Issue Brief

AI Safety and Automation Bias

The Downside of Human-in-the-Loop

Authors
Lauren Kahn
Emelia S. Probasco
Ronnie Kinoshita

CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

November 2024

Executive Summary

Automation bias is the tendency for an individual to over-rely on an automated system. It can lead to increased risk of accidents, errors, and other adverse outcomes when individuals and organizations favor the output or suggestion of the system, even in the face of contradictory information.

Automation bias can endanger the successful use of artificial intelligence by eroding the user's ability to meaningfully control an AI system. As AI systems have proliferated, so too have incidents where these systems have failed or erred in various ways, and human users have failed to correct or recognize these behaviors.

This study provides a three-tiered framework to understand automation bias by examining the role of users, technical design, and organizations in influencing automation bias. It presents case studies on each of these factors, then offers lessons learned and corresponding recommendations.

User Bias: Tesla Case Study

Factors influencing bias:

- User's personal knowledge, experience, and familiarity with a technology.
- User's degree of trust and confidence in themselves and the system.

Lessons learned from case study:

- Disparities between user perceptions and system capabilities contribute to bias and may lead to harm.

Recommendation:

- **Create and maintain qualification standards for user understanding.** User misunderstanding of a system's capabilities or limitations is a significant contributor to incidents of harm. Since user understanding is critical to safe operation, system developers and vendors must invest in clear communications about their systems.

Technical Design Bias: Airbus and Boeing Design Philosophies Case Study

Factors influencing bias:

- The system's overall design, user interface, and how it provides user feedback.

Lessons learned from case study:

- Even with highly trained users such as pilots, systems interfaces contribute to automation bias.
- Different design philosophies have different risks. No single approach is necessarily perfect, and all require clear, consistent communication and application.

Recommendation:

- **Value and enforce consistent design and design philosophies that account for human factors, especially for systems likely to be upgraded.** When necessary, justify and make clear any departures from a design philosophy to legacy users. Where possible, develop common design criteria, standards, and expectations, and consistently communicate them (either through organizational policy or industry standard) to reduce the risk of confusion and automation bias.

Organizational Policies and Procedure Bias: Army Patriot Missile System vs. Navy AEGIS Combat System Case Study

Factors influencing bias:

- Organizational training, processes, and policies.

Lessons learned from case study:

- Organizations can employ the same tools and technologies in very different ways based on protocols, operations, doctrine, training, and certification. Choices in each of these areas of governance can embed automation biases.
- Organizational efforts to mitigate automation bias can be successful but mishaps are still possible, especially when human users are under stress.

Recommendation:

- Where autonomous systems are used by organizations, **design and regularly review organizational policies appropriate for technical capabilities and organizational priorities.** Update policies and processes as technologies change to best account for new capabilities and mitigate novel risks. If there is a mismatch between the goals of the organization and policies governing how capabilities are used, automation bias and poor outcomes are more likely.

Across these three case studies, it is clear that "human-in-the-loop" cannot prevent all accidents or errors. Properly calibrating technical and human fail-safes for AI, however, poses the best chance for mitigating the risks of using AI systems.

Table of Contents

Introduction

In contemporary discussions about artificial intelligence, a critical but often overlooked aspect is automation bias—the tendency of human users to overly rely on AI systems. Left unaddressed, automation bias can and has harmed both AI and autonomous system users and innocent bystanders in examples that range from false legal accusations to death. Automation bias, therefore, presents a significant challenge in the real-world application of AI, particularly in high-stakes contexts such as national security and military operations.

Successful deployment of AI systems relies on a complex interdependence between AI systems and the humans responsible for operating them. Addressing automation bias is necessary to ensure successful, ethical, and safe AI deployment, especially when the consequences of overreliance or misuse are most severe. As societies incorporate AI into systems, decision-makers thus need to be prepared to mitigate the risks associated with automation bias.

Automation bias can manifest and be intercepted at the user, technical design, and organizational levels. We provide three case studies that explain how factors at each of these levels can make automation bias more or less likely, derive lessons learned, and highlight possible mitigation strategies to alleviate these complex issues.

What Is Automation Bias?

Automation bias is the tendency for a human user to overly rely on an automated system, reflecting a cognitive bias that emerges from the interaction between a human and an AI system.

When affected by automation bias, users tend to decrease their vigilance in monitoring both the automated system and the task it is performing.[1] Instead, they place excessive trust in the system's decision-making capabilities and inappropriately delegate more responsibility to the system than it is designed to handle. In severe instances, users might favor the system's recommendations even when presented with contradictory evidence.

