of the application.

4. *Experiment with different accountability methods.* AI is a rapidly evolving technical field, and the interaction between AI and other applications creates a complex ecosystem. Therefore, accountability that works well today may not be equally effective as future technologies change that ecosystem. And as an organization's structure and culture evolves, so too may its accountability efficacy.[299]

   One example experiment comes from Microsoft, which established an AI, Ethics and Effects in Engineering and Research (AETHER) Committee in 2018. Wary of the suspicion that such a move would be viewed primarily as an attempt to improve public relations, Microsoft required direct participation in the committee by senior leadership. Microsoft also asked employees with different backgrounds to provide recommendations to senior leadership on challenging and sensitive AI outcomes and to help develop implementable policy and governance structures in conjunction with the company's legal team. The committee also set up an "Ask AETHER" phone line for employees to raise concerns.[300]

   The impacts from experiments like these are still being assessed, but their existence signals a growing willingness by organizations to implement oversight and accountability mechanisms.

AI has real consequences and is certain to continue to produce unintended outcomes. That is why we must explore all the possible perspectives to address this accountability challenge and to do our best to position our organizations to be proactive against, and responsive to, undesirable outcomes.

Lesson #14. Envision Safeguards for AI Advocates

If ethical outcomes are part of our organization's values, we need to devote resources and establish accountability among ourselves and our teams to ensure those values are upheld, and to protect those who fight to uphold those values.

Employees in AI organizations, both commercial and government, are organizing and protesting in response to perceived harmful outcomes arising from the products and organizational decisions of their leadership. Through walkouts,[301] advocacy,[302] and expressions of general concern[303] these employees are representing and reinforcing the ethical principles that their organizations proclaim. When these employees are punished or fired,[304,305] sometimes unlawfully,[306] they need stronger safeguards and top cover.

What might those safeguards look like? The AI Now Institute at New York University (a research institute dedicated to understanding the social implications of AI technologies) lays out specific approaches that

organizations should adopt to avoid social, economic, and legal penalties, including "clear policies accommodating and protecting conscientious objectors, ensuring workers the right to know what they are working on, and the ability to abstain from such work without retaliation or retribution. Workers raising ethical concerns must also be protected, as should whistleblowing in the public interest."[307] Support for workers would also include 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.[308]

Lesson #15. 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 representatives need to understand the entire lifecycle of the AI-enabled system: from evaluating the origins and 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.[309,310]

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

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.[311] Or think of "green energy" companies that respond to consumer preference for sustainable businesses and products, and enjoy more profits at the same time.[312] Both models center on a recognized, consensual set of criteria, as well as an (ideally, independent) evaluative body that confirms compliance with the standard. ForHumanity, a non-profit group that advocates for increased human rights awareness and protection as AI spreads, describes what such a program might look like with its SAFEAI Seal.[313]

Following these examples, evaluators should come from a range of 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[314,315] and promoting truth in advertising requirements for AI products and services.[316] O3VV agencies could take the form of government auditing programs, 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.[317] One approach calls for a licensing program for O3VV professionals, and another calls for increasing monetary incentives for deploying certified systems.[318] 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.[319,320]

Lesson #16. Entrust Sector-specific Agencies to Establish AI Standards for Their Domains

AI is increasingly integrated into more domains, including national defense, healthcare, education, criminal justice, and others. Establishing a global approach to AI governance is challenging because the legislative and social histories and policies in each domain differ drastically.[321] New technologies will be 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. In other words, doctors could use the AI's queues while still employing their own judgment, decision making, and authority to achieve improved outcomes.[322,323]

> 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

Second, as introduced earlier, state and local jurisdictions in the US have deployed COMPAS, a black box tool that assesses the risk of prison 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. The tool 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 presenting 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.[324,325,326]

In each of these cases, the expert could discount or act on the AI's recommendation. 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,[327] or because the court uses twelve jurors rather than a single judge, a practice established as representative of a good cross-section of perspectives.[328] In contrast, the public might be more at ease with AI offering predictions on medical diagnoses because doctors routinely use "evidence-based medicine"[329] 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,[330] 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. In early 2020, The White House 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."[331] It is incumbent on individual agencies to permit, regulate, temper, and even ban[332] AI-enabled systems as determined by the experts and established practices in each domain.

The French Data Protection Authority (the government agency responsible for the protection of personal data)[333] provides an example of two founding principles for AI standards:

- "A principle of fairness applied to all sorts of algorithms, which takes into account not only their personal outcomes but their collective ones as well. In other words, an algorithm… should be fair towards its users, not only as consumers but also as citizens, or even as communities or as an entire society.
- A principle of continued attention and vigilance: its point is to organize the ongoing state of alert that our societies need to adopt as regards the complex and changing socio-technical objects that algorithmic systems represent. It applies to every single stakeholder (designers, businesses, end-users) involved in 'algorithmic chains.'"

Government legislation on AI 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.[334]

Return to Table of Contents

CONCLUSION

Given the increasing integration of AI-enabled systems into most areas of daily life, we must remember that the decisions we make as we design and deploy AI systems, and the values and assumptions that shape those decisions, can have a profound impact on individuals and entire societies. We must constantly remind ourselves to evaluate the pedigree, type, and comprehensiveness of the data on which we base our AI designs; to include the broadest possible range of perspectives in our teams; to examine the impacts of our systems; and to ensure the proper balance between algorithmic decisions and human checks and balances.

We must also remember that the eventual users of AI systems lack our understanding of the maturity and reliability of the technology. As a result, they may view the outputs of our systems as “truth” and base important decisions upon those outputs, when in fact even the best-designed AIs vary in performance as environments or conditions change. Therefore, we should ensure that our systems are rigorously tested in controlled environments, and designed in ways that promote human partnership and disclosure sharing of information that would help stakeholders appropriately calibrate their trust in the AI.

Most fundamentally, we must always ask ourselves whether an AI-enabled system is even appropriate for meeting a given need. AI developers and deployers aren’t omniscient, and the AI we create can never be perfect, in the sense of always producing optimal outcomes for all users, all domains, and society at large. In our rapidly changing world, we cannot predict user needs, expectations, and requirements for AI-enabled systems, or anticipate all the possible ways users may apply – or misapply – the systems we produce, or all the possible personal and social consequences. But the examples of AI fails described in this paper, and the lessons learned from them, can guide us to create the best possible AI for a given problem, domain, and set of users and stakeholders, and for the societies in which we live.

AUTHORS

	***Mr. 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.*
	***Mr. Ron Hodge*** *is a national security strategist who provides strategic and technical leadership across multiple disciplines. He focuses on early identification of disruptive technologies and acts on opportunities to conceptualize and deploy new ideas to address the hardest challenges facing our nation.*
	***Dr. Lura Danley*** *is an applied psychologist who uses psychology-based principles and scientific methods to bridge gaps between human behavior, cybersecurity and technology. She specializes in providing research-based insights and data-driven analysis to address critical national security challenges.*

CONTRIBUTORS: Michael Aisenberg, Hassan Bermiss, Cindy Domniguez, Stan Drozdetski, Ron Ferguson, Ryan Fitzgerald, Sheila Gagen, Josh Kiihne, Marilyn Kupetz, Margaret MacDonald, Patrick Martin, Patty McDermott, Lidia Sabatini, Steve Stone, John Ursino, and Emma Williams

REVIEWERS: Eric Bloedorn, Chuck Howell, and Julie Steinke

Thank you to Richard Games, Lisa Bembenick, Eric Bloedorn, Aaron Lesser, and Chris Magrin for their support.

The MITRE Corporation (MITRE)—a not-for-profit organization—operates federally funded research and development centers (FFRDCs). These are unique organizations sponsored by government agencies under the Federal Acquisition Regulation to assist with research and development, study and analysis, and/or systems engineering and integration.

Approved for public release, distribution is unlimited. Case number 20-1365.

