July 2021

AI Accidents: An Emerging Threat

What Could Happen and What to Do

CSET Policy Brief

CSET
CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

AUTHORS
Zachary Arnold
Helen Toner

Executive Summary

Modern machine learning is powerful in many ways, but profoundly fragile in others. Because of this fragility, even the most advanced artificial intelligence tools can unpredictably fail, potentially crippling the systems in which they are embedded. As machine learning becomes part of critical, real-world systems, from cars and planes to financial markets, power plants, hospitals, and weapons platforms, the potential human, economic, and political costs of AI accidents will continue to grow.

Policymakers can help reduce these risks. To support their efforts, this brief explains how AI accidents can occur and what they are likely to look like "in the wild." Using hypothetical scenarios involving AI capabilities that already exist or soon will, we explain three basic types of AI failures—robustness failures, specification failures, and assurance failures—and highlight factors that make them more likely to occur, such as fast-paced operation, system complexity, and competitive pressure. Finally, we propose a set of initial policy actions to reduce the risks of AI accidents, make AI tools more trustworthy and socially beneficial, and support a safer, richer, and healthier AI future. Policymakers should:

- **Facilitate information sharing about AI accidents and near misses**, working with the private sector to build a common base of knowledge on when and how AI fails.

- **Invest in AI safety research and development (R&D)**, a critical but currently underfunded area.

- **Invest in AI standards development and testing capacity**, which will help develop the basic concepts and resources needed to ensure AI systems are safe and reliable.

- **Work across borders to reduce accident risks**, including through R&D alliances and intergovernmental organizations.

Table of Contents

1. What are AI accidents?

We are on the threshold of a new industrial revolution. Artificial intelligence—the capability of machines to reason, communicate, and make decisions as only humans could before—will be at its center.[1] Technical achievements over the past several years, especially in the machine learning subfield of AI, have produced vastly more powerful AI systems.[2] Advances in complementary fields, such as robotics and networking, are unlocking new real-world applications for these systems, from autonomous fighter jets,[3] to fiction written by computers,[4] to novel, AI-optimized medicines.[5]

In the years to come, AI is expected to pervade our lives, much like electricity in the twentieth century and the internet in the twenty-first.[6] Today, we are at the very beginning of this process. Thomas Edison received the patent on his light bulb in 1880; it took until 1925 to electrify half of U.S. homes.[7]

Deploying AI is an ongoing process that holds tremendous promise—and equally tremendous danger. Today's cutting-edge AI systems are powerful in many ways, but profoundly fragile in others. They often lack any semblance of common sense, can be easily fooled or corrupted, and fail in unexpected and unpredictable ways. It is often difficult or impossible to understand why they act the way they do.[8]

Despite these problems, AI systems are becoming integrated into the real world at a pace that is only expected to accelerate in the next decade.[9] These systems may be fragile, but as companies, governments, and militaries decide when and how to deploy them, their huge potential benefits will often overshadow uncertain risks. Leaders in these organizations also may not be fully aware of these risks, and may face pressure from competitors willing to move quickly.[10] To be sure, some industries are already deploying AI much faster than others, and a few sensitive sectors may remain "walled off" for some time.[11] But eventually, the powerful incentives driving the spread of AI today are likely to make it pervasive. As our economy, security, and health become more and more dependent on AI systems, these systems' fragilities will put lives at stake.

Today, many are worried about AI being misused *intentionally*. An adversary could attack with swarms of drones; authoritarian governments are already using AI algorithms to discriminate on the basis of race or ideology. These risks are real, and they deserve attention.

But unintended, *accidental* AI disasters are also an urgent concern. AI-related accidents are already making headlines, from inaccurate facial recognition systems causing false arrests to unexpected racial and gender discrimination by machine learning software.[12] This is especially striking since AI has so far mostly been deployed in seemingly lower-stakes settings, such as newsfeed rankings, ad targeting, and speech recognition, with less deployment in higher-stakes areas such as autonomous driving.

Despite these initial accidents, governments, businesses, and militaries are preparing to use today's flawed, fragile AI technologies in critical systems around the world. Future versions of AI technology may be less accident-prone, but there is no guarantee—and regardless, if rollout continues as expected, prior versions of the technology may already have been deployed at massive scale. The machine learning models of 2020 could easily still be in use decades in the future, just as airlines, stock exchanges, and federal agencies still rely today on COBOL, a programming language first deployed in 1960.[13]

In retrospect, even the most extreme technological accidents, from the *Challenger* disaster to the meltdown at Chernobyl, can seem both predictable and preventable.[14] History is full of accidents that seem obvious in retrospect, but "no one could have seen coming" at the time. In other cases, known risks are brushed aside, or obvious fixes go unmade. Unless we act, there is no reason to think that the advent of AI will be any different.

In fact, there are reasons to think AI could cause *more* accidents than other technologies that have caused high-profile disasters. Unlike the space shuttle or nuclear power plants, for example, AI will be pervasive throughout society, creating endless opportunities for things to go awry. What's more, modern AI is so good at some tasks that even sophisticated users and developers can come to trust it implicitly.[15] This degree of trust, placed in pervasive, fallible systems without any common sense, could have terrible consequences.

To avoid these consequences, we first need to understand how AI can unexpectedly fail and what the real-world effects could be. In Section 2, we define several potential types of AI accidents, using hypothetical scenarios to illustrate how each type might play out in the real world. These scenarios are fictional, but plausible. In most cases, they are based on incidents that have already happened in the real world, and all of them involve AI technology that already exists or likely soon will. The exact scenarios we describe may or may not ever actually occur, but we should expect many like them to unfold in the coming years.

Today, the risk of large-scale, real-world AI accidents may seem hypothetical. But if we wait until AI is pervasive before we try to understand and address this risk, it will be too late. Policymakers can do a great deal—right now—to help ensure that tomorrow's AI-enabled society is safe and secure. To help speed these efforts, in Section 3, we identify risk factors that make AI accidents more likely, and in Section 4, we highlight initial actions for U.S. policymakers concerned about AI accidents. These measures would not only reduce accident risks, but also help make AI tools more trustworthy and socially beneficial, supporting a safer, richer, and healthier future.