Automation bias most often presents in two ways: as an error of omission, when a human fails to take action because the automation did not alert them (as discussed in the first case study on vehicles); or as an error of commission, when a human follows incorrect directions from the automation (as discussed in the case study on the Patriot Missile System).[2] In this analysis, we also discuss an instance where a bias against the automation causes harm (i.e., the third case study on the AEGIS weapons system). Automation bias does not always result in catastrophic events, but it increases the likelihood of such outcomes. Mitigating automation bias can help to improve human oversight, operation, and management of AI systems and thus mitigate some risks associated with AI.

The challenge of automation bias has only grown with the introduction of progressively more sophisticated AI-enabled systems and tools across different application areas including policing, immigration, social welfare benefits, consumer products, and militaries (see Box 1). Hundreds of incidents have occurred where AI, algorithms, and autonomous systems were deployed without adequate training for users, clear communication about their capabilities and limitations, or policies to guide their use.[3]

Box 1. Automation Bias and the UK Post Office Scandal

In a notable case of automation bias, a faulty accounting system employed by the UK Post Office led to the wrongful prosecution of 736 UK sub-postmasters for embezzlement. Although it did not involve an AI system, automation bias and the myth of "infallible systems" played a significant role—users willingly accepted system errors despite substantial evidence to the contrary, favoring the unlikely case that hundreds of postmasters were involved in theft and fraud.[4] As one author of an ongoing study into the case highlighted, "This is not a scandal about technological failing; it is a scandal about the gross failure of management."[5]

While automation bias is a challenging problem, it is a tractable issue that society can tackle throughout the AI development and deployment process. The avenues through which automation bias can manifest—namely at the user, technical, and organizational levels—also represent points of intervention to mitigate automation bias.

A Framework for Understanding and Mitigating Automation Bias

Technology must be fit for purposes, and users must understand those purposes to be able to appropriately control systems. Furthermore, knowing when to trust AI and when and how to closely monitor AI system outputs is critical to its successful deployment.[6] A variety of factors calibrate trust and reliance in the minds of operators, and they generally fall into one of three categories (though each category can be shaped by the context within which the interaction may occur, such as situations of extreme stress or, conversely, fatigue):[7]

- factors intrinsic to the human user, such as biases, experience, and confidence in using the system;
- factors inherent to the AI system, such as its failure modes (the specific ways in which it might malfunction or underperform) and how it presents and communicates information; and,
- factors shaped by organizational or regulatory rules and norms, mandatory procedures, oversight requirements, and deployment policies.

Organizations implementing AI must avoid myopically focusing only on the technical "machine" side to ensure the successful deployment of AI. Management of the human aspect of these systems deserves equal consideration, and management strategies should be adjusted according to context.

Recognizing these complexities and potential pitfalls, this paper presents case studies for three controllable factors affecting automation bias (user, technical, organizational) that correspond to the aforementioned factors that shape the dynamics of human-machine interaction (see Table 1).

Table 1. Factors Affecting Automation Bias

Factors	Description	Case Study
User	User's personal knowledge, experience, and familiarity with a technology User's degree of trust and confidence in themselves and the system	Tesla and driving automation
Technical Design	The system's overall design, the structure of its user interface, and how it provides user feedback	Airbus and Boeing design philosophies
Organization	Organizational processes shaping AI use and reliance	U.S. Army's management and operation of the Patriot Missile System vs. U.S. Navy's management and operation of the AEGIS Combat System

An additional layer of task-specific factors, such as time constraints, task difficulty, workload, and stress, can exacerbate or alternatively reduce automation bias.[8] These factors should be duly considered in the design of the system, as well as training and organizational policies, but are beyond the scope of this paper.

Case Studies

Case Study 1: How User Idiosyncrasies Can Lead to Automation Bias

Individuals bring their personal experiences—and biases—to their interactions with AI systems.[9] Research shows that greater familiarity and direct experience with self-driving cars and autonomous vehicle technologies make individuals more likely to support autonomous vehicle development and consider them safe to use. Conversely, behavioral science research demonstrates that a lack of technological knowledge can lead to fear and rejection, while having only a little familiarity with a particular technology can result in overconfidence in its capabilities.[10] The case of increasingly "driverless" cars illustrates how the individual characteristics and experiences of users can shape their interactions and automation bias. Furthermore, as the case study on Tesla below illuminates, even system improvements designed to mitigate the risks of automation bias may have limited effectiveness in the face of a person's bias.