ENDNOTES

1 "Benjamin Franklin quotable quote," Goodreads. Accessed March 16, 2020. [Online]. Available: https://www.goodreads.com/quotes/460142-if-you-fail-to-plan-you-are-planning-to-fail

2 Department of Defense, "Summary of the 2018 Department of Defense Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity," *defense.gov,* February 12, 2019. [Online]. Available: https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF

3 ichristianization, "Microsoft build 2017 translator demo," YouTube, June 13, 2017. [Online]. Available: https://www.youtube.com/watch?v=u4cJoX-DoiY

4 N. Martin, "Artificial intelligence is being used to diagnose disease and design new drugs," *Forbes*, Sept. 30, 2019. [Online]. Available: https://www.forbes.com/sites/nicolemartin1/2019/09/30/artificial-intelligence-is-being-used-to-diagnose-disease-and-design-new-drugs/#8874c44db51f

5 "Meet the AI robots helping take care of elderly patients," *Time Magazine*, Aug. 23, 2019. [Online]. Available: https://time.com/5660046/robots-elderly-care/

6 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

7 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/

8 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/

9 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

10 S. Gibbs, "Tesla Model S cleared by auto safety regulator after fatal Autopilot crash," *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

11 E. Hunt, "Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter," *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

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 J. C. F. de Winter & D. Dodou, "Why the Fitts list has persisted throughout the history of function allocation," *SpringerLink*, August 25, 2011. [Online]. Available: https://link.springer.com/article/10.1007/s10111-011-0188-1

14 E. Brynjolfsson and A. McAfee, The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies. New York, NY, USA: W. W. Norton, 2016.

15 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

16 N. D. Sarter, D. D. Woods, and C. E. Billings, "Automation surprises," in G. Salvendy (Ed.), Handbook of Human Factors & Ergonomics (2nd ed., pp. 1926-1943). New York, NY, USA: John Wiley, 1997.

17 S. M. Casner and E. L. Hutchins, "What do we tell the drivers? Toward minimum driver training standards for partially automated cars," *Journal of Cognitive Engineering and Decision Making*, March 8, 2019. [Online]. Available: https://journals.sagepub.com/doi/full/10.1177/1555343419830901

[18] 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

[19] Data & Society, “Algorithmic accountability: A primer,” Tech Algorithm Briefing: How Algorithms Perpetuate Racial Bias and Inequality, prepared for the Congressional Progressive Caucus, April 18, 2018. [Online]. Available: https://datasociety.net/wp-content/uploads/2018/04/Data_Society_Algorithmic_Accountability_Primer_FINAL-4.pdf

[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] N. D. Sarter, D. D. Woods, and C. E. Billings, “Automation surprises,” in G. Salvendy (Ed.), *Handbook of Human Factors & Ergonomics* (2nd ed., pp. 1926-1943). New York, NY, USA: John Wiley, 1997.

[22] S. M. Casner and E. L. Hutchins, “What do we tell the drivers? Toward minimum driver training standards for partially automated cars,” *Journal of Cognitive Engineering and Decision Making*, March 8, 2019. [Online]. Available: https://journals.sagepub.com/doi/full/10.1177/1555343419830901

[23] J. Dastin, “Amazon scraps secret AI recruiting tool that showed bias against women,” *Reuters*, 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

[24] Defense Science Board, *Task Force Report: The Role of Autonomy in DoD Systems*, Washington, D.C., June 2016. [Online]. Available: https://www.hsdl.org/?abstract&did=722318

[25] M. McDonough, “Business-focus on artificial intelligence rising,” Twitter, Feb. 28, 2017. [Online]. Available: https://twitter.com/M_McDonough/status/836580294484451328

[26] C. O’Neil, “The era of blind faith in big data must end,” TED, April 2017. [Online]. Available: https://www.ted.com/talks/cathy_o_neil_the_era_of_blind_faith_in_big_data_must_end

[27] “Here to help,” *xkcd*. Accessed March 18, 2020. [Online]. Available: https://www.xkcd.com/1831/

[28] J. Brownlee, “A gentle introduction to transfer learning for deep learning,” Machine Learning Mastery, Sept. 16, 2019. [Online]. Available: https://machinelearningmastery.com/transfer-learning-for-deep-learning/

[29] S. Schuchmann, “History of the second AI winter,” towards data science, May 12, 2019. [Online]. Available: https://towardsdatascience.com/history-of-the-second-ai-winter-406f18789d45

[30] Defense Science Board, *Task Force Report: The Role of Autonomy in DoD Systems*, Washington, D.C., June 2016. [Online]. Available: https://www.hsdl.org/?abstract&did=722318

[31] A. Gregg, J. O'Connell, A. Ba Tran, and F. Siddiqui. “At tense meeting with Boeing executives, pilots fumed about being left in dark on plane software,” *Washington Post*, March 13, 2019. [Online]. Available: https://www.washingtonpost.com/business/economy/new-software-in-boeing-737-max-planes-under-scrutinty-after-second-crash/2019/03/13/06716fda-45c7-11e9-90f0-0ccfeec87a61_story.html

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

[33] 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

[34] M. McFarland, “My seat keeps vibrating. Will it make me a better driver before driving me insane?” *Washington Post*, Jan. 12, 2015. [Online]. Available: https://www.washingtonpost.com/news/innovations/wp/2015/01/12/my-seat-keeps-vibrating-will-it-make-me-a-better-driver-before-driving-me-insane/?noredirect=on&utm_term=.31792eb87c03

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

[36] Barry Friedman: Is technology making police better—or…” *Recode Decode podcast*, Nov. 24, 2019. [Online]. Available: https://www.stitcher.com/podcast/vox/recode-decode/e/65519494?curator=MediaREDEF

[37] 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/

[38] N. Bilton, “Nest thermostat glitch leaves users in the cold,” *New York Times*, Jan. 13, 2016. [Online]. Available: https://www.nytimes.com/2016/01/14/fashion/nest-thermostat-glitch-battery-dies-software-freeze.html

[39] A. J. Hawkins, “Everything you need to know about the Boeing 737 Max airplane crashes,” *The Verge*, March 22, 2019. [Online]. Available: https://www.theverge.com/2019/3/22/18275736/boeing-737-max-plane-crashes-grounded-problems-info-details-explained-reasons

[40] E. Ongweso, “Samsung Galaxy S10 ‘vault-like security’ beaten by a $3 screen protector,” *Vice*, Oct. 17, 2019. [Online]. Available: https://www.vice.com/en_us/article/59nqdb/samsung-galaxy-s10-vault-like-security-beaten-by-a-dollar3-screen-protector

[41] “Airplane redundancy systems” *Poente Technical*. Accessed April 3, 2020. [Online]. Available: https://www.poentetechnical.com/aircraft-engineer/airplane-redundancy-systems/

[42] AJ Vicens, “An Amazon Echo recorded a family’s private conversation and sent it to some random person,” *Mother Jones*, May 24, 2018. [Online]. Available: https://www.motherjones.com/politics/2018/05/an-amazon-echo-recorded-a-familys-private-conversation-and-sent-it-to-some-random-person/

[43] J. Oates, “Japanese hotel chain sorry that hackers may have watched guests through bedside robots,” *Register*, Oct. 22, 2019. [Online]. Available: https://www.theregister.co.uk/2019/10/22/japanese_hotel_chain_sorry_that_bedside_robots_may_have_watched_guests

[44] T. G. Dietterich and E. J. Horvitz, “Rise of Concerns about AI: Reflections and Directions,” *Communications of the ACM*, vol. 58, no. 10, pp. 38-40, October 2015. [Online]. Available: http://erichorvitz.com/CACM_Oct_2015-VP.pdf

[45] J. S. McEwen and S. S. Shapiro, “MITRE’S Privacy Engineering Tools and Their Use in a Privacy Assessment Framework,” The MITRE Corporation, McLean, VA, Nov. 2019. [Online]. Available: https://www.mitre.org/publications/technical-papers/mitre%E2%80%99s-privacy-engineering-tools-and-their-use-in-a-privacy

[46] University of Michigan Engineering, “Watch engineers hack a ‘smart home’ door lock,” YouTube, May 2, 2016. [Online]. Available: https://www.youtube.com/watch?v=Iwm6nvC9Xhc

