UC BERKELEY

CENTER FOR LONG-TERM CYBERSECURITY

CLTC WHITE PAPER SERIES

The Flight to Safety-Critical AI

LESSONS IN AI SAFETY FROM THE AVIATION INDUSTRY

WILL HUNT

CLTC WHITE PAPER SERIES

The Flight to Safety-Critical AI

LESSONS IN AI SAFETY FROM THE AVIATION INDUSTRY

WILL HUNT
Graduate Researcher, AI Security Initiative
UC Berkeley Center for Long-Term Cybersecurity

CLTC
Center for Long-Term Cybersecurity
UC Berkeley

UC BERKELEY

CENTER FOR LONG-TERM CYBERSECURITY

Contents

Executive Summary

Rapid progress in the field of artificial intelligence (AI) over the past decade has generated both enthusiasm and rising concern. The most sophisticated AI models are powerful — but also opaque, unpredictable, and accident-prone. Policymakers and AI researchers alike fear the prospect of a “race to the bottom” on AI safety, in which firms or states compromise on safety standards while trying to innovate faster than the competition. Yet the empirical record suggests that races to the bottom are uncommon, and previous research on AI races has been almost entirely theoretical.

This paper therefore assesses empirically how competitive pressures — economic and political — have affected the speed and character of AI research and development (R&D) in an industry with a history of both extensive automation and impressive safety performance: aviation. Based on interviews with a wide range of experts, findings show limited evidence of an AI race to the bottom and some evidence of a (long, slow) race to the top. In part because of extensive safety regulations, the industry has begun to invest in AI safety R&D and standard-setting, focusing on hard technical problems like robustness and interpretability, but has been characteristically cautious about using AI in safety-critical applications. This dynamic may also be at play in other domains, such as the military. These results have implications for policymakers, regulators, firms, and researchers seeking to maximize the upside while minimizing the downside of continued AI progress.

Key findings:

- **In most industries, the empirical evidence of racing to the bottom is limited.** Previous work looking for races to the bottom on environmental, labor, and other standards suggests that race-to-the-top dynamics may be equally or more common. In the case of AI specifically, few researchers have attempted to evaluate the empirical evidence of a race to the bottom.

- **In the aviation industry, the lack of AI-based standards and regulations has prevented the adoption of safety-critical AI.** Many modern AI systems have a number of features, such as data-intensivity, opacity, and unpredictability, that pose serious challenges for traditional safety certification approaches. Technical safety standards for AI are only in the early stages of development, and standard-setting bodies have thus far focused on less safety-critical use cases, such as route planning, predictive maintenance, and decision support.

- **There is some evidence that aviation is engaged in a race to the top in AI safety.** Industry experts report that representatives from firms, regulatory bodies, and academia have engaged in a highly collaborative AI standard-setting process, focused on meeting rather than relaxing aviation's high and rising safety standards. Meanwhile, firms and governments are investing in research on building certifiably safe AI systems.

- **Extensive regulations, high regulatory capacity, and cooperation across regulators all make it hard for aviation firms to either cut corners or make rapid progress on AI safety.** Despite the doubts raised by the tragic Boeing 737 crashes, regulatory standards for aviation are high and relatively hard to shirk. The maintenance of high safety standards depends in part on regulators' power to impose significant consequences on firms when they do attempt to cut corners.

Recommendations:

- **Policymakers: Increase funding for research into testing, evaluation, verification, and validation (TEVV) for AI/autonomous systems.** Expeditious progress in the TEVV research agenda will unlock significant economic and strategic benefits, in aviation as well as other safety-critical industries. Aviation firms will invest in parts of the TEVV research agenda unprompted, but universities and AI labs are more likely to drive much of the fundamental progress required for safety-critical AI.

- **Regulators: Provide incentives for firms to share information on AI accidents and near-misses.** Aviation regulators have deliberately developed forums, incentives, and requirements for sharing information about possible safety hazards. Historically, firms have recognized that they will not be punished for being open about mistakes, and that they benefit from learning about others' safety difficulties.

- **Firms: Pay the costs of safety in advance.** Conventional wisdom in the aviation industry holds that software defects that cost $1 to fix in requirements or design cost $10 to fix during a traditional test phase and $100 to fix after a product goes into use. As the capital costs of training AI systems increase, and AI use cases become higher-stakes, firms will need to invest early in verification and validation of AI systems, which may include funding basic AI safety research.

- **Researchers: Analyze the relationship between competition and AI safety on an industry-by-industry and issue-by-issue basis.** This paper's findings affirm that the competitive dynamics surrounding AI development will likely vary from one industry and issue area to the next. Microsoft's recent call for regulation to prevent a race to the bottom in facial recognition technologies suggests that safety is not the only area in which race dynamics could have socially harmful effects. And different industries vary considerably in their norms, market structures, capital requirements, and regulatory environments, all of which affect competitive dynamics. Of special interest is the military avionics industry: preliminary findings from this paper suggest, contrary to media accounts, that the U.S. military may be even slower to adopt AI than the commercial aviation industry, and has made significant investments in AI safety research.

Introduction

Concerns are rising about a possible race to the bottom on AI safety.[1] AI systems are often opaque and display unpredictable behavior, making it difficult to evaluate their reliability or safety.[2] Yet politicians, defense officials, and police departments have sometimes shown more enthusiasm for novel applications of AI than awareness of the accident risks these applications might pose.[3] Some observers worry, in particular, that the popular but misleading narrative of an "AI arms race" between the United States and China could lead the two countries to take greater risks with safety as each hurries to develop and deploy ever-more powerful AI systems before the other.[4] In the words of Paul Scharre, former Special Assistant to the U.S. Under Secretary of Defense for Policy, "For each country, the real danger is not that it will fall behind its competitors in AI but that the perception of a race will prompt everyone to rush to deploy unsafe AI systems."[5]

In the private sector, too, AI developers have expressed worries that economic competition might lead to the sale of AI systems with impressive capabilities but weak safety assurances. AI research lab OpenAI, for example, has warned that artificial general intelligence (AGI) development might become "a competitive race without time for adequate safety precautions."[6] And while fears of an AI race to the bottom often center on safety, other issues, like deliberate misuse of AI, raise similar concerns. Consider Microsoft, which has actively advocated for new regulations for AI-based facial recognition technologies on this basis, with president Brad Smith stating, "We believe that the only way to protect against this race to the bottom is to build a floor of responsibility that supports healthy market competition."[7]

But current discussions of the existing or future race to the bottom in AI elide two important observations. First, different industries and regulatory domains are characterized by a wide range of competitive dynamics — including races to the top and middle — while claims about races to the bottom often lack empirical support.[8] Second, AI is a general-purpose technology with applications across every industry; we should therefore expect significant variation in competitive dynamics and consequences for AI from one industry to the next. For example, self-driving car firms entering the highly competitive automotive industry have invested heavily in AI safety research, and fully autonomous vehicles will likely make driving far safer in the long run.[9] Differences in AI use cases, safety norms, market structures, capital requirements, and, perhaps especially, regulatory environments all plausibly affect the willingness of firms and states to invest in, or compromise on, standards, regulations, and norms surrounding the use and abuse of AI systems.

This paper therefore proposes analyzing the nature of competitive dynamics surrounding AI safety on an issue-by-issue and industry-by-industry basis. Rather than discuss the risk of "AI races" in the abstract, this research focuses on the issue of AI safety within commercial aviation, an industry where safety is critically important and automation is common. Do the competitive dynamics shaping the aviation industry's development and rollout of safety-critical AI systems and technical standards constitute a race to the bottom, a race to the top, or a different dynamic entirely?

To answer this question, the paper draws on interviews with more than twenty subject-matter experts, including commercial pilots, system safety engineers, standard-setters, regulators, academics, and air traffic controllers. The results suggest that the aviation industry has so far approached AI with great caution. For safety-critical use cases, such as autonomous flight or air traffic control, AI simply will not be used in the foreseeable future. And while timelines are long for safety-critical AI for aviation, firms like Airbus and Boeing are investing in AI-related R&D in hopes of eventually developing autonomous systems that can meet the industry's high and ever-rising safety standards. In short, the industry is engaged in a (long, slow) race to the top.

The findings from this research have implications for both policymakers and researchers. They suggest the need for further investment in AI safety, to speed up the race to the top and ultimately unlock significant benefits in industries like aviation. The results also highlight the critical role that industry-specific standards and regulatory environments play in shaping racing dynamics and suggest the value of an industry-by-industry exploration of AI races. We should expect important variation in how different industries respond to continued AI progress: there will be not one, but multiple AI races worthy of study.

The next section briefly reviews existing literature on races to the bottom, middle, and top, and argues for the value of exploring these dynamics in the aviation industry specifically. The subsequent section marshals evidence from interviews and publicly available data suggesting that aviation is engaged in a "race to the top" toward AI safety. It then explores the factors underlying this dynamic, and the extent to which they might apply — or be desirable — in other industries. The penultimate section makes recommendations for policymakers, regulators, firms, and researchers seeking to accelerate the flight to safety-critical AI. The paper concludes with an overview of implications for future policy research focused on accident risks from AI technologies.

