MITRE | Center for Data-Driven Policy

FIVE AI FAILS

(AND HOW WE CAN LEARN FROM THEM)

by Jonathan Rotner, Ron Hodge, Lura Danley (Ph.D.)

AUGUST 2021

Executive Summary

In this important new look at how artificial intelligence (AI) can fail in the real world, Mr. Jonathan Rotner, Mr. Ron Hodge, and Dr. Lura Danley investigate several categories of AI failures to determine what lessons planners and programmers can and should draw from them. This paper should be of interest to AI designers and developers along with policymakers and anyone interested in how this exciting, evolving technology plays out in the real world. Some of the lessons Rotner, Hodge, and Danley reach include the realizations that:

Our success with AI hinges on how we learn from failures. Yet planning for 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 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 and decision making. **"Five 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.

For each fail in this paper, the first page 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 page offers one recommendation on practical steps that can be taken, right now, to apply these insights.

Overall, the key lessons from a human-centric mindset regarding AI are:

Developing AI is not just a technical challenge or a "human behavior" issue. It is a multidisciplinary problem, and by including multidisciplinary perspectives, AI planners and programmers can more clearly articulate the design tradeoffs that must be considered when evaluating different priorities and outcomes.

Many AI applications affect 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. A broader set of stakeholders 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.

Our assumptions shape AI and there is no such thing as a neutral, impartial, or unbiased application. Underlying assumptions about the data, model, user behaviors, and environment affect the AI's objectives and outcomes. An AI system can unintentionally replicate and encode values. Given the current composition of the AI development workforce, those values typically represent how young, White, technically oriented, Western men see the world.

Documentation can be a key tool for success. End-users and consumers will want to use good products, and other AI developers may want to repurpose those products for their own domains. To do so appropriately and safely, they will need to know the original intentions, design tradeoffs, and risks and mitigations. Therefore, original developers need to capture their assumptions and tradeoff decisions, and organizations must enable ongoing outreach.

AI is not always the best solution to a problem so accountability must be part of the equation. Oversight, accountability, and enforcement mechanisms can facilitate ethical outcomes and encourage implementation of the previous lessons. The more the AI application could influence people's behavior and livelihoods, the more care is needed.

Contents

Introduction

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."[1]

AI has tremendously valuable applications; for instance, when it promises to translate a person's conversation into another language in real time, more accurately diagnose patients and propose treatments, or take care of the elderly. 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, that mass surveillance leads to imprisonment and suppression of populations, or that self-driving cars have caused deaths, people realize that AI can also lead to real harm. In these cases, the belief in AI's inevitability can elicit terror. AI developers and deployers experience and observe both extremes of this continuum, and everything in between. 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. Although many of these fails and lessons apply to more than AI, collectively they represent the systemic challenges faced by AI developers and practitioners.

AI is different from other technologies in several ways, notably that decisions aren't static, since data and model versions are updated all the time, and 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 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 its reach is vast enough that a single AI with a single programmed objective can scale to affect human decisions at a global level.

We can't build AI in a vacuum. 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. Studying these automated technology case studies can help provide context for understanding today's challenges.