2. What could AI accidents look like?

When AI systems unexpectedly fail, the failure often fits into one of the following categories:

- Failures of **robustness**: the system receives abnormal or unexpected inputs that cause it to malfunction.

- Failures of **specification**: the system is trying to achieve something subtly different from what the designer or operator intended, leading to unexpected behaviors or side effects.

- Failures of **assurance:** the system cannot be adequately monitored or controlled during operation.

In this section, we briefly explain each of these types of failure, and describe how they might unfold in the real world.

For a more detailed overview of robustness, specification, and assurance, see CSET's Key Concepts in AI Safety.

Robustness

If a system malfunction could cause serious harm, we want that system—and each of its components—to work reliably under a wide range of circumstances. The field of reliability engineering has a long history of ensuring that nuclear facilities, chemical plants, and other safety-critical systems continue to operate safely under unusual conditions ranging from sensor failures to natural disasters.

AI "robustness" is referring to the same basic concept: will the system still function as intended under unexpected or unfamiliar circumstances? Systems involving AI can make use of many of the basic concepts and principles of reliability engineering, but they also face new challenges.

For example, AI systems are especially likely to malfunction when used in contexts that differ systematically—even if subtly—from what they were designed for, or when given inputs different from those used in "training." This is called "distributional shift," referring to a change in the types of data the system is given.

> **Cancer detector misdiagnoses Black users:** *A new smartphone app uses your phone camera to identify early-stage signs of skin cancer, with highly accurate results in the developer's field tests. Millions of Americans download and use the app to decide whether to consult their doctors about potentially concerning symptoms. A few years later, public health researchers detect a sharp upward trend in late-stage skin cancer diagnoses among Black patients, corresponding to thousands of additional diagnoses and hundreds of deaths. An inquiry reveals that the self-screening app was trained and field-tested mainly on data from northern Europe, and is much less accurate at detecting cancers on dark skin tones.*[16]
>
> **Bus ad triggers facial recognition system:** *To improve safety and boost public trust in its new driverless iTaxis, IntelliMotor designs the vehicles' AI-based vision system to recognize human faces within a short distance of the windshield. If a face is detected with high certainty, the iTaxi automatically decelerates to minimize harm to the human. To prove it works, several of the engineers step in front of speeding iTaxis on the IntelliMotor test range—the cars brake, and the engineers are unharmed.*
>
> *IntelliMotor pushes a software update with the new facial recognition capability to all deployed iTaxis. Meanwhile, in several U.S. cities, city buses are plastered with ads for Bruce Springsteen's upcoming concert tour. The updated iTaxis identify the Boss's printed face as a nearby pedestrian and begin stopping short whenever they come near buses, quickly causing thousands of collisions across the country.*[17]
>
> **Phantom missile launches:** *In missile defense, seconds of delay can spell the difference between an interception and a*

> *miss. U.S. Strategic Command's new missile defense system, Global Eye, eliminates delay by scanning gigabytes of real-time data every second. If the system's algorithms detect a missile launch with high certainty, the system can quickly and autonomously trigger an interceptor launch in order to shoot down the missile.*
>
> *One day, unusual atmospheric conditions over the Bering Strait create an unusual glare on the horizon. Global Eye's visual processing algorithms interpret the glare as a series of missile launches, and the system fires interceptors in response. As the interceptors reach the stratosphere, China's early-warning radar picks them up. Believing they are under attack, Chinese commanders order a retaliatory strike.*[18]

Modern AI systems are also very sensitive to interference with their inputs; even small problems with the data fed to a system can, in some cases, completely throw off the results.[19]

> **AI-driven blackouts, part 1:** *Enercorp, a large public utility, uses OptiVolt demand-response software to dispatch power from its generating stations. The software continuously collects a wide range of real-time data, from weather forecasts to macroeconomic trends, and processes it through a machine learning model trained on years of archived real-world energy market data. Based on this history and the processed real-time data, OptiVolt predicts energy prices and expected profits on a minute-by-minute basis. When expected profits are high, OptiVolt autonomously "spins up" the plants it projects will be best able to meet demand; when expected profits fall below zero, plants are automatically "spun down."*
>
> *One day, during a routine debugging exercise, a software engineer at the regional grid operator accidentally introduces simulated data showing a massive oversupply of energy on the grid into a public feed monitored by Enercorp's software. Before the mistake can be corrected a few seconds later, OptiVolt has already triggered a spin-down of several major plants, leading to a region-wide blackout.*[20]

Chemical controller fails in windstorm: *After a series of highly publicized operator errors at its Cleveland plant, OxyCorp, a chemical manufacturer, installs a software-based control system to prevent accidental releases of toxic substances. The system relies on a machine learning model trained on millions of hours of operating data from OxyCorp's facilities. Using sensor data from the plant, the model can identify when it is safe to open the plant's exhaust vents. Thanks to its extensive "experience," the model adapts seamlessly to process changes and physical modifications within the complex plant, which were blamed for confusing human operators in the past. The new software system proves highly reliable and becomes a trusted tool within OxyCorp.*

Months later, a windstorm disrupts several of the plant's sensors. Based on the flawed sensor input, the control system continues to read "safe," and the plant operators act accordingly, leaving the vents open, even as managers elsewhere in the plant begin an unscheduled production run in response to an urgent customer request. The run produces a cloud of lethal chlorine gas, which escapes through the open exhaust vents and drifts toward downtown.

In many cases, bad actors may even be able to exploit AI systems' data sensitivity, by introducing "adversarial" data inputs designed to cause havoc. In one well-known study, for example, researchers tricked state-of-the-art computer vision systems into ignoring stop signs by applying a few small stickers to the signs.[21]

Insurgents trick targeting system: *U.S. Air Force software engineers create Elendil, a targeting assistance system built on state-of-the-art image recognition technology and annotated drone footage. Elendil processes gigabytes of overhead imagery per second, identifying enemy vehicles with much greater accuracy and at far greater speed than human analysts. The Air Force deploys Elendil in a combat zone during a period of high alert. Intelligence suggests that an insurgent leader is planning to move to a new safe house in the near future.*

> *Unknown to the Air Force, the insurgent group has stolen a copy of the Elendil source code from a contractor's server. They use the code to develop "adversarial" graphics that Elendil will reliably identify as enemy and non-enemy. They paint the "non-enemy" graphics on their own vehicles' roofs, and paint "enemy" graphics on several public school buses parked in a poorly secured lot nearby. The next morning, the insurgent leader's convoy moves out as the buses make their morning rounds, triggering Elendil alerts. In the heat of the moment, Air Force targeting analysts order strikes on several of the buses, killing 140 schoolchildren; meanwhile, the convoy slips away undetected.*[22]

For more information on "adversarial attacks" on AI systems, see CSET's introduction to robustness and adversarial examples.