[47] M. Hanrahan, “Ring security camera hacks see homeowners subjected to racial abuse, ransom demands,” *ABC News*, Dec. 12, 2019. [Online]. Available: https://abcnews.go.com/US/ring-security-camera-hacks-homeowners-subjected-racial-abuse/story?id=67679790

[48] “Cybersecurity Vulnerabilities Affecting Medtronic Implantable Cardiac Devices, Programmers, and Home Monitors: FDA Safety Communication.” US Food & Drug Administration, March 2019. [Online]. Available: https://www.fda.gov/medical-devices/safety-communications/cybersecurity-vulnerabilities-affecting-medtronic-implantable-cardiac-devices-programmers-and-home

[49] J. Herrman, “Google knows where you’ve been but does it know who you are,” *New York Times Magazine*, Sept. 12, 2018. [Online]. Available: https://www.nytimes.com/2018/09/12/magazine/google-maps-location-data-privacy.html

[50] A. Greenberg, “Hackers remotely kill a Jeep on the highway—with me in it,” *Wired*, July 21, 2015. [Online]. Available: https://www.wired.com/2015/07/hackers-remotely-kill-jeep-highway/

[51] L. Rocher, J. M. Hendrickx, and Y.-A. de Montjoye, “Estimating the success of re-identifications in incomplete datasets using generative models,” *Nature Communications*, July 23, 2019. [Online]. Available: https://www.nature.com/articles/s41467-019-10933-3

[52] “pwned,” *Urban Dictionary*. Accessed on: March 11, 2020. [Online]. Available: https://www.urbandictionary.com/define.php?term=pwned

[53] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, “A general framework for adversarial examples with objectives,” *arXiv.org*, April 4, 2019. [Online]. Available: https://arxiv.org/abs/1801.00349

[54] M. Fredrikson, S. Jha, T. Ristenpart, “Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures,” *CCS '15: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security,* October 2015, pp. 1322–1333. [Online]. Available: https://www.cs.cmu.edu/~mfredrik/papers/fjr2015ccs.pdf

[55] M. James, “Adversarial attacks on voice input,” *I Programmer*, Jan. 31, 2018. [Online]. Available: https://www.i-programmer.info/news/105-artificial-intelligence/11515-adversarial-attacks-on-voice-input.html

[56] G. Ateniese et al., “Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers,” *arXiv.org*, June 19, 2013. [Online]. Available: https://arxiv.org/abs/1306.4447

[57] A. Polyakov, “How to attack Machine Learning (Evasion, Poisoning, Inference, Trojans, Backdoors),” towards data science, August 6, 2019. [Online]. Available: https://towardsdatascience.com/how-to-attack-machine-learning-evasion-poisoning-inference-trojans-backdoors-a7cb5832595c

[58] K. Eykholt et al., “Robust physical-world attacks on deep learning models,” *arXiv.org*, April 10, 2018. [Online]. Available: https://arxiv.org/abs/1707.08945

[59] M. James, “Adversarial attacks on voice input,” *I Programmer*, Jan. 31, 2018. [Online]. Available: https://www.i-programmer.info/news/105-artificial-intelligence/11515-adversarial-attacks-on-voice-input.html

[60] A. Dorschel, “Rethinking data privacy: The impact of machine learning,” *Medium*, April 24, 2019. [Online]. Available: https://medium.com/luminovo/data-privacy-in-machine-learning-a-technical-deep-dive-f7f0365b1d60

[61] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, “A general framework for adversarial examples with objectives,” *arXiv.org*, April 4, 2019. [Online]. Available: https://arxiv.org/abs/1801.00349

[62] 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

[63] B. Barrett, “Lawmakers can’t ignore facial recognition’s bias anymore,” *Wired*, July 26, 2018. [Online]. Available: https://www.wired.com/story/amazon-facial-recognition-congress-bias-law-enforcement/

[64] P. Egan, “Data glitch was apparent factor in false fraud charges against jobless claimants,” *Detroit Free Press*, July 30, 2017. [Online]. Available: https://www.freep.com/story/news/local/michigan/2017/07/30/fraud-charges-unemployment-jobless-claimants/516332001/

[65] S. Mullainathan, “Biased algorithms are easier to fix than biased people,” *New York Times*, Dec. 6, 2019. [Online]. Available: https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html?searchResultPosition=1

[66] 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

[67] 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/

[68] Data & Society, “Algorithmic accountability: A primer,” Tech Algorithm Briefing: How Algorithms Perpetuate Racial Bias and Inequality, Prepared for the Congressional Progressive Caucus, April 18, 2018. [Online]. Available: https://datasociety.net/wp-content/uploads/2018/04/Data_Society_Algorithmic_Accountability_Primer_FINAL-4.pdf

[69] N. Barrowman, “Why data is never raw,” *New Atlantis*, Summer/Fall 2018. [Online]. Available: https://www.thenewatlantis.com/publications/why-data-is-never-raw

[70] 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/

[71] K. Crawford and R. Calo, “There is a blind spot in AI research,” *Nature*, Oct. 13, 2016. [Online]. Available: https://www.nature.com/articles/538311a

[72] D. Amodei, “Concrete problems in AI safety,” *arXiv.org*, July 25, 2016. [Online]. Available: https://arxiv.org/pdf/1606.06565.pdf

[73] N. V. Patel, “Why doctors aren’t afraid of better, more efficient AI diagnosing cancer,” *Daily Beast*, Dec. 22, 2017. [Online]. Available: https://www.thedailybeast.com/why-doctors-arent-afraid-of-better-more-efficient-ai-diagnosing-cancer

[74] T. Murphy VII, “The first level of Super Mario Bros. is easy with lexicographic orderings and time travel... after that it gets a little tricky,” April 1, 2013. [Online]. Available: http://www.cs.cmu.edu/~tom7/mario/mario.pdf

[75] J. Vincent, “OpenAI has published the text-generating AI it said was too dangerous to share,” *The Verge,* November 7, 2019. [Online]. Available: https://www.theverge.com/2019/11/7/20953040/openai-text-generation-ai-gpt-2-full-model-release-1-5b-parameters

[76] “GPT-2: 1.5B Release,” *OpenAI*, November 5, 2019. [Online]. Available: https://openai.com/blog/gpt-2-1-5b-release/

[77] Data & Society, “Algorithmic accountability: A primer,” Tech Algorithm Briefing: How Algorithms Perpetuate Racial Bias and Inequality, Prepared for the Congressional Progressive Caucus, April 18, 2018. [Online]. Available: https://datasociety.net/wp-content/uploads/2018/04/Data_Society_Algorithmic_Accountability_Primer_FINAL-4.pdf

[78] 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

[79] “College Board Announces Improved Admissions Resource,” *College Board,* August 27, 2019. [Online]. Available: https://www.collegeboard.org/releases/2019/college-board-announces-improved-admissions-resource

[80] A. Jenkins, “This town is fining drivers to fight 'horrific' traffic from Google Maps and Waze,” *Travel + Leisure*, Dec. 26, 2017. [Online]. Available: https://www.travelandleisure.com/travel-news/leonia-waze-google-maps-fines

[81] A. Feng and S. Wu, “The myth of the impartial machine,” *Parametric Press*, no. 01 (Science + Society), May 1, 2019. [Online]. Available: https://parametric.press/issue-01/the-myth-of-the-impartial-machine/

[82] 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/

[83] “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

[84] M. Heid, “The unsettling ways tech is changing your personal reality,” *Elemental*, Oct. 3, 2019. [Online]. Available: https://elemental.medium.com/technology-is-fundamentally-changing-the-ways-you-think-and-feel-b4bbfdefc2ee

[85] 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

[86] W. Oremus, “Who controls your Facebook feed,” *Slate*, Jan. 3, 2016. [Online]. Available: http://www.slate.com/articles/technology/cover_story/2016/01/how_facebook_s_news_feed_algorithm_works.html

[87] “Tech experts: What you post online could be directly impacting your insurance coverage,” *CBS New York*, March 21, 2019. [Online]. Available: https://newyork.cbslocal.com/2019/03/21/online-posting-dangerous-selfies-insurance-coverage/