Theory and Methodology

The logic of "races to the bottom" is intuitive and appealing. Like firms engaged in a price war, investment-strapped states gradually reduce the constraints on firm behavior — typically standards or regulations governing labor, the environment, safety, and other variables — until the benefits of attracting foreign investment are outweighed by the social costs of compromised labor laws, increased pollution, and other regulatory compromises.[10]

Consider, for example, the widely-cited competition between New Jersey and Delaware in the late 19th century, in which the two states competed to slash corporate regulations in order to attract more investment.[11] This dynamic — now often referred to as "The Delaware Effect" — left Delaware with some of the most business-friendly corporate laws in the United States, and today more than two-thirds of Fortune 500 companies are incorporated in Delaware.[12] A similar logic has been applied to firms: in order to remain competitive, corporations might cheat on existing standards or lobby for lower standards, in order to save on the time and cost of compliance.

Claims of such races to the bottom have featured prominently in many important policy debates of the last few decades. In 1969, for example, images of the beaches of Santa Barbara, California in the aftermath of a major oil spill came to symbolize the race to the bottom in environmental standards, contributing to the passage of U.S. environmental regulations in the 1970s.[13] More recently, critics of trade agreements such as the Trans-Pacific Partnership have argued that trade agreements may induce races to the bottom on a range of issues: for example, if firms can easily relocate their business to new countries, this might induce countries to lower their labor standards in an effort to lure and retain foreign direct investment.[14]

The notion of a possible race to the bottom in safety standards for AI specifically is relatively new but rising in prominence, perhaps especially within the U.S. national security community. For example, Larry Lewis, Director of the Center for Autonomy and Artificial Intelligence at CNA, wrote in a recent article, "A transformative technology like AI can be used responsibly and safely, or it could fuel a much faster race to the bottom."[15] Similarly, former Secretary of the Navy Richard Danzig, in a report on risks from rapid development of AI, synthetic biology, and other emerging technologies, concludes that "superiority is not synonymous with security: There are substantial risks from the race."[16]

Increased attention to the prospect of races to the bottom in AI safety is importantly related to the confused, albeit attention-grabbing narrative of an “AI arms race” between the United States and China. To some U.S. defense officials, China’s recent progress and eagerness to invest in AI suggest analogies to the Cold War build-up of nuclear weapons. The “AI arms race” narrative has gained traction across the web: prior to 2016, a Google search for the phrase “AI arms race” yielded just 300 hits, but in 2020, the same phrase yields more than 100,000 hits.[17]

This Cold War analogy is flawed, likely doing more to exacerbate tensions with China than to clarify the competitive dynamics surrounding AI development.[18] General purpose technologies like electricity likely provide a more appropriate analogy to AI than nuclear weapons do.[19] Even so, some fear that the arms race narrative could nevertheless contribute to a race to the bottom between the United States and China, especially given the souring of relations between the two nations in recent years.[20]

Considering the substantial influence that the race to the bottom narrative has had in policy debates, we might expect the empirical record to show strong evidence of such races, both within AI safety and more broadly. In fact, however, the evidence of races to the bottom is surprisingly elusive across most industries and issue areas.[21] Indeed, the race to the bottom dynamic may not explain even the eponymous case of Delaware.[22] Meanwhile, some scholars have documented a “California Effect,” in which larger firms actively encourage states and countries to impose more extensive regulations, which can serve as barriers to entry for start-up firms lacking the capital and know-how to achieve compliance.[23] Recent literature on the closely related “Brussels Effect” shows that a similar race-to-the-top dynamic obtains at the global level.[24]

The academic literature considering races to the bottom specifically in the domain of artificial intelligence is thin and almost entirely theoretical. Work on the subject typically avoids claiming that a race to the bottom is, in fact, likely.[25] As one recent paper notes, “Instead of offering predictions, this paper should be thought of as an analysis of more pessimistic scenarios.”[26]

Thus, despite the growing prominence of the AI race to the bottom narrative, previous work has left largely unexamined the empirical question of whether any industries currently show signs of compromising on AI safety standards.[27] This is an oversight: AI systems are in ever-wider use, and firms, regulators, and states are actively grappling with the serious challenges posed by AI safety. This paper thus starts from the premise that, while AI remains an emerging technology, it is possible and valuable to make an empirical study of early efforts to set standards for safety-critical AI.[28]

WHY AVIATION?

The first challenge for any empirical analysis of AI racing dynamics is that AI is a general purpose technology with applications across every industry, and competitive dynamics will vary from one industry to the next. This paper therefore focuses on a single industry — aviation — rather than attempting to explore multiple industries at once. To further narrow the aperture, this paper focuses specifically on "safety-critical" AI applications, which face a different set of regulatory requirements from non-safety-critical applications (Box 1).

Box 1. What is "safety-critical AI"?

This paper follows the definition of AI used by the Organisation for Economic Co-operation and Development (OECD): "An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy."[29] Note that while many systems captured by this definition pose safety problems, many more do not. See Hernandez-Orallo et al. (2019) for a survey of safety risks associated with AI.[30]

"Safety-critical AI" describes any AI system for which unintended behavior could be extremely costly. Safety-critical systems — from control systems for trains and planes to nuclear power plants — typically go through extensive and costly testing, evaluation, verification, and validation (Box 2) before being certified for use. Some AI applications in the aviation industry do not qualify as "safety-critical." For example, a faulty AI-based route planning system could result in flight delays, but likely not a serious accident. By contrast, a fully autonomous AI system responsible for managing plane takeoffs and landings would qualify as safety-critical, because failure could result in a collision or other accident.

A second challenge is data availability. As discussed in the next section, commercial applications of AI — though quickly rising in popularity — remain relatively rare, especially in safety-critical domains. In most industries, standard-setting bodies have only just begun work on modifying existing standards to account for AI. Quantitative data on AI adoption and use cases in a given industry often does not exist. But qualitative data gathered from subject-matter experts, though less precise, can offer insights into the current status and future trajectory of racing dynamics in a given domain.

The aviation industry's experiences with AI safety are especially interesting for two reasons. First, aviation is exceptionally safety-conscious. The tragic Max 737 crashes have justifiably cast doubt on the continued reliability of both Boeing and the Federal Aviation Administration

(Box 4). Yet these crashes stand out against an exceptional safety track record. Table 1 presents statistics from Barnett (2020), disaggregated into three groups of nations: "first world," "advancing," and "less-developed." Across all three groups, accident rates have fallen dramatically over each of the last three decades, typically by a factor of two or more. As Barnett notes, accident rates in the "traditional first world" between 2008-2017 were so low that a child boarding a flight in the United States had a higher chance of growing up to be a U.S. president than of dying in a plane crash.[31] Fatality rates are far higher in less-developed countries, though China and Eastern Europe have achieved "first-world" levels of safety over the past decade.

Table 1. Fatality risk per flight boarding across three groups of countries, 1988–2017[32]

	1988–1997	1998–2007	2008–2017
Traditional first world	1 in 4.4 million	1 in 10.8 million	1 in 28.8 million
Advancing	1 in 1 million	1 in 1.9 million	1 in 10.9 million
Less developed	1 in 200,000	1 in 400,000	1 in 1.3 million

Second, relative to other industries with exceptional safety performance — for example, the nuclear power industry — aviation is much more exposed to market forces. While heavy regulations, high capital requirements, and other features have led to significant market concentration, even large duopolistic firms like Boeing and Airbus face pressure to innovate at the cutting edge while saving costs wherever possible. The pressure to automate is especially fierce: most air accidents are at least partly the result of human error, thus reducing reliance on humans for safety-critical functions can reduce the risk of accidents and their financial and reputational consequences.[33] Partly for these reasons, the jobs of air traffic controllers and pilots are among the most heavily automated in the U.S. economy.[34]

In short, aviation's combination of safety-criticality and competitiveness make it an ideal industry in which to explore AI safety racing dynamics. The next section therefore takes a deep dive into the industry, drawing on interviews with a range of experts to determine both the intensity and nature of AI racing dynamics.

A Long, Slow Flight to AI Safety

What might the aviation industry have to gain from AI — and what makes "safety-critical AI" a difficult problem? What are the current status and trajectory of AI in the aviation industry, and are there any signs of a race to the bottom? Finally, what are the consequences — positive and negative — of extensive safety regulations in the aviation industry? This section draws on publicly available materials and interviews with a range of aviation experts to probe these questions.

Because of the wide range of stakeholders involved in aviation safety and the lack of existing research in this area, interviewees were sourced from a wide range of backgrounds. They included air traffic controllers, academics working at the intersection of technical AI safety and aviation, safety engineers at both large and small manufacturers of aircraft, developers supplying AI-based applications to aviation firms, experts in cybersecurity for aviation, and a commercial pilot. A number of interviewees had experience across multiple industries aside from aviation, including self-driving cars, utilities, and semiconductors. Given the focus of this paper on commercial aviation, just two interviewees had experience in military aviation: further research in this domain would benefit from exploring the extent to which the findings of this paper generalize to the military setting.