Specification

Machine learning systems implement the instructions their designers provide: for example, "score as many points as possible," "identify which photos have cats in them," "predict which word will occur next in this sentence." This is accomplished by specifying a rule that captures what the AI system is supposed to do. For example, in the case of "identify which photos have cats in them" the rule could be "minimize the number of photos incorrectly labeled as 'cat.'" Specification problems arise when there is no simple rule or instruction that captures all of what we want an AI system to do.

Trickiest of all are cases where it *seems* like there is a rule that captures what we want, but in fact that rule only partially captures what we care about. As one prominent researcher has put it, AI responds like "the genie in the lamp [...] you get exactly what you ask for, not what you want."[23] For example, social media platforms use algorithms to recommend engaging content to users, hoping to maximize the users' entertainment and boost revenue. Sadly, conspiracy theories, hate speech, and other noxious types of content are highly engaging to many users, so the algorithms will heavily recommend them if left unchecked. To fix this, the platforms have had to continually add emergency updates and patches.[24] A human employee could have inferred that the goal of “maximizing engagement” did not justify promoting illegal or harmful content, but a machine learning system can only follow the rules it is given.

> **AI-driven blackouts, part 2:** *A year after the OptiVolt blackout, Enercorp has deployed a new version of the software—this time, with new processes added to detect and discard obviously erroneous data inputs. A few months later, the nation experiences an unprecedented heat wave. As air conditioners, fans, and freezers work overtime across the country, wholesale electricity prices skyrocket. OptiVolt2 tirelessly ramps up, throttles, and shifts generation capacity across Enercorp’s plants in order to maintain steady production and exploit local variations in prices, earning the company huge profits. But as the heat wave wears on, OptiVolt2’s rapid-fire production commands stress Enercorp’s turbines past their physical limits; the software’s algorithm does not account for wear and tear on the equipment. During the fifth week of hot weather, dozens of turbines fail, destabilizing the power grid and triggering another wave of blackouts.*

> **Wall of fire:** *Summer brings wildfires to the Los Angeles area, forcing evacuations along Interstate 15. One morning, a truck overturns on the freeway, blocking all northbound lanes. Navigation apps detect low traffic on nearby side roads and begin redirecting drivers accordingly. Unfortunately, these roads are empty because the surrounding neighborhoods have been evacuated; the apps’ routing algorithms do not take fire safety conditions into account. As traffic fills the side roads, the wind picks up. Wildfire quickly spreads into the evacuated area, trapping the rerouted vehicles in the flames.*[25]

A related type of specification problem is known as *reward hacking:* when an AI system finds a way to meet the exact objective as specified, but in a way that totally misses the actual goal intended. In other words, it optimizes for the letter of the law rather than the spirit. Researchers have observed this behavior in the lab over and over again—from a boat in a video game learning to set itself on fire to earn points, to a robot learning to trick its human monitor into thinking it was succeeding at its task.[26] Hopefully, most failures of this kind will be identified in testing and fixed before entering real-world use. But even the most rigorous testing cannot possibly anticipate all of the ways AI systems might misinterpret instructions. As these systems become more and more common in society, and are exposed to an ever wider range of operating conditions, even the rarest potential malfunctions are bound to occur.

> ***Microelectronic meltdown:*** *ChipCorp's new software, Optimizr, uses reinforcement learning to optimize production at its computer chip factory. Instead of giving precise commands—"speed up Belt 4 if a backlog exists at the assembly stage"—like their previous software required, plant managers can give Optimizr high-level goals, such as "improve energy efficiency." Running thousands of simulations a minute based on plant schematics and sensor data, the software itself identifies the best ways to achieve these objectives, then implements them through network interfaces with the plant equipment.*
>
> *Shortly after being activated in production for the first time, Optimizr sends an unusual series of instructions to the facility's six lithography machines, each valued over $100 million. The commands trigger a previously unknown mechanical flaw, and the machines overheat, melting down ChipCorp's production capacity in a matter of minutes. The company goes bankrupt. Forensic analysis later reveals that Optimizr sent the fatal commands after being programmed to "minimize unplanned outages on the packaging line this quarter." Apparently, the system "reasoned" that if the lithography machines were destroyed by overheating, they would not produce any chips to package, and the packaging line would never start up—eliminating any chance of an unplanned outage.*[27]

Assurance

As with other technologies, we need to be sure AI systems deployed in high-stakes settings are acting safely, and will continue to act safely in the future. Unfortunately, at present, it is difficult or even impossible for us to keep track of the workings of AI systems and how they could malfunction.

For many older types of automated systems, engineers use exhaustive testing or mathematical analysis to “validate” that the system will behave within reasonable bounds. But modern AI systems are far more complex than older generations of automated systems, with millions or billions of calculations behind every action. As such, they cannot be exhaustively tested like older systems—there are simply too many possibilities to test. AI models used in industry are often partially validated, with a small sample of decisions manually checked for accuracy, but while a sampling approach might ensure the system behaves acceptably on average, it cannot give us confidence about extreme cases.

As an alternative, or in addition, if we could determine *why* AI systems operate as they do, we could anticipate how they will act in particular situations. This would help us identify and address their problems before they cause real-world consequences. Unfortunately, it is currently extremely difficult for us to understand what is behind modern AI systems’ actions. Machine learning algorithms do not “reason” like humans, and their inner workings often cannot be explained in the familiar terms of logic and motivation.[28] This “black box” problem, sometimes referred to as the problem of AI interpretability or explainability, is currently the subject of a great deal of academic research. However, practical solutions are still far off, and in some cases, they may never be found.[29]

Finally, even if the inner workings of modern AI systems can be deciphered, the systems need to be designed to clearly and consistently communicate this information to their human monitors. Even with simpler systems, it has proven difficult to design user interfaces that allow humans to effectively monitor and intervene.