[88] R. Deller, “Book review: Automating inequality: How high-tech tools profile, police and punish the poor by Virginia Eubanks,” *LSE Review of Books blog*, July 2, 2018. [Online]. Available: https://blogs.lse.ac.uk/lsereviewofbooks/2018/07/02/book-review-automating-inequality-how-high-tech-tools-profile-police-and-punish-the-poor-by-virginia-eubanks/

[89] “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

[90] “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

[91] “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

[92] P. Scharre, “Killer apps: The real dangers of an AI arms race,” *Foreign Affairs*, March/April 2019. [Online]. Available: https://www.foreignaffairs.com/articles/2019-04-16/killer-apps

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

[94] 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/

[95] "Ahsan v The Secretary of State for the Home Department (Rev 1) [2017] EWCA Civ 2009 (05 December 2017)." British and Irish Legal Information Institute, December 5, 2017. [Online]. Available: http://www.bailii.org/ew/cases/EWCA/Civ/2017/2009.html

[96] P. Wu, "Test your machine learning algorithm with metamorphic testing," *Medium*, Nov. 13, 2017. [Online]. Available: https://medium.com/trustableai/testing-ai-with-metamorphic-testing-61d690001f5c

[97] I. Goodfellow and N. Papernot, "The challenge of verification and testing of machine learning," *Cleverhans blog*, June 14, 2017. [Online]. Available: http://www.cleverhans.io/security/privacy/ml/2017/06/14/verification.html

[98] Raphael, "Introducing tf-explain, interpretability for TensorFlow 2.0," *Sicara blog*, July 30, 2019. [Online]. Available: https://blog.sicara.com/tf-explain-interpretability-tensorflow-2-9438b5846e35

[99] "Fit interpretable machine learning models. Explain blackbox machine learning," GitHub. Accessed March 13, 2020. [Online]. Available: https://github.com/Microsoft/interpret

[100] Y. Sun et al., " Structural test coverage criteria for deep neural networks," in *2019 IEEE/ACM 41st International Conference on Software Engineering: Companion Proceedings*, 2019. [Online]. Available: https://www.kroening.com/papers/emsoft2019.pdf

[101] L. M. Strickhart and H.N.J. Lee, "Show your work: Machine learning explainer tools and their use in artificial intelligence assurance," The MITRE Corporation, McLean, VA, June 2019, unpublished.

[102] D. Sculley et al., "Machine learning: The high interest credit card of technical debt," in *SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)*. Accessed March 16, 2020. [Online]. Available: https://ai.google/research/pubs/pub43146

[103] A. Madan, "3 practical ways to future-proof your IoT devices," *IoT Times*, July 2, 2019. [Online]. Available: https://iot.eetimes.com/3-practical-ways-to-future-proof-your-iot-devices/

[104] 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

[105] D. Sculley et al., "Machine learning: The high interest credit card of technical debt," in *SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)*. Accessed March 16, 2020. [Online]. Available: https://ai.google/research/pubs/pub43146

[106] D. Sculley et al., "Machine learning: The high interest credit card of technical debt," in *SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)*. Accessed March 16, 2020. [Online]. Available: https://ai.google/research/pubs/pub43146

[107] R. Potember, "Perspectives on Research in Artificial Intelligence and Artificial General Intelligence Relevant to DoD," Defense Technical Information Center, Jan. 1, 2017. [Online]. Available: https://apps.dtic.mil/docs/citations/AD1024432

[108] J. Zittrain, "The hidden costs of automated thinking," *The New Yorker*, July 23, 2019. [Online]. Available: https://www.newyorker.com/tech/annals-of-technology/the-hidden-costs-of-automated-thinking

[109] N. Carne, "Blaming the driver in a 'driverless' car," *Cosmos*, Oct. 29. 2019. [Online]. Available: https://cosmosmagazine.com/technology/blaming-the-driver-in-a-driverless-car

[110] S. Captain, "Humans were to blame in Google self-driving car crash, police say," Fast Company, May 4, 2018. [Online]. Available: https://www.fastcompany.com/40568609/humans-were-to-blame-in-google-self-driving-car-crash-police-say

[111] J. Stewart, "Tesla's autopilot was involved in another deadly car crash," *Wired*, March 30, 2018. [Online]. Available: https://www.wired.com/story/tesla-autopilot-self-driving-crash-california/

[112] J. Stewart, "Why Tesla's autopilot can't see a stopped firetruck," *Wired*, Aug. 27, 2018. [Online]. Available: https://www.wired.com/story/tesla-autopilot-why-crash-radar/

[113] M. McFarland, "Uber self-driving car kills pedestrian in first fatal autonomous crash," *CNN Business*, March 19, 2018. [Online]. Available: https://money.cnn.com/2018/03/19/technology/uber-autonomous-car-fatal-crash/index.html

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

[115] S. M. Casner and E. L. Hutchins, “What do we tell the drivers? Toward minimum driver training standards for partially automated cars,” *Journal of Cognitive Engineering and Decision Making*, March 8, 2019. [Online]. Available: https://journals.sagepub.com/doi/full/10.1177/1555343419830901

[116] W. Langewiesche, “The human factor,” *Vanity Fair*, Oct. 2014. [Online]. Available: https://www.vanityfair.com/news/business/2014/10/air-france-flight-447-crash

[117] “A320, vicinity Tel Aviv Israel, 2012,” *SKYbrary*. Accessed on: March 11, 2020. [Online]. Available: https://www.skybrary.aero/index.php/A320,_vicinity_Tel_Aviv_Israel,_2012

[118] S. Gibbs, “Tesla Model S cleared by auto safety regulator after fatal Autopilot crash,” *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

[119] C. Ross and I. Swetlitz, “IBM’s Watson supercomputer recommended ‘unsafe and incorrect’ cancer treatments, internal documents show,” *STATnews,* July 25, 2018. [Online]. Available: https://www.statnews.com/wp-content/uploads/2018/09/IBMs-Watson-recommended-unsafe-and-incorrect-cancer-treatments-STAT.pdf

[120] S. Fussell, “Pearson Embedded a 'Social-Psychological' Experiment in Students' Educational Software [Updated],” *Gizmodo*, April 18, 2018. [Online]. Available: https://gizmodo.com/pearson-embedded-a-social-psychological-experiment-in-s-1825367784

[121] 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

[122] 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

[123] R. Wexler, “When a computer program keeps you in jail,” *New York Times*, June 13, 2017. [Online]. Available: https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html

[124] C. Langford, “Houston Schools Must Face Teacher Evaluation Lawsuit,” *Courthouse News Service,* May 8, 2017. [Online]. Available: https://www.courthousenews.com/houston-schools-must-face-teacher-evaluation-lawsuit/

[125] B. Khaleghi, “The what of explainable AI,” Element AI, Sept. 3, 2019. [Online]. Available: https://www.elementai.com/news/2019/the-what-of-explainable-ai

[126] 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

[127] A. Yoo, “UPS: Driving performance by optimizing driver behavior,” Harvard Business School Digital Initiative, April 5, 2017. [Online]. Available: https://digital.hbs.edu/platform-digit/submission/ups-driving-performance-by-optimizing-driver-behavior/

[128] K. Hill, “Facebook recommended that this psychiatrist's patients friend each other,” *Splinternews*, Aug. 29, 2016. [Online]. Available: https://splinternews.com/facebook-recommended-that-this-psychiatrists-patients-f-1793861472

[129] P. L. McDermott, “Human-machine teaming systems engineering guide,” The MITRE Corporation, Dec. 2018. [Online]. Available: https://www.mitre.org/publications/technical-papers/human-machine-teaming-systems-engineering-guide

[130] D. Gunning, “Explainable artificial intelligence (XAI),” Defense Advanced Research Projects Agency, Nov. 2017. [Online]. Available: https://www.darpa.mil/attachments/XAIProgramUpdate.pdf?source=post_page----------

[131] Z. C. Lipton, “The mythos of model interpretability,” *arXiv.org*, March 6, 2017. [Online]. Available: https://arxiv.org/abs/1606.03490

[132] 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

[133] Z. C. Lipton, “The mythos of model interpretability,” *arXiv.org*, March 6, 2017. [Online]. Available: https://arxiv.org/abs/1606.03490