In light of the diversity of individuals interviewed for the paper, interviews were unstructured, focusing on each interviewee's area of expertise. The results from this paper should therefore be understood as preliminary: future work might profitably test this paper's conclusions more systematically, perhaps through structured interviews or an expert survey.

Overall, the results suggest that, despite the economic and safety benefits of AI, aviation has taken a (characteristically) conservative approach to AI adoption. Regulators and firms alike have focused their efforts on setting standards, especially for less safety-critical applications of AI, such as predictive maintenance and route planning. At the same time, they are investing in AI safety research, which promises to unlock safety-critical applications that are currently off the table. The section concludes with a discussion of the likelihood and potential consequences — both positive and negative — of extensive AI safety regulations in other industries.

THE APPEAL OF SAFETY-CRITICAL AI FOR AVIATION

Safety-critical AI offers a range of potential benefits to both aviation firms and regulators, which could plausibly induce a race-to-the-bottom dynamic in the absence of regulation. First among those benefits are the cost savings that AI applications could enable. As noted in the previous section, compared with other safety-critical industries, aviation faces significant pressures to cut costs. One interviewee with experience in both the aviation and utilities industries emphasized that aviation is more highly exposed to market forces relative to other safety-critical industries such as the nuclear industry: "Nuclear is a very high-risk operation and extremely technical—but it operates within a fence, which can provide more operational predictability compared to aviation," the expert argued. "Nuclear operations are also somewhat more insulated from radical market forces. Since they are part of utilities, which are most often natural monopolies, costs for maintaining safety margins are set within strong regulatory cost controls. . . . With airlines, if you get spikes in fuel costs or coronavirus, you don't know if you'll survive. Margins have always been extremely tight and uncertain; there's incredible pressure to improve efficiency, while making sure the safety margin remains acceptable. As in all safety-sensitive industries, catastrophic loss can mean the loss of the company."

Given the pressure to cut costs, safety-critical AI may eventually become a necessity for airlines, manufacturers, and suppliers hoping to remain cost-competitive. For example, regulations previously required three pilots on any commercial flight; today, only two are required. As one interviewee said, "In the end [for manufacturers and airlines], everything is about money. One experienced pilot can cost $300,000 per year — that's a huge figure." Interviewees said they believe that AI will allow airlines to remove the remaining co-pilot from most planes, and in the long run, will replace human pilots entirely. Airbus, for example, successfully executed a fully automated takeoff in January, 2020, with help from an AI-based vision system.[35] The company has plans to complete a fully automated taxi and landing by the mid-2020s.[36] A similar trend holds in air traffic control, which in the United States ranked as the sixth-most heavily automated job over the past two decades ("pilots, copilots, and flight engineers" ranked third).[37]

Another potential advantage of safety-critical AI is the speed with which it can be developed, relative to more traditional software. Hard-coding safety-critical software is time-intensive, requiring consideration of innumerable edge cases (Box 2). By contrast, AI systems can serve as a relatively lightweight alternative. An interviewee who had worked on implementing AI in a decision-support context noted, "You can either hard-code systems manually to do certain functions, or use AI so you can do it quicker. . . . You can spot a specific problem, then train an AI model quite quickly, test it, then get significant benefits." This is possible in part because the aviation industry collects a tremendous amount of data, which makes it possible to quickly train data-hungry AI models. As one expert working on air traffic control (ATC) said: "ATC analysis, radar data analysis, capacity studies—it won't be long before others reach into this space. We have so much structured data."

Box 2. Costs of safety-critical software certification

Some interviewees expect certifying AI systems to be faster and cheaper than traditional software certification in many cases. If true, this could be a major benefit of AI within the aviation industry, because traditional software certification is very expensive. For experienced teams, certification — which involves rigorous design, documentation, and testing to ensure that a software tool is virtually failsafe — can increase development costs by 20 to 40 percent. Most teams lack the requisite experience, however: on average, software certification adds 75 to 150 percent to total development costs.[38] Among the most important cost drivers is documentation. One interviewee, discussing past experience working with a software company supplying the aviation industry, underscored the documentation challenge: "We did an estimate on numbers of lines, comparing documentation and code, and it's a tenfold difference. You have to create all these documents whenever you introduce a change. You have to use all this tracking."

Of course, even if certifying AI systems does reduce cost relative to traditional software, certification will continue to be expensive, perhaps giving firms incentives to take shortcuts where they can. This suggests the importance of giving regulators the power to impose real costs on shirkers when accidents do happen (see Box 4 for more).

In addition to the economic pressures faced by companies, regulators face political pressure to proactively prepare for an AI future. One interviewee, for example, suggested that if the U.S. Federal Aviation Administration were to reject a company's AI application because it simply did not understand how the application worked, the company might complain to its local member of Congress, with potential political consequences for the FAA. As an interviewee from FAA put it, "We are trying to get out ahead of somebody showing up and saying, 'Oh, by the way, we have this disruptive technology that we're using.'"

Additional pressure stems from fears of falling behind regulators in other regions. In particular, one interviewee from Europe highlighted concerns about falling behind China and the United States. When asked what the consequences of falling behind might be, she referenced a meeting with regulators from the Civil Aviation Authority of China (CAAC): “They have no problem managing the whole aviation sector from China. But what about autonomy? What about sovereignty? Behind AI, there are these security dimensions. How much information do you want to share?” Other EU-based interviewees did not echo these concerns about China, but affirmed that European aviation regulators and air traffic control providers pay close attention to other regulatory bodies, particularly the FAA in the United States.

Firms thus have economic incentives to develop and implement safety-critical AI, while regulators face pressures to keep up with both the pace of AI development in the private sector, as well as the pace of regulators in other nations. For now, however, significant bureaucratic and technical obstacles stand in the way of widespread adoption of AI. The next section discusses these obstacles in greater detail.

TECHNICAL AND BUREAUCRATIC OBSTACLES TO SAFETY-CRITICAL AI

While the promise of AI is significant, interviewees agreed that the certification of AI systems remains a core obstacle. Software certification for aviation is expensive and time-intensive in general, and AI introduces additional bureaucratic and technical hurdles.

First, current aviation safety standards, regulations, and approaches simply were not designed with AI in mind. Existing standards will thus require significant revisions before AI systems can be pronounced sufficiently safe to deploy, at least in safety-critical functions. As one expert noted, certifying AI in aviation is “absolutely a challenge, because there’s no guidance or requirements that I can point to and say, ‘I’m using that particular requirement.’”

Why does AI pose a problem for existing standards, regulations, and protocols? Experts interviewed consistently pointed to two major concerns, which also feature prominently in technical AI safety research (see Box 3):

Opaque, unpredictable algorithms. Many AI systems, particularly deep learning systems, are unpredictable in the sense that they react in surprising ways to slight perturbations of input. They are also opaque, meaning that regulators cannot easily probe the logic of a system to understand when such reactions are likely to occur. As one interviewee noted, “You can test the model one million times, but you still don’t know what might happen the next time. . . . You might develop a very sophisticated algorithm that appears to work every time. But in the end, is it certifiable?” Another interviewee suggested a significant distinction between automation and

AI: “Automation is very simple, you want a system that always does the same thing. AI is just the opposite! You’re never sure if it will reproduce the same thing.”

Heavy reliance on data. Another point of concern is that most modern AI systems rely heavily on data. The aviation industry does not have well-established procedures for certifying data-intensive, safety-critical systems. One FAA systems safety engineer recounted a meeting with a firm that had developed an application using data from an aircraft engine. The data contained many missing values, and in the process of filling in the gaps, the firm had made a number of unjustified assumptions. In the process, the FAA expert noted, they “may have missed a potential event that could eventually turn into an accident.”

Box 3. The nascent AI safety research subfield

Despite the significant technical challenges standing in the way of safety-critical AI, recent developments in research labs and academia provide grounds for optimism. Technical AI safety research agendas have proliferated over the last three years, laying the groundwork for progress on thorny problems with AI, such as the problems of explaining how a machine learning model arrived at a given prediction (“explainability”) or predicting whether an algorithm will behave sensibly in novel environments (“robustness”).[39] Because many of these issues pose significant hurdles to bringing AI-based products to market, the researchers working on them have begun to attract private investment, including from the aviation industry.

Meanwhile, in the public sector — especially at large agencies like the Department of Defense — researchers have begun work on the testing, evaluation, verification, and validation (TEVV) of AI-based systems. Indeed, the Defense Advanced Research Projects Agency has made TEVV for AI/autonomy a core focus through a range of programs, including a program dedicated to “Explainable AI and Assured Autonomy.”[40]

The significant increase in AI safety research in academia, for-profit companies, and public agencies suggests we will likely see progress on some of the vexing technical problems presented by AI. The question, of course, is whether increased safety research will keep up with the exploding number of AI applications across the global economy.