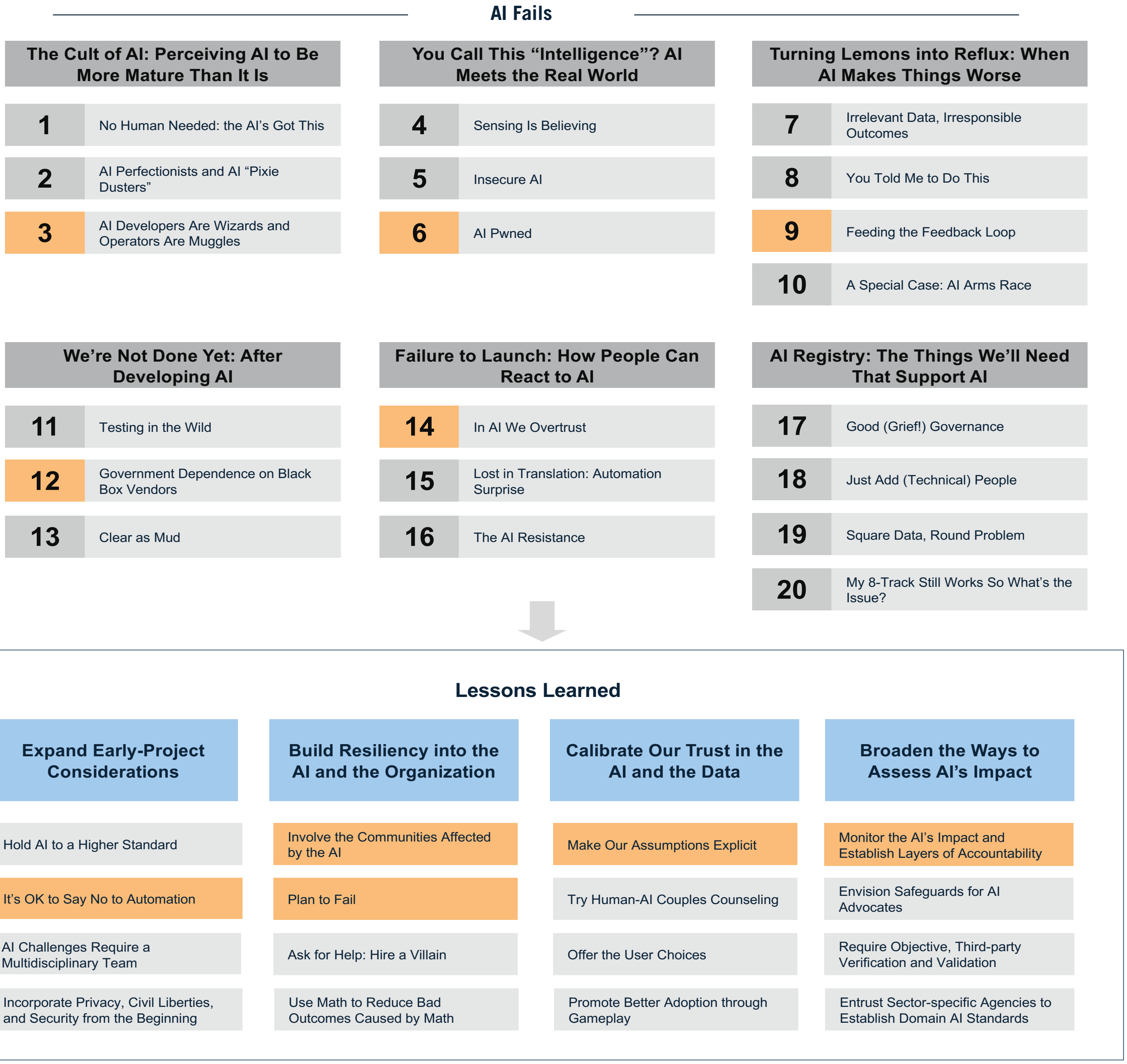
Failures abound. Here's where else to see them

Where have we seen AI fails? How should these fails be viewed from a human-centric perspective? What practical steps that can be taken, right now, to apply these insights?

Stay tuned, and also keep in mind that the five fails and five lessons learned included in this writeup represent a sample set of the many fails included in the full research paper and our website from which this white paper is adapted.

The additional fails demonstrate the impacts of human biases and assumptions, illustrate obstacles resulting from the wrong equipment and an untrained workforce, and characterize how different people can react to AI – with 40 more real-world stories of things gone wrong.

Please dive in and look for ways to learn from others, apply these lessons to your own AI development, and reach out to the authors with your experiences!



AI Fails

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

1 No Human Needed: the AI's Got This
2 AI Perfectionists and AI "Pixie Dusters"
3 AI Developers Are Wizards and Operators Are Muggles

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

4 Sensing Is Believing
5 Insecure AI
6 AI Pwned

Turning Lemons into Reflux: When AI Makes Things Worse

7 Irrelevant Data, Irresponsible Outcomes
8 You Told Me to Do This
9 Feeding the Feedback Loop
10 A Special Case: AI Arms Race

We're Not Done Yet: After Developing AI

11 Testing in the Wild
12 Government Dependence on Black Box Vendors
13 Clear as Mud

Failure to Launch: How People Can React to AI

14 In AI We Overtrust
15 Lost in Translation: Automation Surprise
16 The AI Resistance

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

17 Good (Grief!) Governance
18 Just Add (Technical) People
19 Square Data, Round Problem
20 My 8-Track Still Works So What's the Issue?