[134] 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

[135] “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

[136] “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

[137] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner, “Overtrust of robots in emergency evacuation scenarios,” presented at 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Christchurch, 2016, pp. 101-108. [Online]. Available: https://www.cc.gatech.edu/~alanwags/pubs/Robinette-HRI-2016.pdf

[138] Georgia Tech, “In emergencies, should you trust a robot?” YouTube. Accessed March 13, 2020. [Online]. Available: https://www.youtube.com/watch?v=frr6cVBQPXQ

[139] M. Heid, “The unsettling ways tech is changing your personal reality,” *Elemental*, Oct. 3, 2019. [Online]. Available: https://elemental.medium.com/technology-is-fundamentally-changing-the-ways-you-think-and-feel-b4bbfdefc2ee

[140] M. Vazquez, A. May, A. Steinfeld, and W.-H. Chen, “A deceptive robot referee in a multiplayer gaming environment,” Conference Paper, *Proceedings of 2011 International Conference on Collaboration Technologies and Systems (CTS)*, pp. 204-211, May 2011. [Online]. Available: https://www.ri.cmu.edu/publications/a-deceptive-robot-referee-in-a-multiplayer-gaming-environment/

[141] L. Hansen, ”8 drivers who blindly followed their GPS into disaster,” *The Week*, May 7, 2013. [Online]. Available: https://theweek.com/articles/464674/8-drivers-who-blindly-followed-gps-into-disaster

[142] P. Madhavan and D. A. Wiegmann, “Similarities and differences between human-human and human-automation trust: An integrative review,” *Theoretical Issues in Ergonomics Science*, vol. 8, no. 4, pp. 277-301, 2007).

[143] “Appeal to authority,” Legally Fallacious. Accessed March 25, 2020. [Online]. Available: https://www.logicallyfallacious.com/logicalfallacies/Appeal-to-Authority

[144] M. Chalabi, “Weapons of math destruction: Cathy O’Neil adds up the damage of algorithms,” *Guardian*, Oct. 27, 2016. [Online]. Available: https://www.theguardian.com/books/2016/oct/27/cathy-oneil-weapons-of-math-destruction-algorithms-big-data

[145] S. M. Casner and E. L. Hutchins, “What do we tell the drivers? Toward minimum driver training standards for partially automated cars,” *Journal of Cognitive Engineering and Decision Making*, March 8, 2019. [Online]. Available: https://journals.sagepub.com/doi/full/10.1177/1555343419830901

[146] Data & Society, “Algorithmic accountability: A primer,” Tech Algorithm Briefing: How Algorithms Perpetuate Racial Bias and Inequality, Prepared for the Congressional Progressive Caucus, April 18, 2018. [Online]. Available: https://datasociety.net/wp-content/uploads/2018/04/Data_Society_Algorithmic_Accountability_Primer_FINAL-4.pdf

[147] “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

[148] B. Aguera y Arcas, “Physiognomy’s new clothes,” *Medium*, May 6, 2017. [Online]. Available: https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a

[149] Synced, “2018 in review: 10 AI failures,” *Medium*, Dec. 10, 2018. [Online]. Available: https://medium.com/syncedreview/2018-in-review-10-ai-failures-c18faadf5983

[150] 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

[151] S. Levin, “New AI can guess whether you're gay or straight from a photograph,” *Guardian*, Sept. 7, 2017. [Online]. Available: https://www.theguardian.com/technology/2017/sep/07/new-artificial-intelligence-can-tell-whether-youre-gay-or-straight-from-a-photograph

[152] Synced, “2018 in review: 10 AI failures,” *Medium*, Dec. 10, 2018. [Online]. Available: https://medium.com/syncedreview/2018-in-review-10-ai-failures-c18faadf5983

[153] N. D. Sarter, D. D. Woods, and C. E. Billings, “Automation surprises,” in G. Salvendy (Ed.), *Handbook of Human Factors & Ergonomics* (2nd ed., pp. 1926-1943). New York, NY, USA: John Wiley, 1997.

[154] S. M. Casner and E. L. Hutchins, “What do we tell the drivers? Toward minimum driver training standards for partially automated cars,” *Journal of Cognitive Engineering and Decision Making*, March 8, 2019. [Online]. Available: https://journals.sagepub.com/doi/full/10.1177/1555343419830901

[155] S. M. Casner and E. L. Hutchins, “What do we tell the drivers? Toward minimum driver training standards for partially automated cars,” *Journal of Cognitive Engineering and Decision Making*, March 8, 2019. [Online]. Available: https://journals.sagepub.com/doi/full/10.1177/1555343419830901

[156] “A320, vicinity Tel Aviv Israel, 2012,” *SKYbrary*. Accessed on: March 11, 2020. [Online]. Available: https://www.skybrary.aero/index.php/A320,_vicinity_Tel_Aviv_Israel,_2012

[157] R. Nieva, “Facebook put cork in chatbots that created a secret language,” *CNET*, July 31, 2017. [Online]. Available: https://www.cnet.com/news/what-happens-when-ai-bots-invent-their-own-language/

[158] G. Klien et al., “Ten challenges for making automation a ‘team player’ in joint human-agent activity,” *IEEE: Intelligent Systems*, vol. 19, no. 6, pp. 91-95, Nov./Dec. 2004. [Online]. Available: http://jeffreymbradshaw.net/publications/17._Team_Players.pdf_1.pdf

[159] J. B. Lyons, “Being transparent about transparency: A model for human-robot interaction,” in *2013 AAAI Spring Symposium Series,* 2013. [Online]. Available: https://www.semanticscholar.org/paper/Being-Transparent-about-Transparency%3A-A-Model-for-Lyons/840080df8a02de6aab098e7eabef84831ac95428

[160] D. Woods, “Generic support requirements for cognitive work: laws that govern cognitive work in action,” *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* vol. 49, pp. 317-321, Sept. 1, 2005. [Online]. Available: https://journals.sagepub.com/doi/10.1177/154193120504900322

[161] “Luddite,” *Merriam-Webster*. Accessed April 11, 2020. [Online]. Available: https://www.merriam-webster.com/dictionary/Luddite

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

[163] D. Simberkoff, “How Facebook's Cambridge Analytica scandal impacted the intersection of privacy and regulation,” *CMS Wire*, Aug. 30, 2018. [Online]. Available: https://www.cmswire.com/information-management/how-facebooks-cambridge-analytica-scandal-impacted-the-intersection-of-privacy-and-regulation/

[164] D. Wray, “The companies cleaning the deepest, darkest parts of social media,” *Vice*, June 26, 2018. [Online]. Available: https://www.vice.com/en_us/article/ywe7gb/the-companies-cleaning-the-deepest-darkest-parts-of-social-media

[165] “Why a #Google walkout organizer left Google,” *Medium*, June 7, 2019. [Online]. Available: https://medium.com/@GoogleWalkout/why-a-googlewalkout-organizer-left-google-26d1e3fbe317

[166] “Technology adoption life cycle,” Wikipedia. Accessed March 17, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Technology_adoption_life_cycle

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

[168] “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

[169] C. F. Kerry, “Protecting privacy in an AI-driven world,” Brookings, Feb. 10, 2020. [Online]. Available: https://www.brookings.edu/research/protecting-privacy-in-an-ai-driven-world/

[170] 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/

[171] S. Browne, *Dark Matters: On the Surveillance of Blackness*, Durham, NC, USA: Duke University Press Books, 2015. [Online]. Available: https://www.dukeupress.edu/dark-matters

[172] 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

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

[174] 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

[175] 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/

[176] E. Ongweso, "Google is investigating why it trained facial recognition on 'dark skinned' homeless people," *Vice*, Oct. 4, 2019. [Online]. Available: https://www.vice.com/en_us/article/43k7yd/google-is-investigating-why-it-trained-facial-recognition-on-dark-skinned-homeless-people

[177] J. Stanley, "Secret Service announces test of face recognition system around White House," *ACLU bog*, Dec. 4, 2018. [Online]. Available: https://www.aclu.org/blog/privacy-technology/surveillance-technologies/secret-service-announces-test-face-recognition