These and other technical features common to many AI systems pose problems for safety certification. Regulators cannot evaluate the details of the safety case of every novel aviation application from scratch; instead, they typically delegate much of the certification process to firms, which becomes much more straightforward with the benefit of standardized performance benchmarks. One interviewee, an FAA safety delegate, affirmed that it is simply ineffective for regulators to dissect the details of the safety case for every new aviation application: “The regulators can’t look at everything we look at; it’s not humanly possible.” Another interviewee, an AI engineer working on decision support for aviation, echoed this sentiment: “In our industry, they [regulators] don’t really care about Caffe1 vs Caffe2, how many layers, and [other details of a given machine learning model]. We get questions about the entire safety case. It doesn’t matter what the values of the neurons are, just what the end-to-end performance is.” Interviewees working on safety certification for AI described the need for extensive conversations with regulators, focused on building trust and developing shared expectations around how to measure safety performance. While these conversations are necessary at this early stage, they come at a price paid in time, money, and regulatory overhead.

Simply put, standards for safety in aviation are exceptionally high. As one safety expert put it, “No combination of failures can be more likely than one in a billion. Which seems like a crazy number — but it does drive architecture schemes.” Continuing to drive down accident risks, especially in the first world, will be difficult. In 2017, for example, there were just eight deaths worldwide attributable to commercially scheduled flight accidents. (In 2018 and 2019, by contrast, the Max 737 crashes caused more than one hundred deaths each.)[41] The bar is therefore high for novel applications of technologies like AI, especially in safety-critical functions.

ASSESSING THE EVIDENCE OF AN AI RACE IN AVIATION

Given the attractions and challenges of safety-critical AI for aviation, this section takes stock of the industry’s approach to AI thus far. At a high level, while experts interviewed were optimistic about the prospects for AI in the long term, all expected the industry to proceed slowly with safety-critical AI for the foreseeable future. The focus, for now, is on standard-setting for non-safety-critical uses of AI. And while there is limited evidence of a race to the bottom, there is some evidence of a race to the top. Firms are investing in the development of technical standards for safety-critical AI, as well as R&D for AI-enabled systems, with the understanding that meeting aviation’s safety standards will be a difficult, decades-long process.

Interviewees were virtually unanimous in the assessment that, for the time being, safety-critical AI is close to non-existent in the aviation industry. In explaining why, they cited first and foremost a lack of technical safety standards and safety regulations. In an industry where regulatory approval is a necessity, standards and regulations are crucial for reducing firms' uncertainty about whether expensive R&D will result in certifiably safe, marketable applications. Unfortunately, setting up technical standards, regulations, and certification processes for safety-critical AI is likely to be a years- or decades-long process. Because so much air traffic flows across national borders, international coordination on aviation safety standards is critical, and this takes time. The long journey toward safety-critical AI is only just beginning.

Early work on standard-setting has begun within the joint working group on artificial intelligence for aviation set up by U.S.-based SAE International and EUROCAE, an EU-based organization.[42] Formed in 2019, this group is working at the cutting edge in standard-setting for AI in aviation. Yet so far, the group has focused on non-safety-critical use cases, such as predictive maintenance and route planning. Because progress requires continuous input, and a measure of consensus, from stakeholders from academia, industry, and national and regional regulatory bodies, standard-setting is a slow process.

In focusing on less safety-critical applications, the EUROCAE-SAE working group aims to test out new regulatory approaches in contexts where lessons learned could be applied to more safety-critical contexts. This is typical of the industry's treatment of new technologies, according to one expert from the EU: "As usual in aviation, you start using it in the best conditions, then build confidence and see how far you can go," she said.[43] Another expert affirmed that maintenance is a particularly promising place to test out new approaches to standard-setting for AI, which could then transfer to other, more complex areas: "You might think these [maintenance vs. flight control] are two different topics. But if you look at the technical challenges, the underlying problems are similar."

Once developed in working groups, AI standards will take years to propagate into real-world safety certification processes. Bodies like the International Civil Aviation Organization will draw on privately developed standards to develop their own standards and recommendations, coordinating across the many UN member states. Then, national or regional regulators such as the U.S. Federal Aviation Administration and the European Union Aviation Safety Agency (EASA) will translate these and other standards into binding regulations and certification procedures for firms in their respective jurisdictions. This process, too, will involve time-intensive coordination: regulators typically make arrangements for the reciprocal recognition of safety certifications, but this is only feasible insofar as regulators trust and understand each others' processes.

Firms do have the option of developing AI applications before safety standards have been modified. Doing so, however, requires building a compelling "safety case" in the absence of widely accepted benchmarks and certification protocols, introducing uncertainty and delay to the development and certification processes. Of course, firms can and do sometimes attempt to obscure important technical changes to regulators, as the Max 737 crashes made tragically clear. But when accidents happen, they must deal with often severe financial consequences, as well as intense scrutiny by both regulators and the public (Box 4).

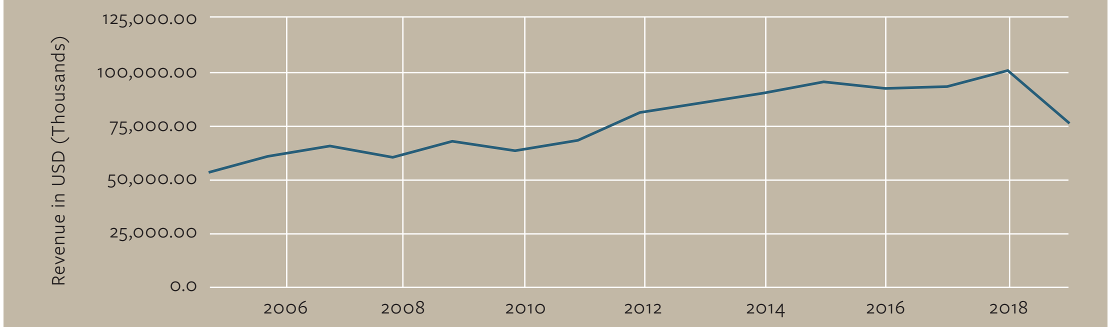
Box 4. The Max 737 crashes and the importance of empowering regulators

On October 29, 2018, a Boeing Max 737 plunged into the Java Sea, thirteen minutes after takeoff from Jakarta, killing all 189 people onboard. Four months later, on March 10, 2019, another Max 737 crashed — this one six minutes after takeoff from Addis Ababa, killing all 157 people onboard.

The Max 737 crashes demonstrate the tragic consequences of cutting corners on safety. Reporting on the Max 737 crashes showed that Boeing had deliberately obscured from regulators a major change to the flight control software involved in both crashes.[44] Airbus' release of its A320neo aircraft generated intense pressure at Boeing to build the Max 737 as quickly as possible.[45] Had the change been flagged, airlines would have been required to retrain their pilots in flight simulators, which is expensive and time intensive.[46]

The crashes also make it clear why such corner-cutting behavior is so rare in aviation. Boeing could, in principle, obscure a wide range of novel technologies from regulators, in order to save time and money on the front end. But the company answers to regulators when things go wrong. Boeing has estimated that the grounding of its 737 planes has cost $20 billion dollars (Figure 1).[47] If the aviation industry is conservative in advancing new technologies, this is in part because of regulators' power to make firms pay the price for their recklessness.

Figure 1. Boeing Revenue, 2005–2019



The serious financial consequences that attend any aviation accident mean that all stakeholders, including for-profit firms, have a genuine interest in improving safety standards to the greatest extent possible. But there are important differences in incentives between regulators and the firms they regulate. One interviewee, a participant in the EUROCAE-SAE working group, described those differences as follows: "Companies want to push forward with this agenda [developing AI standards], because then they can develop new products that are more appealing to customers. That's very roughly their approach. On the other hand, research institutes or universities approach the process from a scientific point of view: knowledge creation is their main objective. Researchers want to understand the technicalities and specifics of these areas, without an interest in implementation. When it comes to regulators, they want to make sure that safety is not compromised, because they are accountable for safety. They want to make sure they have a 360-degree view of this problem, with everyone having a voice. So for them, safety is the number one priority, even if that compromises development speed." The same speaker also reflected, however, that despite these differences in incentives, the standard-setting process is relatively collaborative and productive: "You might expect that there are different stakeholders with different objectives with different points of view. But in the end it works pretty well."

National regulators have, for the most part, not yet begun the process of developing safety regulations for AI. But if and when such regulations are developed, experts indicated they see little chance of a "Delaware Effect" emerging. Indeed, their assessments of the history of aviation safety regulation more often described something like the "Brussels" or "California" effects. One expert in air traffic control in Europe noted that the FAA, in particular, is often viewed as a model by other regulatory bodies: "With the FAA being such a strong, independent regulator, quite often people will look at FAA to see what its position or policy is."

This relative alignment of incentives around raising rather than compromising safety standards and regulations is quite necessary, because aviation manufacturers largely handle certification themselves. The 737 crashes generated skepticism about the wisdom of delegating certification to firms, but the practice goes back decades and has historically worked well. As one expert working on certifying AI for air traffic control noted, "It's not that you're trying to overcome a hurdle. . . . If you want to do something like this, then you have to prove it to your own organization as much as to the regulator."