Poor interface design has been blamed for incidents from U.S. Navy ship collisions to airplane crashes.[30]

For more information on the problem of explaining AI decisions, see CSET's introduction to interpretability in machine learning.

Other barriers are psychological. Despite their flaws and limitations, AI systems are incredibly effective at some tasks. As human operators interact with AI systems whose workings they do not understand, but that seem to work reliably, many come to trust them implicitly—even in situations the systems were not designed for. In turn, they stop carefully monitoring the systems, or do not intervene even when they notice something that does not look right. This pattern has been documented again and again in the real world.[31]

> ***AI fails on the high seas:*** *Morsen Shipping Lines installs a new computer vision system on its tankers. In low-visibility settings, the system can pick out obstructions and oncoming vessels with superhuman speed and accuracy. One foggy night, for reasons Morsen's technical teams are still working to understand, the vision system on one tanker fails to sound alarms as the ship approaches semi-submerged debris off the Florida coast. (Normally, a crew member would be keeping watch as an extra precaution, but since the computer vision system is so effective, captains have started skipping this extra precaution from time to time.) Relying on the system, the tanker's captain maintains course. The debris tears a gash in the ship's hull, spilling carcinogenic chemicals.*[32]
>
> **Ambulance chaos:** *Faced with a surge of emergency room visits during an unusually bad flu season, New York City's hospitals turn to Routr, a machine learning platform. Reading data from first responders, public health agencies, and member hospitals in real time, Routr redirects incoming 911 calls from hospitals that could fill up soon to hospitals*

> *that are likely to have enough room. The software is based on AI algorithms that have been "trained" on terabytes of historical occupancy data, allowing them to identify patterns that no human could have recognized.*
>
> *Thanks to Routr, during November and December, city hospitals have beds to spare even as cases skyrocket. However, when the clock turns over to a new year on January 1, the software inexplicably begins routing calls throughout the city to only a few hospitals in Queens. By morning, the hospitals are overwhelmed—and in ambulances outside the hospital entrances, patients are suffering, and in some cases dying, in snarled traffic.*
>
> *Months later, a state-ordered investigation finds, among other lapses, that human dispatchers monitoring Routr were aware of the unusual routing pattern on New Year's Eve as it unfolded, but they did not intervene. In an interview, one dispatcher explained that "the system had made weird decisions before that always turned out to be genius…We didn't know exactly what was going on, but we just figured the AI knew what it was doing."*[33]

Finally, even when a human wants to intervene, it may not be possible. AI systems often make and execute decisions in microseconds, far faster than any human in the loop can act. In other cases, the system's user interface may make intervening difficult. An AI system might even actively resist being controlled, whether by design or as a strategy "learned" by the system itself during training.

> **Autopilot fights back:** *On descent into Dallas, faulty wiring triggers a glitch in Flight 77's heading indicator system. The plane's recently upgraded autopilot system, which is in control of the landing process, banks hard in response. The pilots pull back on the control wheel, but it is not enough—in these situations, the autopilot's "smart stabilization" feature modulates sudden inputs from the wheel in order to avoid destabilizing the plane. A few miles short of the runway, the plane crashes into a hotel, killing hundreds.*[34]

3. When are AI accidents more likely?

The destructive scenarios we have described are hypothetical—for now. Even in these early days of AI adoption and deployment, accidents involving AI systems are already widespread. To cite just a few publicly reported examples:

- Self-driving cars have been involved in crashes across the United States, with problems in the cars' AI software blamed in several cases.[35]

- Left to their own devices, algorithms built into popular social media platforms have unexpectedly boosted disturbing and harmful content, contributing to violence and other serious harm in the real world.[36]

- People have been wrongly arrested based on "false positive" identifications by police facial recognition systems.[37]

- An algorithm used by many hospitals to identify high-risk patients was found to be racially biased, meaning that patients of color in those hospitals may have received worse care.[38]

As AI is integrated into more and more critical systems, the dangers of AI accidents will grow. In practice, we expect these accidents will be more likely and more severe in some situations than in others. Identifying these risky situations ahead of time is challenging, but based on AI accidents that have already occurred and historical accidents involving other technologies, we expect risk factors for severe AI accidents will include:

- **Competitive pressure**. When *not* using AI could mean falling behind competitors or losing profits, companies, militaries, and governments are more likely to deploy buggy AI systems, use them in reckless ways, or cut corners on testing and operator training.[39] The infamous Boeing 737 MAX, though it does not use machine learning, is an example of this dynamic. The aircraft was developed, tested, and certified under extreme time pressure, aiming to compete with a comparable Airbus system.[40] Ultimately, this haste led to two crashed planes and hundreds of deaths.

- **System complexity.** When AI is integrated into a system in which many components depend on each other in opaque ways, the AI's flaws or unexpected behaviors will have "ripple effects" throughout the system—with unpredictable and possibly catastrophic results. In such complex systems—for example, a complicated industrial machine with thousands of interacting sensors, some of which are AI-enabled—it is also harder to detect AI-related errors as they occur, much less understand and address their causes.[41]

- **Systems that operate too quickly for human intervention.** Modern AI can operate at superhuman speed, and the systems built around it are often designed to turn that speed into split-second action. In the event of an AI glitch, these systems could cause severe real-world harm before human operators even realize that there is a problem.[42]

- **Untrained or distracted users.** For the end user, modern AI systems can seem deceptively simple. In many cases, users can even interact with the systems through a straightforward question-and-answer interface.[43] But no matter how streamlined the interface, modern AI systems are complex, error-prone tools. Applying them safely and effectively in safety-critical environments requires just as much training as other complex technologies do. Untrained users may trust the system too much, unaware of its weaknesses and biases; when the system goes wrong, they may not recognize the problem or know how to fix it. When a system is used routinely for long stretches of time, even well-trained users' eyes are likely to begin to glaze over at some point, further increasing the chance of mistakes.[44]

- **Systems with many instances.** When a single AI model is used in many different real-world settings at once, a single error can create havoc on a much larger scale. For example, if all of the self-driving cars in a fleet use the same image recognition algorithm, a flaw in that algorithm could make any of the cars crash. This is a well-known problem in cybersecurity; hackers often target a single, widely used system, such as the Microsoft Windows operating system, in order to compromise a large number of users, such as thousands of PC users.