[178] R. Courtland, "Bias detectives: The researchers striving to make algorithms fair," *Nature*, June 20, 2018. [Online]. Available: https://www.nature.com/articles/d41586-018-05469-3

[179] D. Robinson and L. Koepke, "Stuck in a pattern: Early evidence on 'predictive policing' and civil rights," Upturn, Aug. 2016. [Online]. Available: https://www.upturn.org/reports/2016/stuck-in-a-pattern/

[180] R. Courtland, "Bias detectives: The researchers striving to make algorithms fair," *Nature*, June 20, 2018. [Online]. Available: https://www.nature.com/articles/d41586-018-05469-3

[181] D. Robinson and L. Koepke, "Stuck in a pattern: Early evidence on 'predictive policing' and civil rights," Upturn, Aug. 2016. [Online]. Available: https://www.upturn.org/reports/2016/stuck-in-a-pattern/

[182] D. Robinson and L. Koepke, "Stuck in a pattern: Early evidence on 'predictive policing' and civil rights," Upturn, Aug. 2016. [Online]. Available: https://www.upturn.org/reports/2016/stuck-in-a-pattern/

[183] "An ethics guidelines global inventory," Algorithm Watch. Accessed on: Jan. 17, 2020. [Online]. Available: https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/

[184] "An ethics guidelines global inventory," Algorithm Watch. Accessed on: Jan. 17, 2020. [Online]. Available: https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/

[185] T. Hagendorff, "The ethics of AI ethics: An evaluation of guidelines," *arXiv.org*, Oct. 11, 2019. [Online]. Available: https://arxiv.org/abs/1903.03425

[186] 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

[187] "Wrestling with AI governance around the world," *Forbes*, March 27, 2019. [Online]. Available: https://www.forbes.com/sites/insights-intelai/2019/03/27/wrestling-with-ai-governance-around-the-world/#7d3f84ed1766

[188] 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*, July 24, 2019. [Online]. Available: https://www.governing.com/topics/public-justice-safety/gov-cities-ban-government-use-facial-recognition.html

[189] P. Martineau, "Cities examine proper—and improper—uses of facial recognition," *Wired*, Nov. 10, 2019. [Online]. Available: https://www.wired.com/story/cities-examine-proper-improper-facial-recognition/

[190] "Ban facial recognition." Accessed March 17, 2020. [Online]. Available: https://www.banfacialrecognition.com/map/

[191] "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

[192] 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

[193] J. Spitzer, "IBM's Watson recommended 'unsafe and incorrect' cancer treatments, STAT report finds," *Becker's Health IT*, July 25, 2018. [Online]. Available: https://www.beckershospitalreview.com/artificial-intelligence/ibm-s-watson-recommended-unsafe-and-incorrect-cancer-treatments-stat-report-finds.html

[194] A. Liptak, "The US Navy will replace its touchscreen controls with mechanical ones on its destroyers," *The Verge*, Aug. 11, 2019. [Online]. Available: https://www.theverge.com/2019/8/11/20800111/us-navy-uss-john-s-mccain-crash-ntsb-report-touchscreen-mechanical-controls

[195] T. Simonite, "When It Comes to Gorillas, Google Photos Remains Blind," *Wired*, January 11, 2018. [Online]. Available: https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

[196] "NICE cybersecurity workforce framework resource center," National Institute of Standards and Technology. Accessed March 17, 2020. [Online]. Available: https://www.nist.gov/itl/applied-cybersecurity/nice/nice-cybersecurity-workforce-framework-resource-center

[197] S. Anand and T. Bärnighausen, "Health workers at the core of the health system: Framework and research issues," Global Health Workforce Alliance, 2011. [Online]. Available: https://www.who.int/workforcealliance/knowledge/resources/frameworkandresearch_dec2011/en/

[198] Lippincott Solutions, "Interdisciplinary care plans: Teamwork makes the dream work," *Calling the Shots blog*, Sept. 6, 2018. [Online]. Available: http://lippincottsolutions.lww.com/blog.entry.html/2018/09/06/interdisciplinaryca-z601.html

[199] M. Mahdizadeh, A. Heydari, and H. K. Moonaghi, "Clinical interdisciplinary collaboration models and frameworks from similarities to differences: A systematic review," *Global Journal of Health Science*, vol. 7, no. 6, pp. 170-180, Nov. 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4803863/

[200] C. Hagel, "Reagan national defense forum keynote," Secretary of Defense Speech, Ronald Reagan Presidential Library, Simi Valley, CA, Nov. 15, 2014. [Online]. Available: https://www.defense.gov/Newsroom/Speeches/Speech/Article/606635/

[201] "Reports," National Security Commission on Artificial Intelligence. Accessed March 18, 2020. [Online]. Available: https://www.nscai.gov/reports

[202] "Bad data costs United Airlines $1B annually," Travel Data Daily. Accessed March 16, 2020. [Online]. Available: https://www.traveldatadaily.com/bad-data-costs-united-airlines-1b-annually/

[203] B. Vergakis, "The Navy, Air Force and Army collect different data on aircraft crashes. That's a big problem," *Task & Purpose*, Aug. 16, 2018. [Online]. Available: https://taskandpurpose.com/aviation-mishaps-data-collection

[204] B. Marr, "How much data do we create every day? The mind-blowing stats everyone should read," *Forbes*, May 21, 2018. [Online]. Available: https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/

[205] "No AI until the data is fixed," *Wired*, Feb. 22, 2019. [Online]. Available: https://www.wired.co.uk/article/no-ai-until-the-data-is-fixed

[206] D. Robinson and L. Koepke, "Stuck in a pattern: Early evidence on 'predictive policing' and civil rights," Upturn, Aug. 2016. [Online]. Available: https://www.upturn.org/reports/2016/stuck-in-a-pattern/

[207] "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

[208] U.S. Government Accountability Office, "Information technology: Federal agencies need to address aging legacy systems," GAO-16-696T, May 25, 2016. [Online]. Available: https://www.gao.gov/products/GAO-16-696T

[209] D. Cassel, “COBOL is everywhere. Who will maintain it?” *The New Stack*, May 6, 2017. [Online]. Available: https://thenewstack.io/cobol-everywhere-will-maintain/

[210] J. Uchill, “How did the government’s technology get so bad?” *The Hill*, Dec. 13, 2016. [Online]. Available: https://thehill.com/policy/technology/310271-how-did-the-governments-technology-get-so-bad

[211] B. Balter, “19 reasons why technologists don’t want to work at your government agency,” April 21, 2015. [Online]. Available: https://ben.balter.com/2015/04/21/why-technologists-dont-want-to-work-at-your-agency/

[212] U.S. Government Accountability Office, “Information technology: Federal agencies need to address aging legacy systems,” GAO-16-696T, May 25, 2016. [Online]. Available: https://www.gao.gov/products/GAO-16-696T

[213] D. Cassel, “COBOL is everywhere. Who will maintain it?” *The New Stack*, May 6, 2017. [Online]. Available: https://thenewstack.io/cobol-everywhere-will-maintain/

[214] “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

[215] 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

[216] S. Gibbs, “Tesla Model S cleared by auto safety regulator after fatal Autopilot crash,” *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

[217] D. Tomchek and S. Krawlzik, “Looking beyond the technical to fill America's cyber workforce gap,” *Nextgov*, Sept. 27, 2019. [Online]. Available: https://www.nextgov.com/ideas/2019/09/looking-beyond-technical-fill-americas-cyber-workforce-gap/160222/

[218] 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

[219] “Ethics & algorithms toolkit.” Accessed March 13, 2020. [Online]. Available: http://ethicstoolkit.ai/

[220] S. Ferro, “Here’s why facial recognition tech can’t figure out black people,” *HuffPost*, March 2, 2016. [Online]. Available: https://www.huffpost.com/entry/heres-why-facial-recognition-tech-cant-figure-out-black-people_n_56d5c2b1e4b0bf0dab3371eb

[221] S. J. Freedberg, “’Guess what, there’s a cost for that’: Getting cloud & AI right,” Breaking Defense, Nov. 26, 2019. [Online]. Available: https://breakingdefense.com/2019/11/guess-what-theres-a-cost-for-that-getting-cloud-ai-right/