Lessons Learned

Expand Early-Project Considerations

- Hold AI to a Higher Standard
- It's OK to Say No to Automation
- AI Challenges Require a Multidisciplinary Team
- Incorporate Privacy, Civil Liberties, and Security from the Beginning

Build Resiliency into the AI and the Organization

- 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

Calibrate Our Trust in the AI and the Data

- Make Our Assumptions Explicit
- Try Human-AI Couples Counseling
- Offer the User Choices
- Promote Better Adoption through Gameplay

Broaden the Ways to Assess AI's Impact

- 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 Domain AI Standards

Fail: In AI We Overtrust

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.

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.[2]

Consumers who received a digital ad said they were more interested in products specifically targeted for them, and even adjusted their own preferences to align with what the ad suggested about them.[3]

Why is this a fail? During the design process developers make conscious and unconscious assumptions about what the AI's goals and priorities should be and which data the AI should learn from. Lots of times, developers' incentives and user incentives align, so this works out wonderfully. 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.[4]

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

1. Humans can assume automation is perfect; therefore, they have high initial trust.[5] This "automation bias" leads users to trust automated and decision-support systems even when that 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.[6] In this case, the AI is perceived as the expert.

3. Humans use mental shortcuts to make sense of complex information, which can lead to overtrusting an AI if it behaves in a way that conforms to expectations, or if it is unclear how the AI works.

Therefore, the more an AI is associated with a supposedly flawless, data-driven authority, the more likely that humans will overtrust the AI. Even professionals in a given field can cede their authority despite their specialized knowledge.[7] Another outcome of overtrust is that the AI reinforces aligning with the model's solution rather than the individual's, pushing AI predictions to become self-fulfilling.[8] Take the many examples of drivers who overrode their own intuition and blindly followed their GPS, including a driver who drove into a body of water and another driver who ran straight into a house![9] 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. See the "Feeding the Feedback Loop" fail for examples.

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 supposedly assess that person's tendencies toward criminality or violence,[10] sexual orientation,[11] and IQ or personality traits.[12] These phrenology and physiognomy products and claims are unethical, irresponsible, and dangerous.

Lesson Learned: Make Our Assumptions Explicit

As developers, we are best positioned to articulate the strengths and weaknesses of our systems, but other perspectives are needed to highlight AI risks and design tradeoffs that we may not have considered. End-users, lawyers, and policymakers (among others) may all have different questions to inform decisions about the AI's appropriate uses, and they offer different considerations for mitigating potential risks. In addition, knowing what's been considered earlier helps new development teams appreciate previous discussions and avoid repeating the same mistakes. Organizations can harness these conversations in two ways:

1. Have the developers fill out standardized templates that capture assumptions and decisions. No one knows the intended and unintended uses for their data and tools better 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," helps document information about the dataset to reduce bias and avoid placing miscalibrated trust in the AI.[13] The second, "model cards for model reporting" clarifies intended use cases and context for the model.[14] Adopting these two templates will go a long way toward helping us achieve transparency, explainability, and accountability in the AI we develop.

2. Structure documentation processes in a way that facilitates proactive and ongoing outreach. The documentation process should prompt us to bring in end-users and representatives of affected communities to ensure they have the information they need and have the opportunity to offer suggestions early enough that developers can incorporate their input. 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. Then, the group as a whole can be proactive in communicating bias and other limitations of systems to potential users.

Checklists aren't enough. But 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.

Read about the other lessons that apply to this fail, at https://sites.mitre.org/aifails

Fail: AI Pwned

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.[15]

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.[16] 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.[17] This kind of attack illustrates that the sensitive information used for training an AI may not be as well protected as desired.

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.[18] Adversaries can also extract sensitive information about individual elements of the training sets[19] or adversaries can make assumptions about which data sources are used and then insert data to bias the learning process.[20]

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[21] and Google Home interpreting a greeting as a command to unlock the front door.[22] 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.[23]

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.[24]

Lesson Learned: Plan to Fail

Benjamin Franklin once said, "If you fail to plan, you are planning to fail."[25] As developers and deployers of AI systems, we must accept that the uncertain and the unexpected are part of reality – and resiliency comes from having ways to *prevent, moderate,* or *recover* from mistakes or failure.[26] Not all resilient methods are technical; they can rely on human participation and partnership. The resiliency needed in an application increases as the AI's success becomes more critical for the overall outcome.

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

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. Many commercial aircraft have, for example, "three flight computers that function independently, with each computer containing three different processors manufactured by different companies."[27] Analog backups, such as old-fashioned paper and pen, can't be hacked or lose power. Finally, experts can proactively familiarize themselves with previous and emerging threats.[28]

Moderate: We should try to include 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."[29] Alternatively, AI projects should make use of human judgment by adding "alerts" both for 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.[30] 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 backups 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.