For now, the lack of established standards and the technical difficulty of assuring AI safety make it difficult to imagine very ambitious safety-critical AI applications for aviation in the near future. Instead, firms and regulators alike are relatively aligned in the goal of setting up technical standards for AI that match or improve upon the the industry's current safety standards. Recognizing both the benefits and difficulties that safety-critical AI presents, firms are investing to some degree in new AI technologies, but may be hesitant to take the risk of commercial implementation until standards and regulations exist. The use of AI so far has been limited to lower-stakes contexts where it presents both economic benefits *and* safety improvements, typically in areas like decision-support and maintenance. The next section illustrates this assessment with a case study of AI for air traffic control.

CASE STUDY: AI FOR AIR TRAFFIC CONTROL

This section documents one of the most ambitious efforts to integrate AI into a safety-critical context: NATS' and Searidge Technologies' experiment with AI for air traffic control at London Heathrow Airport. The case highlights both the benefits of AI, and the limits of safety-critical use for the time being.

The skies of London are exceptionally congested. Approximately 180 million passengers arrive at, or depart from, London's six airports every year, making the city the busiest airspace in the world. Yet despite its high volume of air traffic, London has just eight runways. By comparison, the world's second-busiest airspace — New York City, with 20 percent less air traffic — has twelve runways. Most of London's crowded runways therefore operate continuously at 99 percent capacity, with zero tolerance for human or technical error.[48]

Adding to the challenge, London air traffic control (ATC) must contend with a great deal of cloud cover. Its ATC towers, which rise almost 300 feet into the air, frequently lose visibility, forcing controllers to rely on radar rather than sight. In low-visibility conditions, safety requirements mandate increasing the distance between arriving and departing aircraft by up to 50 percent, dramatically decreasing traffic flow.

The combination of high and increasing flight volumes, few runways, and cloudy skies is a problem for NATS, Britain's sole air traffic control provider. In Europe (in contrast with the United States), funding for air traffic control is tied to air traffic volumes. This gives NATS strong incentives to innovate a way to cope with London's variable weather patterns.

NATS has recently responded to these incentives by developing — in collaboration with Searidge Technologies, a Canadian software firm focused on ATC applications — a set of specialized cameras and video displays that together allow Heathrow's air traffic controllers to operate continuously, even in periods of low visibility. NATS' new "digital tower" setup is equipped with an AI-based image recognition model that flags when an aircraft has cleared its runway.

Verifying the performance of NATS' image recognition model is of critical importance: mistakenly clearing a plane for landing or takeoff before another plane has cleared the runway could lead to a collision. But an established methodology for verifying the performance of an AI-based image recognition model simply does not exist. Moreover, early versions of the model were brittle, failing often and in unexpected ways. For example, British Airways planes have a different paint scheme from most other planes in the model's corpus of examples, and performance worsened considerably as a result.

Fortunately, as with any AI model, adding many more cases to the model's training data has allowed it to achieve far higher levels of accuracy. Perhaps unexpectedly, NATS was able to reach impressively high performance levels in a much shorter time than they would have expected from a more traditional software approach. The latest model has been exposed to 40,000 novel real-world examples, and has achieved very low margins of error. Further testing is required, however, and even if all goes well, the system will be used only to aid air traffic controllers' decision-making.

What can we learn from this example about the aviation industry's appetite for safety-critical AI applications? First, it is striking how serious the need for AI was at the time of implementation. Inclement weather, exceptionally high traffic volumes, and a funding structure that incentivized increased air traffic all helped to generate the demand for an AI application. And despite all of these factors, the image recognition software that NATS ultimately developed, though groundbreaking in the context of air traffic control, is limited to decision support. As one expert knowledgeable about the project emphasized, the goal is not to replace human operators. Instead, "the question is, how do I make the human controller more effective, giving them decision support to handle repetitive tasks?" Other experts echoed that the idea of replacing human operators remains a far-off goal. For now, even in the most ambitious firms working in aviation, the focus is on user support, not replacement.

The effort to bring AI to Heathrow air traffic control marks one of the most ambitious current efforts at bringing AI to a safety-critical use case. It demonstrates both the promise of AI and the relative conservatism of aviation, even at the most innovative end of the spectrum.

CAUSES AND CONSEQUENCES OF AVIATION'S CONSERVATISM

In 1999, *The Economist*, *Fortune Magazine*, and other major outlets ran a seemingly innocuous ad for Airbus' new four-engine A340 aircraft. It showed a single A340 flying over stormy seas under a dark cloudy sky. The caption: "*If you're over the middle of the Pacific, you want to be in the middle of four engines.*"

Airline executives — Airbus' customers — hated the ad, which seemed to imply that crossing the Pacific on an aircraft with just two engines is unsafe. (In reality, a two-engine aircraft is just as safe as a four-engine aircraft.) Gordon Bethune, CEO of Continental Airlines, immediately wrote Airbus President Noel Forgeard complaining that the ad exploited "the unfounded fears of the traveling public." American Airlines VP Tom Horton said the ad "shocked" him.[49]

In the aviation industry, competing on flight safety is taboo.[50] All western airlines and aircraft manufacturers have to meet exceptionally high safety standards; and undermining public confidence in flying's safety hurts all firms in the industry. As a Boeing spokesperson stated in response to the Airbus ad, "To infer that one type of aircraft is safer, or another is riskier, is inaccurate and inappropriate."[51] Airbus stopped its ad campaign, though three years later, it ran a softened version ("4 engines 4 long haul"), which was also quickly canceled after provoking further ire from airlines.[52]

Simply put, flying scares many people. This observation accounts for much of the aviation industry's conservatism and has shaped the industry's approach to safety-critical AI thus far. In surveys measuring the dread that the public feels toward a range of activities and technologies, respondents consistently rank aviation second, behind nuclear power.[53] Aviation safety regulators often morbidly observe that "regulations are written in blood," but it may be more apt to say that they are written in newspaper headlines.

Because accidents in other industries typically provoke much weaker public response, their regulators often have less power to enforce safety standards, and firms have more leeway to use unproven technologies like AI. Experts frequently compare aviation's approach to AI with that of automotive firms developing self-driving vehicles. "The aviation world is a very conservative one," said one expert. "I look at other areas where I would have more concern about the reliance on AI — even down to trials of automated vehicles. Look at the environment they're operating in: it's very uncontrolled." While most interviewees agreed that the automotive industry has taken greater safety risks than the aviation industry, they did not necessarily view this as problematic. As one expert from EUROCONTROL said, "We are quite curious about the automotive industry — we have the impression that they are moving fast. But it's so unsafe to drive that I suspect AI can really help."

The gulf between aviation firms and most internet firms — famously fast adopters of AI — is even bigger. Consider reliability engineering at Netflix, a firm renowned for, among other things, its exceptionally reliable, high-quality streaming services. Those services depend critically on Netflix's ability to deliberately cause system failures and learn from them. The Netflix tech blog avers: "It's no secret that at Netflix we enjoy deliberately breaking things to test our production systems. Doing so lets us validate our assumptions and prove that our mechanisms for handling failure will work when called upon. Netflix has a tradition of implementing a range of tools that create failure."[54] Reliability engineering at Netflix begins with simulations, but ultimately extends to experiments on actual users. Sometimes, deliberate system failures turn out to be more extreme than engineers anticipate, with real consequences for streaming services. But these failures rarely make headlines or inspire new regulations, and thus they typically have limited consequences for the firm's bottom line.[55]

In aviation, by contrast, the costs of air accidents dwarf the learning benefits: the goal is to get as close to failsafe as possible, without ever failing in the real world. This requirement makes development and testing far more expensive and time intensive. It also necessitates a "safety culture" that is foreign to most other industries (Box 5).

Box 5. Crew Resource Management

The safety culture of the aviation industry underpins much of its safety success and is substantially the result of a battery of (expensive) safety training programs.[56] Perhaps the most famous of these is crew (or cockpit) resource management (CRM), a set of training procedures first developed by the U.S. Federal Aviation Administration and now required for airline personnel operating in either U.S. or European airspaces. CRM focuses not on the technical details of flight safety, but on the social and cognitive skills required for gaining and maintaining situational awareness, solving problems, and making decisions. Trainees learn to recognize hazardous attitudes such as "anti-authority," "impulsivity," "invulnerability," and "macho," as well as how to use error management techniques, such as following standard operating procedures, communicating risks, and maximizing safety redundancies.

CRM has applications outside of the cockpit. It has since been adapted to air traffic control, aircraft design and maintenance, and other safety-critical industries such as rail transportation, healthcare, the military, and firefighting. It might be profitably translated to less safety-critical industries as well, but it is unclear whether firms in these industries would willingly pay the cost of training.