4. What to do

AI accidents are already happening.[45] If we do not act, they will become far more common and destructive. Improvements in AI technology and bottom-up market pressure from consumers may help make AI safer and less accident-prone, but they are unlikely to do enough on their own. Policy has an essential role to play. Smart policy can drive research into less accident-prone AI technologies, bring the AI community together to reduce risks, and provide incentives for private actors to use AI safely, saving lives and livelihoods in the future.

Today, the policy effort around AI safety and accident risk is only beginning. There are several federal actions that will be central to any policy agenda. These include:

- **Facilitate information sharing about AI accidents and near misses.** To make AI safer, we need to know when and how it fails. In many other technological domains, shared incident reporting contributes to a common base of knowledge, helping industry and government track risks and understand their causes. Models include the National Transportation Safety Board's database for aviation incidents and the public-private cyber intelligence platforms known as Information Sharing and Analysis Centers.[46] The government should consider creating a similar repository for AI accident reports. As part of this effort, policymakers should explore different ways of encouraging the private sector to actively disclose the details of AI accidents. For example, the government could offer confidentiality protections for sensitive commercial information in accident reports, develop common standards for incident reporting, or even mandate disclosure of certain types of incidents.[47]

- **Invest in AI safety research and development (R&D).** The federal government and private industry invest billions in AI R&D every year, but almost none of this funding goes to AI safety research.[48] Federal R&D funding has led to critical safety and security innovations in many other contexts, from cryptographic protocols that enable secure communication to the sensors behind modern airbags.[49] It will be crucial to make similar investments in AI safety, including research aiming to solve the problems of robustness, specification, and assurance described above, as well as investing in the development of AI engineering as a more rigorous discipline.[50] The 2021 National Defense Authorization Act (NDAA) made a good start in this direction by including provisions calling for the National Science Foundation and the Department of Energy to invest in research into "trustworthy AI."[51] However, it remains to be seen how much funding will actually be invested in these areas.

- **Invest in AI standards development and testing capacity.** Today, there is no commonly accepted definition of safe AI, and no standard way to test real-world AI systems for accident risk. Federal agencies, such as the National Institute of Standards and Technology, as well as more specialized regulators, such as the Food and Drug Administration and the Federal Communications Commission, are well positioned to help build these resources. To begin, Congress should fund, and NIST should create, a National AI Testbed: a digital platform containing standardized datasets, code, and testing environments on which public and private AI systems can be stress-tested for safety and reliability.[52] This could complement the mandate in the 2021 NDAA for NIST to create an AI risk-mitigation framework and technical standards for AI systems.[53]

- **Work across borders to reduce accident risks.** AI is booming around the world, and the United States' AI safety efforts will be far more effective if it can draw on the innovative capacity and market power of its allies. International R&D alliances, standards bodies such as the International Organization for Standardization, and intergovernmental organizations such as the Organisation for Economic Co-operation and Development could be important forums for collaboration around AI safety. Preventing AI accidents could even be an opportunity for engagement with China, which faces the same accident risks as other AI powers.[54]

Authors

Zachary Arnold is a research fellow at CSET, where Helen Toner is director of strategy.

Acknowledgments

Thanks to Ashwin Acharya, Catherine Aiken, Matt Daniels, Peter Henderson, Tim Hwang, Sean McGregor, Matt Mahoney, Vishal Maini, Igor Mikolic-Torreira, and Molly Wasser for helpful feedback. We are also indebted to Pedro A. Ortega, Vishal Maini, and the DeepMind safety team for their work developing the three-part framework used in this brief. However, the authors are solely responsible for the views expressed in this publication and for any errors.

Document Identifier: doi: 10.51593/20200072

Endnotes

[1] Shana Lynch, “Andrew Ng: Why AI is the New Electricity,” *Insights by Stanford Business*, March 11, 2017, https://www.gsb.stanford.edu/insights/andrew-ng-why-ai-new-electricity; George Dvorsky, “Henry Kissinger Warns That AI Will Fundamentally Alter Human Consciousness,” *Gizmodo*, November 5, 2019, https://gizmodo.com/henry-kissinger-warns-that-ai-will-fundamentally-alter-1839642809.

[2] AI Index Steering Committee, “Artificial Intelligence Index Report 2021” (Stanford University’s Human-Centered Artificial Intelligence Institute, March 2021), 41-79, https://aiindex.stanford.edu/wp-content/uploads/2021/03/2021-AI-Index-Report_Master.pdf; Dave Gershgorn, “The data that transformed AI research—and possibly the world,” Quartz, July 26, 2017, https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/.

[3] Rachel S. Cohen, “Air Force to Test Fighter Drone Against Human Pilot,” *Air Force Magazine*, June 4, 2020, https://www.airforcemag.com/air-force-to-test-fighter-drone-against-human-pilot/.

[4] GPT-3 Creative Fiction, https://www.gwern.net/GPT-3 (“GPT-3’s samples are not just close to human level: they are creative, witty, deep, meta, and often beautiful. . . . Chatting with GPT-3 feels uncannily like chatting with a human.”).

[5] Jürgen Bajorath et al., “Artificial Intelligence in Drug Discovery: Into the Great Wide Open,” *Journal of Medical Chemistry* 63, no. 16 (August 2020): 8651-8652, https://pubs.acs.org/doi/10.1021/acs.jmedchem.0c01077; Jo Marchant, “Powerful antibiotics discovered using AI,” Nature, February 20, 2020, https://www.nature.com/articles/d41586-020-00018-3.

[6] Michael Horowitz et al., “Strategic Competition in an Era of Artificial Intelligence” (Center for a New American Security, July 25, 2018), https://www.cnas.org/publications/reports/strategic-competition-in-an-era-of-artificial-intelligence.