Read about the other lessons that apply to this fail, at https://sites.mitre.org/aifails

Fail: Feeding the Feedback Loop

When an AI's prediction is geared toward 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 toward 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.

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.[31]

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 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.[32]

Why is this a fail? The scale of AI deployment can result in substantial disruption to 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.[33,34]

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.[35]

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 incendiary or fake articles for newsfeeds,[36] health insurance companies decide which normal behaviors are deemed risky based on recommendations from AI,[37] and governments allocate social services according to AIs that consider only one set of factors.[38]

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.[39]

Lesson Learned: Monitor the AI's Impact and Establish Layers of Accountability

Those of us who design AI systems have the best of intentions. Yet the reality is that after we deploy an AI, the data, the environment, or how users interact with the AI will change, and the algorithm will work in unexpected ways.

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

When weighing 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. To act quickly against unanticipated outcomes, organizations should:

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. If the AI "drifts" enough from its baseline, we may have to retrain or even scrap the model. Baseline criteria should be both mathematical and contextual, and criteria should include the perspectives of all affected stakeholders.

In parallel, risk assessment criteria should guide decisions about the AI's suitability to an application domain. We should set guidance for higher stakes cases, when legality or ethics may lead to concern.

2. Regularly monitor the AI's impact and require prompt fixes. 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.[40] This will not only help refine the model, but also help us act promptly as harms or biases emerge.

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 the feedback team. In addition, this feedback team can be proactive. The team should broadcast how an individual's data is used and implement processes for discarding old data.[41] In one example for others to model, Google set up an email address and actively guided other researchers looking to build off their work in the case of an open-source application with potential harmful outcomes.[42]

4. Experiment with different accountability methods. Accountability that works well today may not be equally effective as future technologies emerge or an organization's structure and culture evolve.[43] So experimentation can help us find the right mix.

One example comes from Microsoft, which established an AI, Ethics and Effects in Engineering and Research (AETHER) Committee in 2018. Microsoft required direct participation by senior leadership. Microsoft asked employees with different backgrounds to work with their legal team to develop policy and governance structures. The committee also set up an "Ask AETHER" phone line for employees to raise concerns.[44]

AI has real consequences and is certain to continue to produce unintended outcomes. That is why we must do our best to position our organizations to be proactive against, and responsive to, undesirable outcomes.

Read about the other lessons that apply to this fail, at https://sites.mitre.org/aifails

Fail: Government Dependence on Black Box Vendors

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.

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 recommendations.[45,46]

The Houston Independent School District implemented an AI to measure teachers' performances by comparing their students' test scores to the statewide average. The teachers' union won a lawsuit, arguing that the proprietary nature of the product prevents teachers from verifying the results, thereby violating their Fourteenth Amendment right to due process.[47]

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.

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.

Lesson Learned: It's OK to Say No to Automation

As developers and deployers of AI systems, the first things we should ask when starting an AI project are 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.[48]

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

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:

- 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.[49]

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.

Read about the other lessons that apply to this fail, at https://sites.mitre.org/aifails

Fail: When AI Developers Are Wizards and Operators Are Muggles

When AI developers think we know how to solve a problem, we may overlook input from 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 the users 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.

Examples:

After one of the Boeing 737 MAX aircraft crashes, pilots were furious they had not been told 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.[50,51]

Uber's self-driving car was not programmed to recognize jaywalking, only pedestrians crossing in or near a crosswalk.[52] This programming would be acceptable in some areas of the country but runs counter to the norms in others, putting those pedestrians in danger.

Why is this a fail? It's a natural inclination to assume that end-users will act the same way programmers do or will want the same results. 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 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. We need 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.

What happens when things fail? 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 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, communities that have been heavily policed understand the deep historical patterns of abuse and profiling that can result, regardless of technology.

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.

As Betty Medsger, investigative reporter, writes, “being Black was enough [to justify surveillance].”[53] 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.[54]

Lesson Learned: Involve the Communities Affected by the AI

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.

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 the AI will affect. Therefore, we should be sure to include representatives from the communities 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.[55]

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.[56] It includes actions that can help organizations identify appropriate stakeholder groups, run group sessions, and close the loop between developers and the invited communities.