Given the lower cost of accidents in other part of the software industry, software firms attempting to build tools for aviation face a steep learning curve. One interviewee had worked with a software firm that developed multiple products for an aviation manufacturer. He noted that "the second time around was much easier. You have to get used to different working procedures and have a good system. In the end, the products were robust: I understand why the regulations are there. We're just used to managing changes in a much more flexible way." He went on to note that, had his company followed its usual testing process, it "would have made several errors. The guy writing software code can't be the guy writing tests. . . . In a generic software company, it's basically the opposite. For example, it's Microsoft's policy that devs should also do the testing, so they learn more from their mistakes. You can't do that in aviation."

An additional challenge to AI development stems from the financial hardship imposed by the long certification process. In military avionics, for example, where the acquisition process is especially harrowing, firms often find themselves stuck in the so-called "Valley of Death": the transitional period between science and technology experimentation and full-scale implementation, where changes in funding mechanisms often leave firms unfunded for 18 to 24 months.[57] Firms like Boeing have the capital to manage this gap, but smaller software firms very often do not.

For all of these reasons, in addition to tackling the technical AI safety problem, AI firms moving into safety-critical domains will likely need to develop new norms, processes, and even business models in order to make the transition as smooth as possible.

Recommendations

Expeditious progress in the field of artificial intelligence could bring significant economic and even safety benefits. But some of the most valuable applications of AI require not only increased capabilities, but also the confidence — among firms and governments, as well as consumers and the general public — that these systems will behave as expected. The challenge going forward will be to develop the bureaucratic and technical know-how necessary to build and certify safety-critical AI systems. This section presents recommendations for both policymakers and technologists seeking to foster the development of highly reliable AI systems, drawing on the insights gained through interviews with experts.

POLICYMAKERS: SCALE UP INVESTMENTS IN TEVV FOR AI SYSTEMS

The case of aviation demonstrates the urgent need for progress in testing, evaluation, verification, and validation (TEVV) for AI systems. Research agendas are developing in this area: for example, DARPA's TrojAI, Explainable AI, and Assured AI programs represent promising steps in this direction. But further funding is warranted. A range of safety-critical tasks that could benefit from AI solutions cannot be implemented because of the currently weak technical understanding of these systems. Funding should not focus only on achieving breakthroughs in technical AI research; domains like transportation and healthcare will need resources to adapt their usual certification processes to account for specific features of AI, such as its heavy dependence on data. Funding can help accelerate this process, and unlock significant economic benefits across a range of industries.

REGULATORS: COLLABORATE ON AI SAFETY STANDARDS AND INFORMATION-SHARING

The risk of regulatory races to the bottom on AI safety, at least between Europe and the United States, appears relatively low in the aviation industry. At the same time, regulators in Europe did express worries that the United States and China might more quickly adapt to an AI future, and that this might create pressures to compromise on safety. Active efforts to foster collaboration across regulators, including China's Civil Aviation Administration, will help to ensure a unified and uniformly safe set of AI regulations that firms can rise to meet.

More generally, regulators on both sides of the Atlantic expressed great interest in improving communication and information-sharing across regulatory agencies. The same is true of those working on AI safety challenges outside of the aviation industry. Multiple interviewees expressed the sentiment that they are all working on the same problem from different angles, and that they would benefit from further cross-industry collaboration on AI safety.

Meanwhile, regulators in other industries where AI plays an important role could consider applying lessons from the aviation industry. For example, regulators should provide incentives for firms to share information on AI accidents and near-misses.[58] Aviation regulators have deliberately developed forums and incentives for sharing information about possible safety hazards. Firms must recognize that they will not be punished for being open about mistakes, and that they will benefit from learning about others' safety difficulties. Investigations into the Max 737 accidents suggest that, while aviation has historically maintained relatively high levels of openness, stronger incentives may be needed to encourage openness.

FIRMS: PAY THE PRICE FOR SAFETY IN ADVANCE

The aviation industry has learned the hard way that it is always worth paying for safety upfront, rather than on the back end. The received wisdom in the industry is that $1 spent discovering a safety vulnerability can obviate the need for $10 spent patching the vulnerability during testing, or $100 spent patching the vulnerability after rollout of a product. The AI field has largely inherited safety norms from the internet industry, which traditionally has worked on products with a very different safety-efficiency tradeoff. But many of AI's use cases are in physical domains, where far higher levels of safety and reliability are required. Firms willing to invest the time and expense in developing safe and reliable AI systems may have an advantage in transitioning into these more safety-critical industries.

A further challenge lies in establishing improved safety norms surrounding AI development and implementation. Unfortunately, a catastrophic AI accident at one firm likely has spillover consequences for other firms seeking to use AI as well. It is easy to lose the public's trust and hard to regain it, especially in safety-critical domains.

Partnerships might help AI firms learn the safety norms and practices of safety-critical industries. The case of Searidge Technologies' partnership with NATS on air traffic control at Heathrow airport offers an excellent example of the value of partnerships between AI firms and aviation firms. Interviewees attested to the great challenges technology firms face in breaking into safety-critical industries like aviation, due to the higher safety demands and regulatory oversight. But they also noted the abundance of structured data awaiting firms that manage to break into the aviation industry. Partnerships between experienced avionics firms and skilled AI firms appear promising.

RESEARCHERS: EXPLORE AI SAFETY RACING DYNAMICS ON AN INDUSTRY-BY-INDUSTRY AND ISSUE-BY-ISSUE BASIS

The results presented in this paper suggest that the aviation industry has thus far moved slowly toward the development and adoption of safety-critical AI. Yet competition may lead to very different dynamics in other industries and issue areas. Other industries with less mature regulatory regimes will likely be more susceptible to races to the bottom on safety. And Microsoft's call for regulations to prevent a race to the bottom in facial recognition technologies suggests that safety is far from the only area in which AI race dynamics could lead to socially harmful outcomes.

One particularly interesting industry to consider in light of the foregoing analysis is the U.S. defense industry. Given that fears of a race to the bottom have often made analogies to arms racing during the Cold War, a close look at how competition has affected the pace of the U.S. military's adoption of AI seems warranted. Only two experts interviewed in this report had military experience. However, both experts agreed that the U.S. military has in fact proceeded even more slowly than the aviation industry in adopting safety-critical AI, in part because of the especially serious consequences of safety failures within the military.

Conclusion

As AI systems become more capable and begin to diffuse into safety-critical domains, technical researchers have begun to voice serious concerns about accident risks flowing from novel AI applications. AI systems are often unpredictable, opaque, and data-intensive in ways that defy the constraints of existing safety management approaches. Yet AI systems are also economically and strategically valuable, leading some to worry about the prospect of a "race to the bottom" in AI safety.

A deep dive into the aviation industry provides grounds for optimism that, in at least one safety-critical domain, firms and regulators are approaching AI tentatively, with ample awareness of the risks these systems pose. Experts across the aviation industry attested to the importance of learning slowly about AI, experimenting first with the least safety-critical applications and investing time and money in improving understanding of these systems before moving toward valuable but more safety-critical applications. For now, there is little sign that competitive pressures are sufficient to overwhelm safety imperatives.

This examination of the aviation industry has important implications for policymakers, regulators, and firms concerned about AI accident risks. To unlock the benefits of AI, in both aviation and other safety-critical industries, this paper recommends further investment in methods for testing, evaluating, verifying, and validating the safety, security, and reliability of AI systems; encouraging inter-regulator collaboration on tackling the unique challenges posed by AI; and paying the price for AI safety upfront, for example, by investing in independent verification and validation.

The results also have implications for AI policy researchers concerned about a race to the bottom on safety. Regulations will likely play a key role in determining the form of racing dynamics in different industries. Researchers should invest further effort in identifying specific industries where a race to the bottom may be likely or especially harmful, and explore what role standards, regulations, norms, and processes might play in mitigating the risk of AI accidents within each of these industries.

Acknowledgments

This paper would not have been possible without the support of the AI Security Initiative and the Center for Long Term Cybersecurity. Thanks to Ann Cleaveland, Steve Weber, Charles Kapelke, and especially Jessica Cussins Newman for excellent guidance, feedback, and support throughout the research and writing process. For helpful feedback and discussion, many thanks to Ashwin Acharya, Christopher Ansell, Jared Brown, Peter Cihon, Ann Cleaveland, Todd LaPorte, Paul Schulman, Mojmir Stehlik, and Remco Zwetsloot. And special thanks go to the many anonymous interviewees who allowed me to barrage them with questions about AI, aviation, and safety. Any errors are my own.

About the Author

Will Hunt is a graduate researcher at the AI Security Initiative and a PhD student in political science at the University of California, Berkeley. He previously served as a visiting researcher at the Center for Security and Emerging Technologies at Georgetown University, and as a summer research fellow at the University of Oxford. He attended Deep Springs College and holds a B.A. from Yale.

Endnotes

[1] Larry Lewis, “AI Safety: Charting Out the High Road,” *War on the Rocks* (December 9, 2019), https://warontherocks.com/2019/12/ai-safety-charting-out-the-high-road/; Richard Danzig, “Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority,” *Center for a New American Security* (June 2018), https://s3.amazonaws.com/files.cnas.org/documents/CNASReport-Technology-Roulette-DoSproof2v2.pdf?mtime=20180628072101.