[7] U.S. Census Bureau, *Historical Statistics of the United States, Colonial Times to 1970* (Washington, D.C.: U.S. Department of Commerce, 1970), 827, https://www2.census.gov/library/publications/1975/compendia/hist_stats_colonial-1970/hist_stats_colonial-1970p2-chS.pdf.

[8] Dario Amodei et al., “Concrete Problems in AI Safety,” *arXiv [cs.AI]* (June 21, 2016), arXiv, https://arxiv.org/abs/1606.06565.

[9] See, e.g., “Worldwide Spending on Artificial Intelligence Is Expected to Double in Four Years, Reaching $110 Billion in 2024, According to New IDC Spending

Guide," IDC, August 25, 2020, https://www.idc.com/getdoc.jsp?containerId=prUS46794720.

[10] See generally Jacques Bughin and Jeongmin Seong, "How Competition is Driving AI's Rapid Adoption," *Harvard Business Review*, October 17, 2018, https://hbr.org/2018/10/how-competition-is-driving-ais-rapid-adoption; Paul Scharre, "A Million Mistakes a Second," *Foreign Policy*, September 12, 2018, https://foreignpolicy.com/2018/09/12/a-million-mistakes-a-second-future-of-war/ ("Robert Work, then-U.S. deputy defense secretary, colorfully summed up the problem this way: 'If our competitors go to Terminators, and it turns out the Terminators are able to make decisions faster, even if they're bad, how would we respond?'").

[11] See, e.g., Will Hunt, "The Flight to Safety-Critical AI: Lessons in AI Safety from the Aviation Industry" (UC Berkeley Center for Long-Term Cybersecurity, August 11, 2020), https://cltc.berkeley.edu/wp-content/uploads/2020/08/Flight-to-Safety-Critical-AI.pdf.

[12] Artificial Intelligence Incident Database, incidentdatabase.ai.

[13] David A. Powner, "Information Technology: Federal Agencies Need to Address Aging Legacy Systems," Testimony to the Committee on Oversight and Government Reform, House of Representatives, 114th Congress, May 25, 2016, 15, https://www.gao.gov/assets/680/677454.pdf; Phil Teplitzky, "Closing the COBOL Programming Skills Gap," *TechChannel*, October 25, 2019, https://techchannel.com/Enterprise/10/2019/closing-cobol-programming-skills-gap; Kurt Beyer, *Grace Hopper and the Invention of the Information Age* (Cambridge, MA: MIT Press, 2012), 297, https://www.google.com/books/edition/Grace_Hopper_and_the_Invention_of_the_In/dr34DwAAQBAJ?hl=en&gbpv=1&dq=17+August+on+an+RCA+501+cobol&pg=PA297&printsec=frontcover.

[14] Howard Berkes, "Challenger Engineer Who Warned of Shuttle Disaster Dies," *NPR*, March 21, 2016, https://www.npr.org/sections/thetwo-way/2016/03/21/470870426/challenger-engineer-who-warned-of-shuttle-disaster-dies; Andy Gregory, "Chernobyl: How did the world's worst nuclear accident happen?," *The Independent*, April 26, 2020, https://www.independent.co.uk/news/world/europe/chernobyl-anniversary-what-happened-soviet-union-history-cover-effects-a9482431.html.

[15] See, e.g., Stephen Rice, Krisstal Clayton, and Jason McCarley, "The Effects of Automation Bias on Operator Compliance and Reliance," in *Human Factors Issues in Combat Identification* (Routledge, 2017), 265-276; Colin Wood, "In Case of Emergency: Humans Overtrust Robots, See Them as Authority Figure," *Government Technology*, March 9, 2016, https://www.govtech.com/public-safety/In-Case-of-Emergency-Humans-Overtrust-Robots-See-Them-as-

Authority-Figure.html; David Wesley and Luis Alfonso Dau, "Complacency and Automation Bias in the Enbridge Pipeline Disaster," *Ergonomics in Design* 25, no. 1 (2017): 17-22; Khari Johnson, "Confidence, uncertainty, and trust in AI affect how humans make decisions," *VentureBeat*, February 1, 2021, https://venturebeat.com/2021/02/01/confidence-uncertainty-and-trust-in-ai-affect-how-humans-make-decisions/. The canonical modern example can be found at https://www.youtube.com/watch?v=DOW_kPzY_JY.

[16] Inspired by Angela Lashbrook, "AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind," *The Atlantic*, August 16, 2018, https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619/.

[17] Inspired by "Chinese AI caught out by face in bus ad," *BBC News*, November 27, 2018, https://www.bbc.com/news/technology-46357004; Electric Future (@electricfuture5), "Car kept jamming on the brakes thinking this was a person," Twitter, September 25, 2020, https://twitter.com/electricfuture5/status/1309688641157906433; see Catherine Olsson, "Incident Number 36," Artificial Intelligence Incident Database, https://incidentdatabase.ai/cite/36.

[18] Inspired by Kiona N. Smith, "The Computer That Almost Started a Nuclear War, And the Man Who Stopped It," *Forbes*, September 25, 2018, https://www.forbes.com/sites/kionasmith/2018/09/25/the-computer-that-almost-started-a-nuclear-war-and-the-man-who-stopped-it/?sh=861d6b82835d; USS *Vincennes* (CG-49), Wikipedia, https://en.wikipedia.org/wiki/USS_Vincennes_(CG-49)#Iran_Air_Flight_655. See also Roman Yampolskiy, "Incident Number 27," Artificial Intelligence Incident Database, https://incidentdatabase.ai/cite/27.

[19] See, e.g., Kevin Eykholt et al., "Robust Physical-World Attacks on Deep Learning Models," *arXiv [cs.CR]* (July 27, 2017), arXiv, https://arxiv.org/abs/1707.08945; Alexey Kurakin, Ian Goodfellow, and Samy Bengio, "Adversarial examples in the physical world," *arXiv [cs.CV]* (July 8, 2016), https://arxiv.org/abs/1607.02533.