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.[57,58] 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.[59,60]

Read about the other lessons that apply to this fail, at https://sites.mitre.org/aifails

Conclusion

Given the increasing integration of AI-enabled systems into most areas of daily life, we developers, designers, and policymakers 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 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.

Endnotes

1. 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
2. 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. https://www.cc.gatech.edu/~alanwags/pubs/Robinette-HRI-2016.pdf
3. M. Heid, “The unsettling ways tech is changing your personal reality,” Elemental, Oct. 3, 2019. https://elemental.medium.com/technology-is-fundamentally-changing-the-ways-you-think-and-feel-b4bbfdefc2ee
4. “Algorithms and artificial intelligence: CNIL’s report on the ethical issues,” CNIL [Commission Nationale de l’Informatique et des Libertés], May 25, 2018. https://www.cnil.fr/en/algorithms-and-artificial-intelligence-cnils-report-ethical-issues
5. 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.
6. “Appeal to authority,” Legally Fallacious. Accessed March 25, 2020. https://www.logicallyfallacious.com/logicalfallacies/Appeal-to-Authority
7. 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. https://datasociety.net/wp-content/uploads/2018/04/Data_Society_Algorithmic_Accountability_Primer_FINAL-4.pdf
8. “Algorithms and artificial intelligence: CNIL’s report on the ethical issues,” CNIL [Commission Nationale de l’Informatique et des Libertés], May 25, 2018. https://www.cnil.fr/en/algorithms-and-artificial-intelligence-cnils-report-ethical-issues
9. L. Hansen, ”8 drivers who blindly followed their GPS into disaster,” The Week, May 7, 2013. https://theweek.com/articles/464674/8-drivers-who-blindly-followed-gps-into-disaster
10. B. Aguera y Arcas, “Physiognomy’s new clothes,” Medium, May 6, 2017. https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a
11. S. Levin, “New AI can guess whether you’re gay or straight from a photograph,” Guardian, Sept. 7, 2017. https://www.theguardian.com/technology/2017/sep/07/new-artificial-intelligence-can-tell-whether-youre-gay-or-straight-from-a-photograph
12. Synced, “2018 in review: 10 AI failures,” Medium, Dec. 10, 2018. https://medium.com/syncedreview/2018-in-review-10-ai-failures-c18faadf5983
13. T. Gebru et al., “Datasheets for datasets,” arXiv.org, Jan. 14, 2020. https://arxiv.org/abs/1803.09010
14. M. Mitchell et al., “Model cards for model reporting,” arXiv.org, Jan. 14, 2019. https://arxiv.org/abs/1810.03993
15. “pwned,” Urban Dictionary. Accessed on: March 11, 2020. https://www.urbandictionary.com/define.php?term=pwned
16. M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, “A general framework for adversarial examples with objectives,” arXiv.org, April 4, 2019. https://arxiv.org/abs/1801.00349
17. M. Fredrikson, S. Jha, and 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. https://www.cs.cmu.edu/~mfredrik/papers/fjr2015ccs.pdf
18. M. James, “Adversarial attacks on voice input,” I Programmer, Jan. 31, 2018. https://www.i-programmer.info/news/105-artificial-intelligence/11515-adversarial-attacks-on-voice-input.html

19. G. Ateniese et al., “Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers,” arXiv.org, June 19, 2013. https://arxiv.org/abs/1306.4447

20. A. Polyakov, “How to attack machine learning (evasion, poisoning, inference, Trojans, backdoors),” towards data science, August 6, 2019. https://towardsdatascience.com/how-to-attack-machine-learning-evasion-poisoning-inference-trojans-backdoors-a7cb5832595c

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

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

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

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

25. “Benjamin Franklin quotable quote,” Goodreads. Accessed March 16, 2020. https://www.goodreads.com/quotes/460142-if-you-fail-to-plan-you-are-planning-to-fail

26. 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. http://www.jeffreymbradshaw.net/publications/56.%20Human-Robot%20Teamwork_IEEE%20IS-2014.pdf

27. 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. https://www.seattletimes.com/business/boeing-aerospace/a-lack-of-redundancies-on-737-max-system-has-baffled-even-those-who-worked-on-the-jet/

28. “Adversarial ML Threat Matrix,” GitHub. Accessed March 19, 2021. [Online]. Available: https://github.com/mitre/advmlthreatmatrix