[2] Danielle C. Tarraf, William Shelton, Edward Parker, Brien Alkire, Diana Gehlhaus, Justin Grana, Alexis Levedahl, Jasmin Leveille, Jared Mondschein, James Ryseff, Ali Wyne, Dan Elinoff, Edward Geist, Benjamin N. Harris, Eric Hui, Cedric Kenney, Sydne Newberry, Chandler Sachs, Peter Schirmer, Danielle Schlang, Victoria M. Smith, Abbie Tingstad, Padmaja Vedula, and Kristin Warren, *The Department of Defense Posture for Artificial Intelligence: Assessment and Recommendations*. Santa Monica, CA: RAND Corporation, 2019. https://www.rand.org/pubs/research_reports/RR4229.html.

[3] Brian Fung and Rachel Metz, “This may be America’s first wrongful arrest involving facial recognition,” *CNN Business,* June 24, 2020, https://www.cnn.com/2020/06/24/tech/aclu-mistaken-facial-recognition/index.html. “Drones, robots, lasers, hypersonic gliders & other high-tech arms: Putin wants Russian military to be up to any future challenge,” *RT* (November 22, 2019), https://www.rt.com/russia/474119-putin-laser-drone-robot-hypersonic/

[4] For a review of this trend and discussion of the shortcomings of the AI arms race narrative, see Remco Zwetsloot, Helen Toner, and Jeffrey Ding, “Beyond the AI Arms Race,” *Foreign Affairs,* November 16, 2018, https://www.foreignaffairs.com/reviews/review-essay/2018-11-16/beyond-ai-arms-race.

[5] Paul Scharre, “Killer Apps: The Real Danger of an AI Arms Race,” *Foreign Affairs*, May/June 2019, https://www.foreignaffairs.com/articles/2019-04-16/killer-apps.

[6] “OpenAI Charter,” *OpenAI*, April 9, 2018, https://openai.com/charter/. See also Amanda Askell, Miles Brundage, and Gillian Hadfield, “The Role of Cooperation in Responsible AI Development,” *arXiv*, July 10, 2019, https://arxiv.org/pdf/1907.04534.pdf.

[7] Brad Smith, “Facial Recognition: It’s Time for Action,” *Microsoft*, December 16, 2018, https://blogs.microsoft.com/on-the-issues/2018/12/06/facial-recognition-its-time-for-action/.

[8] The academic literature asserting and refuting claims of on races to the bottom is extensive and covers many domains. I offer a brief review of some highlights from this literature in the next section.

[9] Daniel Fagnant and Kara M. Kockelman, “Preparing a nation for autonomous vehicles: Opportunities, barriers and policy recommendations,” *Eno Center for Transportation* (October 2013), https://www.enotrans.org/wp-content/uploads/AV-paper.pdf; Mikio Yanagisawa, Wassim G. Najm, and Paul Rau, “Preliminary Estimates of Target Crash Populations for Concept Automated Vehicle Functions,” in *Proceedings of the In 24th International Technical Conference on the Enhanced Safety of Vehicles* (*ESV*), 1–11, 2015, https://www-esv.nhtsa.dot.gov/Proceedings/25/25ESV-000266.pdf.

[10] Aseem Prakash and Matthew Potoski, “Racing to the bottom? Trade, environmental governance, and ISO 14001,” *American Journal of Political Science* 50, no. 2 (2006): 350–364; Beth A. Simmons and Zachary Elkins. “The

globalization of liberalization: Policy diffusion in the international political economy." *American Political Science Review* (2004): 171–189; Michael A. Bailey, Mark Carl Rom, and Matthew M. Taylor, "State competition in higher education: A race to the top, or a race to the bottom?" *Economics of Governance* 5, no. 1 (2004): 53–75. Prakash, Aseem, and Kelly L. Kollman. "Biopolitics in the EU and the US: A Race to the Bottom or Convergence to the Top?" *International Studies Quarterly* 47, no. 4 (2003): 617–641. John Braithwaite and Peter Drahos, *Global Business Regulation*, Cambridge, UK: Cambridge University Press, 2000; Berliner, Daniel, Anne Regan Greenleaf, Milli Lake, Margaret Levi, and Jennifer Noveck. "Governing global supply chains: what we know (and don't) about improving labor rights and working conditions." *Annual Review of Law and Social Science* (November 2015), https://www.annualreviews.org/doi/pdf/10.1146/annurev-lawsocsci-120814-121322.

11 William L. Cary, "Federalism and corporate law: reflections upon Delaware," *The Yale Law Journal* 83, no. 4 (1974): 663–705.

12 "About the Division of Corporations," Delaware Division of Corporations, https://corp.delaware.gov/aboutagency/.

13 Mary Graham, "Environmental Protection & the States: 'Race to the Bottom" or "Race to the Bottom Line?" *Brookings*, December 1, 1998, https://www.brookings.edu/articles/environmental-protection-the-states-race-to-the-bottom-or-race-to-the-bottom-line/.

14 See for example Rose Aguilar, "The Trans-Pacific Partnership Will Lead to a Global Race to the Bottom," *The Guardian*, May 18, 2015, https://www.theguardian.com/commentisfree/2015/may/08/the-trans-pacific-partnership-will-lead-to-a-global-race-to-the-bottom.

15 Lewis, "AI Safety: Charting Out the High Road."

16 Danzig, "Technology Roulette."

17 Zwetsloot, Toner, and Ding, "Beyond the AI Arms Race."

18 Elsa Kania, "The pursuit of AI is more than an arms race." *Defense One*, April 19, 2019, https://www.defenseone.com/ideas/2018/04/pursuit-ai-more-arms-race/147579/; Scharre, "Killer Apps: The Real Danger of an AI Arms Race"; Danzig, "Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority."

19 This is not to say that the development of nuclear weapons has no relevance as an analogy for AI development. But the "arms race" narrative glosses over large differences between AI systems — which comprise a diverse set of enabling, general purpose, highly commercially valuable technologies — and the nuclear bomb — an exceptionally important military innovation of no commercial use,

20 Edward Geist, "It's already too late to stop the AI arms race — We must manage it instead," *Bulletin of the Atomic Scientists* 72, Vol. 5, 318-321, https://www.tandfonline.com/doi/pdf/10.1080/00963402.2016.1216672; Michael Auslin, "Can the pentagon win the AI arms race?" *Foreign Affairs*, October 19, 2018, https://www.foreignaffairs.com/articles/united-states/2018-10-19/can-pentagon-win-ai-arms-race; Gregory Allen and Elsa Kania, "China is using America's own plan to dominate the future of artificial intelligence," Foreign Policy, September 8, 2017, https://foreignpolicy.com/2017/09/08/china-is-using-americas-own-plan-todominate-the-future-of-artificial-intelligence/; Julian E. Barnes and Josh Chin, "The New Arms Race in AI," *The Wall Street Journal,* March 2, 2018, https://www.wsj.com/articles/the-new-arms-race-in-ai-1520009261; Nicholas Thompson and Ian Bremmer, "The AI Cold War That Threatens Us All," *Wired,* October 23, 2018, https://www.wired.com/story/ai-cold-war-china-could-doom-us-all/.

[21] See for example Daniel W. Drezner, "The global governance of the Internet: Bringing the state back in." Political science quarterly 119, no. 3 (2004): 477–498 and Scott J. Bassinger and Mark Hallerberg, "Remodeling the competition for capital: How domestic politics erases the race to the bottom," American Political Science Review (2004): 261–276.

[22] See especially Lynn M. LoPucki and Sara D. Kalin, "Failure of Public Company Bankruptcies in Delaware and New York: Empirical Evidence of a Race to the Bottom," *Vanderbilt Law Review* 54 (2001): 231.

[23] David Vogel, "Trading up and governing across: transnational governance and environmental protection," *Journal of European public policy* 4, no. 4 (1997): 556–571.

[24] Anu Bradford, "The Brussels effect," Northwestern University Law Review 107 (2012): 1–68.

[25] Early work on AI races to the bottom focused on research teams to develop artificial general intelligence (AGI) capabilities, reasoning that the great economic and strategic value of AGI might lead to corner cutting on safety. See in particular Stuart Armstrong, Nick Bostrom, and Carl Shulman, "Racing to the Precipice," FHI Technical Report #2013-1, 2013, https://www.fhi.ox.ac.uk/wp-content/uploads/Racing-to-the-precipice-a-model-of-artificial-intelligence-development.pdf; Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford: Oxford University Press, 2014. This paper focuses instead on narrow AI systems, though the results may offer some insight into the more speculative and longer-term AI trajectories studied in this earlier work.

[26] Askell, Brundage, and Hadfield, "The Role of Cooperation in Responsible AI Development."

[27] For a mapping of current efforts to set global standards for AI, see Peter Cihon, "Standards for AI Governance: International Standards to Enable Global Coordination in AI Research and Development," Future of Humanity Institute Technical Report, 2019, https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf.