[20] Inspired by "2018 Hawaii false missile alert," Wikipedia, https://en.wikipedia.org/wiki/2018_Hawaii_false_missile_alert; Robert Walton, "Google machine learning shifts data center operations to maximize efficiency, renewables use," *Utility Dive*, May 4, 2020, https://www.utilitydive.com/news/google-machine-learning-shifts-data-center-operations-to-maximize-efficienc/577074/.

[21] Eykholt et al., "Robust Physical-World Attacks on Deep Learning Models."

[22] Inspired by Matthew Hutson, “A turtle—or a rifle? Hackers easily fool Ais into seeing the wrong thing,” *Science*, July 19, 2018, https://www.sciencemag.org/news/2018/07/turtle-or-rifle-hackers-easily-fool-ais-seeing-wrong-thing; Evan Ackerman, “Slight Street Sign Modifications Can Completely Fool Machine Learning Algorithms,” *IEEE Spectrum*, August 4, 2017, https://spectrum.ieee.org/cars-that-think/transportation/sensors/slight-street-sign-modifications-can-fool-machine-learning-algorithms; Eykholt et al., “Robust Physical-World Attacks on Deep Learning Models.”

[23] Stuart Russell, “Of Myths and Moonshine,” in “The Myth of AI,” Edge, https://www.edge.org/conversation/the-myth-of-ai#26015.

[24] See, e.g., Kevin Roose, Mike Isaac, and Sheera Frenkel, “Facebook Struggles to Balance Civility and Growth,” *The New York Times*, November 24, 2020, https://www.nytimes.com/2020/11/24/technology/facebook-election-misinformation.html; Manoel Horta Ribeiro, “Auditing radicalization pathways on YouTube,” *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, January 2020, 131-141, https://dl.acm.org/doi/abs/10.1145/3351095.3372879?download=true; Will Carless and Jessica Guynn, “YouTube continues to push dangerous videos to users susceptible to extremism, white supremacy, report finds,” *USA Today*, February 12, 2021, https://www.usatoday.com/story/tech/2021/02/12/youtube-channel-recommendation-white-supremacist-extremist-videos/6712787002/; Sara Fischer, “YouTube makes changes to limit hate speech and boost inclusion,” *Axios*, December 3, 2020, https://www.axios.com/youtube-limits-hate-speech-boosts-inclusion-960639b0-7ee0-4401-8e7c-aa7f4d4fce3f.html.

[25] Inspired by Jefferson Graham and Brett Molina, “California fires: Navigation apps like Waze sent commuters into flames, drivers say,” *CNBC*, originally published in *USA Today*, December 17, 2017, https://www.cnbc.com/2017/12/07/california-fires-navigation-apps-like-waze-sent-commuters-into-flames-drivers-say.html; see Catherine Olsson, “Incident Number 22,” Artificial Intelligence Incident Database, https://incidentdatabase.ai/cite/22.

[26] Jack Clark and Dario Amodei, “Faulty Reward Functions in the Wild,” *OpenAI Blog*, December 21, 2016, https://openai.com/blog/faulty-reward-functions/; Dario Amodei, Paul Christiano, and Alex Ray, “Learning from Human Preferences,” *OpenAI Blog*, June 13, 2017, https://openai.com/blog/deep-reinforcement-learning-from-human-preferences/.

[27] Inspired by Tom Murphy VII, “The First Level of Super Mario Bros. is Easy with Lexicographic Orderings and Time Travel…after that it gets a little tricky,” in *SIGBOVIK 2013*, 20-21, http://www.cs.cmu.edu/~tom7/mario/mario.pdf, and the incident described in Kim Zetter, “A Cyberattack Has Caused Confirmed Physical Damage for the Second Time Ever,” *Wired*, January 8, 2015,

https://www.wired.com/2015/01/german-steel-mill-hack-destruction/. See generally Victoria Krakovna, “Specification gaming examples in AI,” *Victoria Krakovna* (personal site), https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/, an invaluable resource.

[28] “Explainable AI: the basics” (The Royal Society, November 2019), 10, https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf.

[29] See generally Cynthia Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence* 1 (2019): 206-215, https://www.nature.com/articles/s42256-019-0048-x.

[30] T. Christian Miller et al., “Collision Course,” *ProPublica*, December 20, 2019, https://features.propublica.org/navy-uss-mccain-crash/navy-installed-touch-screen-steering-ten-sailors-paid-with-their-lives/; “‘I Couldn’t Trust My Airplane Anymore’ – A UX View on the 737 Max,” *Belveal*, July 8, 2019, https://www.belveal.com/blog/2019/18/i-couldnt-trust-my-airplane-anymore-user-experience-boeing-737-max-mcas.

[31] See, e.g., Shira Ovide, “When the Police Treat Software Like Magic,” *The New York Times*, June 25, 2020, https://www.nytimes.com/2020/06/25/technology/facial-recognition-software-dangers.html; Kate Goddard, Abdul Roudsari, and Jeremy C. Wyatt, “Automation bias: a systematic review of frequency, effect mediators, and mitigators,” *Journal of the American Medical Informatics Association* 19, no. 1 (January-February 2012): 121-127, https://pubmed.ncbi.nlm.nih.gov/21685142/; Paul Robinette et al., “Overtrust of robots in emergency evacuation scenarios,” in *HRI 2016 - 11th ACM/IEEE International Conference on Human Robot Interaction*, April 12, 2016, https://pennstate.pure.elsevier.com/en/publications/overtrust-of-robots-in-emergency-evacuation-scenarios; Kamilla Egedal Andersen et al., “Do We Blindly Trust Self-Driving Cars,” in the Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, March 2017, 67-68, https://dl.acm.org/doi/10.1145/3029798.3038428.

[32] Inspired by Charles Perrow, *Normal Accidents: Living with High-Risk Technologies* (Princeton, NJ: Princeton University Press, 1984), 152.

[33] Inspired by Paul Scharre, “Autonomous Weapons and Operational Risk” (Center for a New American Security, February 2016), 14, https://s3.amazonaws.com/files.cnas.org/documents/CNAS_Autonomous-weapons-operational-risk.pdf; Jane Macfarlane, “Your Navigation App Is Making Traffic Unmanageable,” *IEEE* Spectrum, September 19, 2019,

https://spectrum.ieee.org/computing/hardware/your-navigation-app-is-making-traffic-unmanageable.