[222] A. Campolo et al., *AI Now Report 2017*. New York, NY, USA: AI Now Institute, 2017. [Online]. Available: https://ainowinstitute.org/AI_Now_2017_Report.pdf

[223] R. V. Yampolskiy and M. S. Spellchecker, “Artificial intelligence safety and cybersecurity: A timeline of AI failures,” *arXiv.org*. Accessed March 25, 2020. [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1610/1610.07997.pdf

[224] J. Rotner, “The person at the other end of the data,” *Knowledge-Driven Enterprise blog*, The MITRE Corporation, Oct. 1, 2019. [Online]. Available: https://kde.mitre.org/blog/2019/10/01/the-person-at-the-other-end-of-the-data/

[225] J. Whittlestone, A. Alexandrova, R. Nyrup, and S. Cave, “The role and limits of principles in AI ethics: Towards a focus on tensions,” presented at AIES ’19, Jan. 27–28, 2019, Honolulu, HI, USA. [Online]. Available: https://www.researchgate.net/publication/334378492_The_Role_and_Limits_of_Principles_in_AI_Ethics_Towards_a_Focus_on_Tensions/link/5d269de0a6fdcc2462d41592/download

[226] I. Goodfellow, P. McDaniel, and N. Papernot, “Making machine learning robust against adversarial inputs,” *Communications of the ACM*, vol. 61, no. 7, pp. 56-66, July 2018. [Online]. Available: https://cacm.acm.org/magazines/2018/7/229030-making-machine-learning-robust-against-adversarial-inputs/fulltext

[227] R. V. Yampolskiy and M. S. Spellchecker, “Artificial intelligence safety and cybersecurity: A timeline of AI failures,” *arXiv.org*. Accessed March 25, 2020. [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1610/1610.07997.pdf

[228] "General data protection regulation," European Union. Accessed March 25, 2020. [Online]. Available: https://eugdpr.com/

[229] D. Miralis and P. Gibson, "Australia: Data protection 2019," *ICLG.com*, March 7, 2019. [Online]. Available: https://iclg.com/practice-areas/data-protection-laws-and-regulations/australia

[230] "Data protection laws of the world: New Zealand," DLA Piper. Accessed March 16, 2020. [Online]. Available: https://www.dlapiperdataprotection.com/index.html?t=law&c=NZ

[231] 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

[232] L. Hautala, "California's new data privacy law the toughest in the US," *CNET.com*, June 29, 2018. [Online]. Available: https://www.cnet.com/news/californias-new-data-privacy-law-the-toughest-in-the-us/

[233] 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

[234] "Diverse Voices: A How-To Guide for Facilitating Inclusiveness in Tech Policy." Accessed April 8, 2020. [Online]. Available: https://techpolicylab.uw.edu/project/diverse-voices/

[235] A. Campolo et al., *AI Now Report 2017*. New York, NY, USA: AI Now Institute, 2017. [Online]. Available: https://ainowinstitute.org/AI_Now_2017_Report.pdf

[236] 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

[237] A. Campolo et al., *AI Now Report 2017*. New York, NY, USA: AI Now Institute, 2017. [Online]. Available: https://ainowinstitute.org/AI_Now_2017_Report.pdf

[238] 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

[239] "Benjamin Franklin quotable quote," Goodreads. Accessed March 16, 2020. [Online]. Available: https://www.goodreads.com/quotes/460142-if-you-fail-to-plan-you-are-planning-to-fail

[240] 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

[241] M. Baker and D. Gates, "Lack of redundancies on Boeing 737 MAX system baffles some involved in developing the jet," *Seattle Times*, March 27, 2019. [Online]. Available: https://www.seattletimes.com/business/boeing-aerospace/a-lack-of-redundancies-on-737-max-system-has-baffled-even-those-who-worked-on-the-jet/

[242] 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/

[243] D. Amodei, "Concrete problems in AI safety," *arXiv.org*, July 25, 2016. [Online]. Available: https://arxiv.org/pdf/1606.06565.pdf

[244] "The Netflix Simian Army," *The Netflix Tech Blog,* July 19, 2011. [Online]. Available: https://netflixtechblog.com/the-netflix-simian-army-16e57fbab116

[245] C. A. Cois, "DevOps case study: Netflix and the chaos monkey," *DevOps blog*, Software Engineering Institute, April 30, 2015. [Online]. Available: https://insights.sei.cmu.edu/devops/2015/04/devops-case-study-netflix-and-the-chaos-monkey.html

[246] "White-hat," *Your Dictionary*. Accessed March 13, 2020. [Online]. Available: https://www.yourdictionary.com/white-hat

[247] E. Tittel and E. Follis, "How to become a white hat hacker," *Business News Daily*, June 17, 2019. [Online]. Available: https://www.businessnewsdaily.com/10713-white-hat-hacker-career.html

[248] K. Lerwing, “Apple hired the hackers who created the first Mac firmware virus,” *Business Insider*, Feb. 3, 2016. [Online]. Available: https://www.businessinsider.com/apple-hired-the-hackers-who-created-the-first-mac-firmware-virus-2016-2

[249] HackerOne, “What was it like to hack the Pentagon?” *h1 blog*, June 17, 2016. [Online]. Available: https://www.hackerone.com/blog/hack-the-pentagon-results

[250] J. Talamantes, “What is red teaming and why do I need it?” *RedTeam blog*. Accessed March 16, 2020. [Online]. Available: https://www.redteamsecure.com/what-is-red-teaming-and-why-do-i-need-it-2/

[251] 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/

[252] A. Feng and S. Wu, “The myth of the impartial machine,” *Parametric Press*, no. 01 (Science + Society), May 1, 2019. [Online]. Available: https://parametric.press/issue-01/the-myth-of-the-impartial-machine/

[253] “AI fairness 360 open source toolkit,” IBM Research Trusted AI. Accessed March 13, 2020. [Online]. Available: http://aif360.mybluemix.net/

[254] “Bias and fairness audit toolkit,” GitHub. Accessed March 13, 2020. [Online]. Available: https://github.com/dssg/aequitas

[255] “A Python package that implements a variety of algorithms that mitigate unfairness in supervised machine learning,” GitHub. Accessed March 13, 2020. [Online]. Available: https://github.com/Microsoft/fairlearn

[256] “What-if tool,” GitHub. Accessed March 13, 2020. [Online]. Available: https://pair-code.github.io/what-if-tool/

[257] “Facets,” GitHub. Accessed March 13, 2020. [Online]. Available: https://pair-code.github.io/facets/

[258] 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

[259] 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

[260] A. Feng and S. Wu, “The myth of the impartial machine,” *Parametric Press*, no. 01 (Science + Society), May 1, 2019. [Online]. Available: https://parametric.press/issue-01/the-myth-of-the-impartial-machine/

[261] D. Sculley et al., “Hidden technical debt in machine learning systems,” in *Advances in Neural Information Processing Systems 28 (NIPS 2015)*. Accessed March 16, 2020. [Online]. Available: https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

[262] A. Feng and S. Wu, “The myth of the impartial machine,” *Parametric Press*, no. 01 (Science + Society), May 1, 2019. [Online]. Available: https://parametric.press/issue-01/the-myth-of-the-impartial-machine/

[263] D. Sculley et al., “Hidden technical debt in machine learning systems,” in *Advances in Neural Information Processing Systems 28 (NIPS 2015)*. Accessed March 16, 2020. [Online]. Available: https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

[264] 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/

[265] 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

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

[267] “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

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

[269] “About Us,” Partnership On AI. Accessed May 27, 2020. [Online]. Available: https://www.partnershiponai.org/about/

[270] “Deployed Examples,” Partnership On AI. Accessed May 27, 2020. [Online]. Available: https://www.partnershiponai.org/about-ml/#examples

[271] J. M. Bradshaw, R. Hoffman, M. Johnson, and D. D. Woods, “The seven deadly myths of ‘autonomous systems,’” *IEEE: Intelligent Systems*, vol. 28, no. 3, pp. 54-61, May 2013. [Online]. Available: https://ieeexplore.ieee.org/document/6588858