[28] Allan Dafoe, "AI Governance: A Research Agenda," Center for the Governance of AI, August 27, 2018, https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf.

[29] "Recommendation of the Council on Artificial Intelligence," OECD Legal Instruments, May 21, 2019, https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449.

[30] Jose Hernandez-Orallo, Fernando Martinez-Plumed, Shahar Avin, and Sean O hEigeartaigh, "Surveying Safety-relevant AI Characteristics," Proceedings of SafeAI, Association for the Advancement of Artificial Intelligence conference (AAAI) (2019), http://ceur-ws.org/Vol-2301/paper_22.pdf.

[31] Arnold Barnett, "Aviation Safety: A Whole New World?" Aviation Science 54, No. 1, https://pubsonline.informs.org/doi/10.1287/trsc.2019.0937.

[32] Ibid.

[33] Scott A. Shappell, and Douglas A. Wiegmann, "U.S. naval aviation mishaps 1977–92: Differences between single- and dual-piloted aircraft," *Aviation, Space, and Environmental Medicine* 67 (1996), 65–9.

[34] Keller Scholl and Robin Hanson, "Testing the Automation Revolution Hypothesis," GMU Working Paper in Economics (December 27, 2019), No. 19–42, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3496364.

[35] "Airbus demonstrates first fully automatic vision-based takeoff," *Airbus*, January 16, 2020, https://www.airbus.com/newsroom/press-releases/en/2020/01/airbus-demonstrates-first-fully-automatic-visionbased-takeoff.html.

36 Ibid.

37 Scholl and Hanson, “Testing the Automation Revolution Hypothesis.”

38 Vance Hilderman and Tony Baghai, *Avionics Certification: A Complete Guide to DO-178* (*software*)*, DO-254* (*hardware*)*,* Leesburg, VA: Avionics Communications, Inc., 2007.

39 Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané, “Concrete problems in AI safety,” *arXiv* (2016), https://arxiv.org/pdf/1606.06565.pdf; Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg, “Scalable agent alignment via reward modeling: a research direction,” *arXiv* (2018), https://arxiv.org/pdf/1811.07871.pdf; Ion Stoica, Dawn Song, Raluca Ada Popa, David Patterson, Michael W. Mahoney, Randy Katz, Anthony D. Joseph, Michael Jordan, Joseph M. Hellerstein, Joseph E. Gonzalez, Ken Goldberg, Ali Ghodsi, David Culler, Pieter Abbeel. “A Berkeley View of Systems Challenges for AI.” *arXiv* (2017), https://arxiv.org/pdf/1712.05855.pdf; Jessica Taylor, Eliezert Yudkowsky, Patrick LaVictoire, and Andrew Critch, “Alignment for Advanced Machine Learning Systems,” *Machine Intelligence Research Institute* (2016), https://intelligence.org/files/AlignmentMachineLearning.pdf.

40 Brian A. Haugh, David A. Sparrow, and David M. Tate, “The Status of Test, Evaluation, Verification, and Validation (TEV&V) of Autonomous Systems,” Institute for Defense Analyses (September 2018), https://www.ida.org/-/media/feature/publications/t/th/the-status-of-test-evaluation-verification-and-validation-of-autonomous-systems/p-9292.ashx.

41 Note that annual fatality rates are far higher for privately operated aircraft, ranging from 399 to 1,328 over the past ten years. “Death Rates Per Year,” Bureau of Aircraft Accident Archives, https://www.baaa-acro.com/statistics/death-rate-per-year. Commercially scheduled flights are subject to much stricter regulations than private flights.

42 SAE’s G-34 group and EUROCAE’s WG 114 group had previously worked independently on developing standards for AI in aviation, until making the decision to join forces.

43 It is also consistent with the recommendations from “The FLY AI Report,” European Aviation Artificial Intelligence High-Level Group, March 5, 2020, https://www.eurocontrol.int/publication/fly-ai-report.

44 Dominic Gates, “Inspector General report details how Boeing played down MCAS in original 737 Max certification — and FAA missed it,” *Seattle Times*, June 30, 2020, https://www.seattletimes.com/business/boeing-aerospace/inspector-general-report-details-how-boeing-played-down-mcas-in-original-737-max-certification-and-faa-missed-it/.

45 David Gelles, Natalie Kitroeff, Jack Nicaz, and Rebecca R. Ruiz, “Boeing was ‘Go, go, go’ to beat Airbus with the 737 Max,” *New York Times*, March 23, 2019, https://www.nytimes.com/2019/03/23/business/boeing-737-max-crash.html.

46 Ibid.

47 Dominic Rushe, “Boeing puts cost of 737 Max crashes at $19bn as it slumps to annual loss,” *The Guardian*, January 29, 2020, https://www.theguardian.com/business/2020/jan/29/boeing-puts-cost-of-737-max-crashes-at-19bn-as-it-slumps-to-annual-loss.

48 Why doesn’t London build more runways? Its airports have tried: In the past three decades, for example, London Heathrow Airport alone has submitted three proposals for the construction of a new runway. Objections ranging from noise pollution to environmental concerns have blocked each successive proposal. The result is that London has built no new full-service runways since World War II. Meanwhile, demand for service to London airports is ever-increasing. “Why it has taken 70 years to build a new runway for London,” *The Economist*, October 16, 2016, https://www.economist.com/the-economist-explains/2016/10/14/why-it-has-taken-70-years-to-build-a-new-runway-for-london.

Measuring the exact increase in air traffic to and from London is difficult due to lack of data. However, London Heathrow Airport hosted approximately 62,000 flights per year in 1953. In 2019, it hosted approximately 476,000 flights, a nearly eight-fold increase. (See Alan Gallop, *Time Flies: Heathrow At 60*, Stroud: Sutton Publishing, 2005. p. 96 and Civil Aviation Authority, "Airport Data," https://www.caa.co.uk/Data-and-analysis/UK-aviation-market/Airports/Datasets/.)

49 Daniel Michaels, "Airbus Industrie's Ad Raises Safety Questions," *The Wall Street Journal*, November 18, 1999, https://www.wsj.com/articles/SB942879444311000425.

50 Interestingly, airlines do appear to compete on safety in other, less regulated domains. For example, airlines have differed in their responses to the COVID-19 pandemic: some (mostly higher-end) airlines like Alaska and Delta have blocked passengers from booking middle seats, while so-called "ultra low cost airlines" like Sun Country, Spirit, and Allegiant largely have not.

51 Ibid.

52 Lynn Lunsford, "Critics Say Ad Implies Rival Jet Is Unsuitable for Long Flights," *The Wall Street Journal,* July 26, 2002, https://www.wsj.com/articles/SB1027622698744765040.

53 Katherine T. Fox-Glassman and Elke U. Weber, "What makes risk acceptable? Revisiting the 1978 psychological dimensions of perceptions of technological risks," *Journal of Mathematical Psychology* 75 (December 2016), https://www.sciencedirect.com/science/article/pii/S002224961630027X.

54 Kolton Andrus, Naresh Gopalani, Ben Schmaus, "Failure Injection Testing," *The Netflix Tech Blog*, October 23, 2014, https://netflixtechblog.com/fit-failure-injection-testing-35d8e2a9bb2.

55 Of course, even on the internet, AI accidents can have consequences. For example, in 2015, Google Images infamously returned pictures of black people in response to queries for "gorillas," generating widespread negative media coverage and necessitating a public apology. By 2018, Google still had not found a fix to the problem short of censoring words like "gorilla," "chimp," and "chimpanzee" from its database, noting that "image labeling technology is still early and unfortunately it's nowhere near perfect." This incident is part of a broader trend in which technical difficulties of testing AI systems and the lack of diversity in the tech industry together lead to increased accident risk. Tom Simonite, "When It Comes to Gorillas, Google Photos Remains Blind," *Wired*, January 11, 2018, https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/.

56 There exists an extensive literature on safety culture in the aviation industry. See in particular, Dominic M. Cooper, "Towards a model of safety culture," *Safety science* 36, no. 2 (2000): 111–136; Frank W. Guldenmund, "The nature of safety culture: a review of theory and research," *Safety science* 34, no. 1–3 (2000): 215–257; and Andrew Hopkins, "Safety culture, mindfulness and safe behaviour: Converging ideas?" *National Research Centre for OHS Regulation*, Working Paper 7 (2002).

57 Defense Procurement and Acquisition Policy, "Manager's Guide to Technology Transition in an Acquisition Environment," Office of the Undersecretary of Defense, January 31, 2003, 4–22, https://www.acq.osd.mil/dpap/docs/aq201s1v10complete.pdf.

58 Miles Brundage et al., "Toward trustworthy AI development: mechanisms for supporting verifiable claims," *arXiv preprint arXiv:2004.07213* (2020), https://arxiv.org/abs/2004.07213.

CLTC

Center for Long-Term Cybersecurity

UC Berkeley

Center for Long-Term Cybersecurity
cltc.berkeley.edu
@CLTCBerkeley