[34] Inspired by the Boeing 737 MAX crashes; see Chesley B. "Sully" Sullenberger III, "My Testimony Today Before the House Subcommittee on Aviation," Testimony to the House Committee on Transportation and Infrastructure, 116th Congress, June 19, 2019, http://www.sullysullenberger.com/my-testimony-today-before-the-house-subcommittee-on-aviation/; Catherine Olsson, "Incident Number 3," Artificial Intelligence Incident Database, https://incidentdatabase.ai/cite/3.

[35] See, e.g., Timothy B. Lee, "Report: Software bug led to death in Uber's self-driving crash," *Ars Technica*, May 7, 2018, https://arstechnica.com/tech-policy/2018/05/report-software-bug-led-to-death-in-ubers-self-driving-crash/; Andrew J. Hawkins, "Tesla didn't fix an Autopilot problem for three years, and now another person is dead," *The Verge*, May 17, 2019, https://www.theverge.com/2019/5/17/18629214/tesla-autopilot-crash-death-josh-brown-jeremy-banner; see also "Incident Number 20," Artificial Intelligence Incident Database, https://incidentdatabase.ai/cite/20, Roman Yampolskiy, "Incident Number 52," Artificial Intelligence Incident Database, https://incidentdatabase.ai/cite/52, "Incident Number 70," Artificial Intelligence Incident Database, https://incidentdatabase.ai/cite/70.

[36] See, e.g., Sarah Perez, "Facebook partially documents its content recommendation system," *TechCrunch*, August 31, 2020, https://techcrunch.com/2020/08/31/facebook-partially-documents-its-content-recommendation-system/; Roman Yampolskiy, "Incident Number 1," Artificial Intelligence Incident Database, https://incidentdatabase.ai/cite/1.

[37] See, e.g., Bobby Allyn, "'The Computer Got It Wrong': How Facial Recognition Led To False Arrest Of Black Man," *NPR*, June 24, 2020, https://www.npr.org/2020/06/24/882683463/the-computer-got-it-wrong-how-facial-recognition-led-to-a-false-arrest-in-michig; Sean McGregor, "Incident Number 74," Artificial Intelligence Incident Database, https://incidentdatabase.ai/cite/74.

[38] Shraddha Chakradhar, "Widely used algorithm for follow-up care in hospitals is racially biased, study finds," *Stat*, October 24, 2019, https://www.statnews.com/2019/10/24/widely-used-algorithm-hospitals-racial-bias/.

[39] See generally Paul Scharre, "Killer Apps: The Real Dangers of an AI Arms Race," *Foreign Affairs*, May/June 2019, https://www.foreignaffairs.com/articles/2019-04-16/killer-apps.

[40] David Gelles et al., "Boeing Was 'Go, Go, Go' to Beat Airbus With the 737 Max," *The New York Times*, March 23, 2019, https://www.nytimes.com/2019/03/23/business/boeing-737-max-crash.html.

[41] See generally Perrow, *Normal Accidents*.

[42] Scharre, "A Million Mistakes a Second."

[43] Brian Barrett, "The Year Alexa Grew Up," *WIRED*, December 19, 2018, https://www.wired.com/story/amazon-alexa-2018-machine-learning/.

[44] "People Make Poor Monitors for Computers," *Macroresilience*, December 29, 2011, http://www.macroresilience.com/2011/12/29/people-make-poor-monitors-for-computers/.

[45] Artificial Intelligence Incident Database, incidentdatabase.ai.

[46] "Accident Synopses," National Transportation Safety Board, https://www.ntsb.gov/_layouts/ntsb.aviation/month.aspx; National Council of ISACs, https://www.nationalisacs.org/.

[47] Congress is currently considering a mandatory approach for cybersecurity incidents. Jory Heckman, "Warner says Senate committee working on bill to require mandatory reporting for cyber threats," *Federal News Network*, April 30, 2021, https://federalnewsnetwork.com/cybersecurity/2021/04/warner-says-senate-committee-working-on-bill-to-require-mandatory-reporting-for-cyber-threats/.

[48] Ben Buchanan, "The Future of AI and Cybersecurity," *The Cipher Brief*, October 30, 2019, https://www.thecipherbrief.com/column_article/the-future-of-ai-and-cybersecurity.

[49] David P. Leech, Stacey Ferris, and John T. Scott, *The Economic Impacts of the Advanced Encryption Standard, 1996-2017* (Washington, DC: National Institute of Standards and Technology, September 2018), https://nvlpubs.nist.gov/nistpubs/gcr/2018/NIST.GCR.18-017.pdf; Liz Jacobs, "GPS, lithium batteries, the internet, cellular technology, airbags: A Q&A about how governments often fuel innovation," *TED Blog*, October 28, 2013, https://blog.ted.com/qa-mariana-mazzucato-governments-often-fuel-innovation/.

[50] For more on the need for an AI engineering discipline, see Michael I. Jordan, "Artificial Intelligence—The Revolution Hasn't Happened Yet," *Harvard Data Science Review*, July 1, 2019, https://hdsr.mitpress.mit.edu/pub/wot7mkc1/release/9.

[51] "Summary of AI Provisions from the National Defense Authorization Act 2021," Stanford University Human-Centered Artificial Intelligence Institute, https://hai.stanford.edu/policy/policy-resources/summary-ai-provisions-national-defense-authorization-act-2021.

[52] "Comment on NIST Draft Standards for Reliable, Robust, and Trustworthy Artificial Intelligence," Center for Security and Emerging Technology, May 31, 2019, https://cset.georgetown.edu/research/comment-on-nist-draft-standards-for-reliable-robust-and-trustworthy-artificial-intelligence/.

[53] "Summary of AI Provisions," Stanford University Human-Centered Artificial Intelligence Institute.

[54] Andrew Imbrie and Elsa Kania, "AI Safety, Security, and Stability Among Great Powers: Options, Challenges, and Lessons Learned for Pragmatic Engagement" (Center for Security and Emerging Technology, December 2019), https://cset.georgetown.edu/research/ai-safety-security-and-stability-among-great-powers-options-challenges-and-lessons-learned-for-pragmatic-engagement/.