29. E. Lacey, “The toxic potential of YouTube’s feedback loop,” Wired, July 13, 2019. https://www.wired.com/story/the-toxic-potential-of-youtubes-feedback-loop/

30. D. Amodei, “Concrete problems in AI safety,” arXiv.org, July 25, 2016. https://arxiv.org/pdf/1606.06565.pdf

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

32. E. Lacey, “The toxic potential of YouTube’s feedback loop,” Wired, July 13, 2019. https://www.wired.com/story/the-toxic-potential-of-youtubes-feedback-loop/

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

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

35. M. Whittaker et al., AI Now Report 2018. New York, NY, USA: AI Now Institute, 2018. https://ainowinstitute.org/AI_Now_2018_Report.pdf

36. W. Oremus, “Who controls your Facebook feed,” Slate, Jan. 3, 2016. http://www.slate.com/articles/technology/cover_story/2016/01/how_facebook_s_news_feed_algorithm_works.html

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

38. 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. 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/

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

40. A. Gonfalonieri, “Why machine learning models degrade in production,” towards data science, July 25, 2019. https://towardsdatascience.com/why-machine-learning-models-degrade-in-production-d0f2108e9214

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

42. J. C. Newman, “Decision points in AI governance,” UC Berkeley Center for Long-Term Cybersecurity, May 5, 2020. https://cltc.berkeley.edu/2020/05/05/decision-points-in-ai-governance/

43. T. Hagendorff, “The ethics of AI ethics: An evaluation of guidelines,” arXiv.org, Oct. 11, 2019. https://arxiv.org/abs/1903.03425

44. J. C. Newman, “Decision points in AI governance,” UC Berkeley Center for Long-Term Cybersecurity, May 5, 2020. https://cltc.berkeley.edu/2020/05/05/decision-points-in-ai-governance/

45. 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. 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

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

47. C. Langford, “Houston schools must face teacher evaluation lawsuit,” Courthouse News Service, May 8, 2017. https://www.courthousenews.com/houston-schools-must-face-teacher-evaluation-lawsuit/

48. 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. http://www.jeffreymbradshaw.net/publications/56.%20Human-Robot%20Teamwork_IEEE%20IS-2014.pdf

49. “Ethics & algorithms toolkit.” Accessed March 13, 2020. http://ethicstoolkit.ai/

50. 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. 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

51. A. MacGillis, “The case against Boeing,” New Yorker, Nov. 11, 2019. https://www.newyorker.com/magazine/2019/11/18/the-case-against-boeing

52. P. McCausland, “Self-driving Uber car that hit and killed woman did not recognize that pedestrians jaywalk,“ NBC News, Nov. 9, 2019. https://www.nbcnews.com/tech/tech-news/self-driving-uber-car-hit-killed-woman-did-not-recognize-n1079281

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

54. “Barry Friedman: Is technology making police better—or…” Recode Decode podcast, Nov. 24, 2019. https://us.radiocut.fm/podcast-episode/barry-friedman-is-technology-mak/

55. M. Whittaker et al., AI Now Report 2018. New York, NY, USA: AI Now Institute, 2018. https://ainowinstitute.org/AI_Now_2018_Report.pdf

56. “Diverse Voices: A How-To Guide for Facilitating Inclusiveness in Tech Policy.” Accessed April 8, 2020. https://techpolicylab.uw.edu/project/diverse-voices/

57. A. Campolo et al., AI Now Report 2017. New York, NY, USA: AI Now Institute, 2017. https://ainowinstitute.org/AI_Now_2017_Report.pdf
58. M. Whittaker et al., AI Now Report 2018. New York, NY, USA: AI Now Institute, 2018. https://ainowinstitute.org/AI_Now_2018_Report.pdf
59. A. Campolo et al., AI Now Report 2017. New York, NY, USA: AI Now Institute, 2017. https://ainowinstitute.org/AI_Now_2017_Report.pdf
60. M. Whittaker et al., AI Now Report 2018. New York, NY, USA: AI Now Institute, 2018. https://ainowinstitute.org/AI_Now_2018_Report.pdf

About the Authors

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.

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.

Lura Danley (Ph.D.) 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.

For more information about this paper or the Center for Data-Driven Policy, contact policy@mitre.org

MITRE's Mission

MITRE's mission-driven teams are dedicated to solving problems for a safer world. Through our public-private partnerships and federally funded R&D centers, we work across government and in partnership with industry to tackle challenges to the safety, stability, and well-being of our nation.

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®