[272] J. M. Bradshaw, R. Hoffman, M. Johnson, and D. D. Woods, “The seven deadly myths of ‘autonomous systems,’” *IEEE: Intelligent Systems*, vol. 28, no. 3, pp. 54-61, May 2013. [Online]. Available: https://ieeexplore.ieee.org/document/6588858

[273] S. M. Casner and E. L. Hutchins, “What do we tell the drivers? Toward minimum driver training standards for partially automated cars,” *Journal of Cognitive Engineering and Decision Making*, March 8, 2019. [Online]. Available: https://journals.sagepub.com/doi/full/10.1177/1555343419830901

[274] 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

[275] J. M. Bradshaw, R. Hoffman, M. Johnson, and D. D. Woods, “The seven deadly myths of ‘autonomous systems,’” *IEEE: Intelligent Systems*, vol. 28, no. 3, pp. 54-61, May 2013. [Online]. Available: https://ieeexplore.ieee.org/document/6588858

[276] 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

[277] G. Klein et al., “Ten challenges for making automation a ’team player’ in joint human-agent activity,” *IEEE: Intelligent Systems*, vol. 19, no. 6, pp. 91-95, Nov./Dec. 2004. [Online]. Available: http://jeffreymbradshaw.net/publications/17._Team_Players.pdf_1.pdf

[278] W. Lawless, R. Mittu, D. Sofge, and L. Hiatt, “Artificial intelligence, autonomy, and human-machine teams—Interdependence, context, and explainable AI,” *AI Magazine*, vol. 40, no. 3, pp. 5-13, 2019.

[279] “A framework for discussing trust in increasingly autonomous systems,” The MITRE Corporation, June 2017. [Online]. Available: https://www.mitre.org/sites/default/files/publications/17-2432-framework-discussing-trust-increasingly-autonomous-systems.pdf

[280] 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

[281] 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

[282] L. M. Strickhart and H.N.J. Lee, “Show your work: Machine learning explainer tools and their use in artificial intelligence assurance,” The MITRE Corporation, McLean, VA, June 2019, unpublished.

[283] “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

[284] N. D. Sarter and D. D. Woods, “How in the world did I ever get into that mode? Mode error and awareness in supervisory control,” *Human Factors*, vol. 37, pp. 5-19, 1995.

[285] "Virtuous cycle of AI: Build good product, get more users, collect more data, build better product, get more users, collect more data, etc.," in A. Ng, *AI Transformation Playbook: How to Lead Your Company into the AI Era*, Landing AI, Dec. 13, 2018. [Online]. Available: https://landing.ai/ai-transformation-playbook/

[286] J. Whittlestone, A. Alexandrova, R. Nyrup, and S. Cave, "The role and limits of principles in AI ethics: Towards a focus on tensions," presented at AIES '19, Jan. 27–28, 2019, Honolulu, HI, USA. [Online]. Available: https://www.researchgate.net/publication/334378492_The_Role_and_Limits_of_Principles_in_AI_Ethics_Towards_a_Focus_on_Tensions/link/5d269de0a6fdcc2462d41592/download

[287] "Project ExplAIn interim report," U.K. Information Commissioner's Office, 2019. [Online]. Available: https://ico.org.uk/about-the-ico/research-and-reports/project-explain-interim-report/

[288] "Squad X improves situational awareness, coordination for dismounted units," Defense Advanced Research Projects Agency, Nov. 30, 2018. [Online]. Available: https://www.darpa.mil/news-events/2018-11-30a

[289] DARPAtv, "Squad X experimentation exercise," YouTube, July 12, 2019. [Online]. Available; https://www.youtube.com/watch?v=DgM7hbCNMmU

[290] S. J. Freedberg, "Simulating a super brain: Artificial intelligence in wargames," Breaking Defense, April 26, 2019. [Online]. Available: https://breakingdefense.com/2019/04/simulating-a-super-brain-artificial-intelligence-in-wargames/.

[291] B. Jensen, S. Cuomo, and C. Whyte, "Wargaming with Athena: How to make militaries smarter, faster, and more efficient with artificial intelligence," *War on the Rocks*, June 5, 2018. [Online]. Available: https://warontherocks.com/2018/06/wargaming-with-athena-how-to-make-militaries-smarter-faster-and-more-efficient-with-artificial-intelligence/

[292] J. Whittlestone, A. Alexandrova, R. Nyrup, and S. Cave, "The role and limits of principles in AI ethics: Towards a focus on tensions," presented at AIES '19, Jan. 27–28, 2019, Honolulu, HI, USA. [Online]. Available: https://www.researchgate.net/publication/334378492_The_Role_and_Limits_of_Principles_in_AI_Ethics_Towards_a_Focus_on_Tensions/link/5d269de0a6fdcc2462d41592/download

[293] 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

[294] 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

[295] 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

[296] A. Campolo et al., *AI Now Report 2017*. New York, NY, USA: AI Now Institute, 2017. [Online]. Available: https://ainowinstitute.org/AI_Now_2017_Report.pdf

[297] "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

[298] J. C. Newman, "Decision Points in AI Governance," *UC Berkeley Center for Long-Term Cybersecurity,* May 5, 2020. [Online]. Available: https://cltc.berkeley.edu/2020/05/05/decision-points-in-ai-governance/

[299] T. Hagendorff, "The ethics of AI ethics: An evaluation of guidelines," *arXiv.org*, Oct. 11, 2019. [Online]. Available: https://arxiv.org/abs/1903.03425

[300] J. C. Newman, "Decision Points in AI Governance," *UC Berkeley Center for Long-Term Cybersecurity,* May 5, 2020. [Online]. Available: https://cltc.berkeley.edu/2020/05/05/decision-points-in-ai-governance/

[301] 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/

[302] 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

[303] J. McLaughlin, Z. Dorfman, and S. D. Naylor, “Pentagon intelligence employees raise concerns about supporting domestic surveillance amid protests,” *Yahoo News,* June 4, 2020. [Online]. Available: https://news.yahoo.com/pentagon-intelligence-employees-raise-concerns-about-supporting-domestic-surveillance-amid-protests-194906537.html

[304] J. Menn, “Google fires fifth activist employee in three weeks; complaint filed,” *Reuters,* Dec. 17, 2019. [Online]. Available: https://www.reuters.com/article/google-unions/google-fires-fifth-activist-employee-in-three-weeks-complaint-filed-idUSL1N28R02L

[305] A. Palmer, “Amazon employees plan ‘online walkout’ to protest firings and treatment of warehouse workers,” *CNBC,* April 16, 2020. [Online]. Available: https://www.cnbc.com/2020/04/16/amazon-employees-plan-online-walkout-over-firings-work-conditions.html

[306] J. Eidelson and H. Kanu, “Software Startup Accused of Union-Busting Will Pay Ex-Employees,” *Bloomberg,* Nov. 10, 2018. [Online]. Available: https://www.bloomberg.com/news/articles/2018-11-10/software-startup-accused-of-union-busting-will-pay-ex-employees

[307] 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

[308] “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

[309] “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

[310] 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

[311] ENERGY STAR homepage. Accessed on: Jan. 21, 2020. [Online]. Available: https://www.energystar.gov/

[312] 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

[313] “SafeAI.” Accessed April 2, 2020. [Online]. Available: https://www.forhumanity.center/safeai/

[314] A. Campolo et al., *AI Now Report 2017*. New York, NY, USA: AI Now Institute, 2017. [Online]. Available: https://ainowinstitute.org/AI_Now_2017_Report.pdf

[315] 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

[316] 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

[317] Z. C. Lipton, “The doctor just won’t accept that,” *arXiv.org*, Nov. 24, 2017. [Online]. Available: https://arxiv.org/abs/1711.08037

[318] “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

[319] 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

[320] 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/

[321] 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

[322] 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

[323] 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/

[324] 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/

[325] 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

[326] 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

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

[328] "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/

[329] 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/

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

[331] 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

[332] 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*, July 24, 2019. [Online]. Available: https://www.governing.com/topics/public-justice-safety/gov-cities-ban-government-use-facial-recognition.html

[333] "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

[334] 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