Tesla's Road to Autonomy

Cars have become increasingly automated over time. Manufacturers and engineers have introduced cruise control and a flurry of other advanced driver assistance systems (ADAS) aimed at improving driving safety and reducing the likelihood of human error, alongside other features such as lane drift systems and blind spot sensors. The U.S. National Highway Traffic Safety Administration suggests that full automation has the potential to "offer transformative safety opportunities at their maturity," but caveat that these are a future technology.* As they make clear on their website in bolded capital letters, cars that perform "all aspects of the driving task while you, as the driver, are available to take over driving if requested. . . **ARE NOT AVAILABLE ON TODAY'S VEHICLES FOR CONSUMER PURCHASE IN THE UNITED STATES.**"[11] Even if these

* The Society of Automotive Engineers (SAE) (in collaboration with the International Organization for Standardization, or ISO) has established six levels of driving automation, from 0 to 5. Level 0, or no automation, represents cars without systems such as adaptive cruise control. On the other end of the spectrum, Levels 4 and 5 suggest cars that may not even require a steering wheel to be installed. Levels 1 and 2 include those systems with increasingly competent driver support features like those mentioned above. In all of these systems, however, the human is driving, "even if your feet are off the pedals and you are not steering." It is at Level 3, where automation begins to take over, that the line between "self-driving" and "driverless" becomes fuzzier, with the vehicle relying less on the driver unless the vehicle requests their engagement. Levels 4 and 5 never require human intervention. See "SAE Levels of Driving Automation™ Refined for Clarity and International Audience," SAE International Blog, May 3, 2021, https://www.sae.org/blog/sae-j3016-update.

cars were available, it is important to consider the possibility that while autonomy might eliminate certain kinds of accidents or human errors (like distracted driving), it has the potential to create new ones (like over-trusting autopilot).[12]

Studies suggest that ADAS adoption by drivers is often opportunistic, and simply a byproduct of upgrading their vehicles. Drivers learn about the vehicle's capabilities in an ad-hoc manner, sometimes just receiving an over-the-air software update that comes with written notes. There are no exams or certifications required for these updates.

Studies have also shown that where use of an ADAS system is solely experiential, such as when a driver adopts an autonomous vehicle without prior training, human misuse or misunderstanding of ADAS systems can happen after only a few encounters behind the wheel.[13] Furthermore, at least one study found that drivers who are exposed to more capable automated systems first tended to establish a baseline of trust when interacting with other (potentially less capable) automated systems.[14] This trust and confidence in ADAS vehicles can manifest as distracted driving, to the point of drivers ignoring warnings, taking longer to react to emergencies, or taking risks they would not take in the absence of automation.[15]

Behind the Wheel: Tesla's Autopilot and the Human Element

In the weeks leading up to the first fatal U.S. accident involving Tesla's Autopilot in 2016, the company's then-president, Jon McNeill, personally tested the system in a Model X. In an email following his test, McNeill praised the system's seemingly flawless performance, admitting, "I got so comfortable under Autopilot that I ended up blowing by exits because I was immersed in emails or calls (I know, I know, not a recommended use)."[16]

Despite marketing that suggests the Tesla Full Self-Driving Capability (FSD) might achieve full autonomy without human intervention, these features currently reside firmly within the suite of ADAS capabilities.[17] Investigations into that first fatal accident found that the driver had been watching a movie and had ignored multiple alerts to maintain hands on the wheel when the Autopilot failed to distinguish a white trailer from a bright sky, leading to a collision that killed the driver.[18] Since then, there have been a range of incidents involving Tesla's Autopilot suite of software, which includes what is called a "Full Self-Driving Capability." These incidents led the National Highway Traffic Safety Administration (NHTSA) to examine nearly one thousand crashes and launch over 40 investigations into accidents in which Autopilot features were reported to have been in use.[19] In its initial investigations, NHTSA found "at least 13 crashes involving one or more fatalities and many more involving serious injuries in which foreseeable driver misuse of the system played an apparent role."[20] Also, among NHTSA's conclusions was that "Autopilot's design was not sufficient to maintain drivers' engagement."[21]

In response to NHTSA's investigation and increasing scrutiny, in December 2023 Tesla issued a safety recall of two million of its vehicles equipped with the Autosteer functionality.[22] In its recall announcement, Tesla acknowledged that:

> "In certain circumstances when Autosteer is engaged, the prominence and scope of the feature's controls may not be sufficient to prevent driver misuse of the SAE Level 2 advanced driver-assistance feature."[23]

As a part of this recall, Tesla sought to address the driver engagement problem with an over-the-air software update that added more controls and alerts to "encourage the driver to adhere to their continuous driving responsibility whenever Autosteer is engaged." That encouragement manifested as:

> "increasing the prominence of visual alerts on the user interface, simplifying engagement and disengagement of Autosteer, additional checks upon engaging Autosteer and … eventual suspension from Autosteer use if the driver repeatedly fails to demonstrate continuous and sustained driving responsibility while the feature is engaged."[24]

Training or certification was not included with the software update; however, a text summary of the software update was provided for users to optionally review, and videos of users indicate that the instructions were easy to ignore. Users also had the option to ignore safety features in the update altogether. The efficacy of these specific changes (either individually or in total) is not yet clear. In April 2024, NHTSA launched a new investigation into Tesla's Autosteer and the software update it performed in December 2023 but, as explained earlier, experiential encounters alone can improperly calibrate the trust new drivers place in their autonomous vehicles.[25]

Case Study 1: Key Takeaways from User Level Case Study

- Wider gaps in misalignment between perceived and actual technology capabilities can lead to, or otherwise exacerbate, automation bias.
- Automation bias will be impacted by the user's level of prior knowledge and experience, which should be of particular concern in safety critical situations.

In the U.S., drivers are often considered the responsible party in car accidents, particularly when it comes to the role of the driver and the role of the system.[26]

As David Zipper, Senior Fellow at the MIT Mobility Initiative, explained:

> "In the United States, the responsibility for road safety largely falls on the individual sitting behind the wheel, or riding a bike, or crossing the street. American transportation departments, law enforcement agencies, and news outlets frequently maintain that most crashes—indeed, 94 percent of them, according to the most widely circulated statistic—are solely due to human error. Blaming the bad decisions of road users implies that nobody else could have prevented them."[27]

However, even the most experienced and knowledgeable human users are not free from the risk of overreliance in the face of poor interface and system design, and there is a peculiar dynamic at play with autonomous vehicles: When incidents occur, blame often falls on the software.[28] While the software may not be blameless, the combination of the system and inappropriate human use must also be considered in identifying the causes of harm. Therefore, ways of intervening or monitoring to prevent inappropriate use by drivers should be sought out alongside ways of improving the system's technical features and design.

Case Study 2: How Technical Design Factors Can Induce Automation Bias

A review of crashes in the aviation industry demonstrates that even in cases where users are highly trained, actively monitored, possess a thorough understanding of the technology's capabilities and limitations, and can be assured not to misuse or abuse the technology, a poorly designed interface can make automation bias more likely.

Fields dedicated to optimizing these links between the user and the system, such as human factors engineering and UI/UX design, are devoted to integrating and applying knowledge about human capabilities, limitations, and psychology into the design and development of technological systems.[29] Physical details, from the size and location of a button to the shape of a lever or selection menu to the color of a flashing light or image, seem small or insignificant. Yet these features can play a pivotal role in shaping human interactions with technology and ultimately determining a system's utility.

The importance of considering human interaction in the design and operation of these systems cannot be overstated—neglecting the human element in design can lead to inefficiencies at best, and unsafe and dangerous conditions at worst. Poorly designed interfaces, characterized by features as simple as drop-down menus with a lack of clear distinctions, were, for example, at the core of the accidental issuance of a widespread emergency alert in Hawaii that warned of an imminent, inbound ballistic missile attack.[30]

Design choices, intentionally or not, shape and establish specific behavioral pathways for how humans operate and rely on the systems themselves. In other words, these design choices can directly embed and/or exacerbate certain cognitive biases, including automation bias. These design choices are especially consequential when it comes to hazard alerts, such as visual, haptic, and auditory alarms. The commercial aviation industry illustrates how automation bias can be directly influenced by system designs:

The Human-Machine Interface: Airbus and Boeing Design Philosophies

Automation has been central to the evolution of the airplane since its inception—it took less than ten years from the first powered flight to the earliest iterations of autopilot.[31] In the years since, aircraft flight management systems, including those that are AI-enabled, have become successively capable. Today, a great deal of the routine work of flying a plane is handled by automated systems. This has not rendered pilots obsolete, however.[32] On the contrary, pilots must now incorporate the aircraft system's interpretation and reaction to external conditions before determining the most appropriate response, rather than directly engaging with their surroundings. While overall, flying has become safer due to automation, automation bias represents an ever-present risk factor.[33] As early as 2002, a joint FAA-industry study warned that the significant challenge for the industry would be to manufacture aircraft and design procedures that are less error-prone and more robust to errors involving incorrect human response after failure.[34]

While there are international standards as well as a general consensus among aircraft manufacturers that flight crews are ultimately responsible for safe aircraft operation, the two leading commercial aircraft providers in the United States, Airbus and Boeing, are known for their opposite design philosophies.[35] The differences between them illustrate different approaches to the automation bias challenge.

In Airbus aircraft, the automated system is designed to insulate and protect pilots and flight crews from human error. The pilot's control is bounded by "hard" limits, designed to allow for manipulation of the flight controls but prohibitive of any changes in altitude or speed, for example, that would lead to structural damage or loss of control of the aircraft (in other words, actions to exceed the manufacturer's defined flight envelope).

In contrast, in Boeing aircraft, the pilot is the absolute and final authority and can use natural actions with the systems to essentially “insist” upon a course of action. These “soft” limits exist to warn and alert the pilot but can be overridden and disregarded, even if it means the aircraft will exceed the manufacturer’s flight envelope.

These design differences may help explain why some airlines only operate single-type fleets; pilots typically stick to one type of aircraft, and cross-training pilots is possible but costly and, therefore, uncommon.[36]

Table 2 shows an FAA summary of the different design philosophies:

Table 2: Airbus and Boeing Design Philosophies

Airbus	Boeing
Automation must not reduce overall aircraft reliability; it should enhance aircraft and systems safety, efficiency, and economy. Automation must not lead the aircraft out of the safe flight envelope, and it should maintain the aircraft within the normal flight envelope. Automation should allow the user to use the safe flight envelope to its full extent, should this be necessary due to extraordinary circumstances. Within the normal flight envelope, the automation must not work against operator inputs, except when absolutely necessary for safety.	The pilot is the final authority for the operation of the airplane. Both crew members are ultimately responsible for the safe conduct of the flight. Flight crew tasks, in order of priority, are safety, passenger comfort, and efficiency. Design for crew operations is based on pilot’s past training and operational experience. Design systems are error tolerant. The hierarchy of design alternatives is simplicity, redundancy, and automation. Apply automation as a tool to aid, not replace, the pilot. Address fundamental human strengths, limitations, and individual differences—for both normal and nonnormal operations.

	Use new technologies and functional capabilities only when: 1) They result in clear and distinct operational or efficiency advantages, and 2) There is no adverse effect to the human-machine interface.

Source: Kathy Abbott, “Human Factors Engineering and Flight Deck Design,” in *The Avionics Handbook*, edited by Cary Spitzer, CRC Press LLC, 2001.

Despite the divergence in their design philosophies, both aircraft types maintain high levels of popularity and safety, with “virtually every large passenger plane that is flown in the Western world” being built by either Airbus or Boeing, proving the effectiveness of their respective approaches when consistently applied across design, training, and operations.[37] Neither is immune, however, to accidents or failures, especially when these philosophies are violated, or changes are not adequately communicated to users.

Boeing Incidents

On October 29, 2018, Lion Air Flight 610 crashed. Less than six months later, on March 10, 2019, Ethiopian Airlines Flight 302 crashed. Both incidents, plus a third incident involving another Boeing 737 Max 8 aircraft that narrowly avoided a crash, resulted in a combined 346 fatalities. While the exact nature of these accidents varied, all three were ultimately attributed to complications arising from Boeing’s introduction of new software—the Maneuvering Characteristics Augmentation System, or MCAS.

The MCAS system was engineered to assist in maintaining the 737 Max’s stability during flight and prevent conditions that could lead to a stall. While the system was designed to assist the pilot and could be overridden, the update was not well communicated to the pilots and thus may have violated one of Boeing’s principles to “design for crew operations based on pilot’s past training and operational experience.”[38]

While Boeing’s failure to communicate the change adequately is reminiscent of issues Tesla has faced communicating updates to drivers, the deviation from past design principles may have further served to undermine the pilot’s control.[39] Indeed, a review by the National Transportation Safety Board determined that “in all three flights, the pilot responses differed and did not match the assumptions of pilot responses to unintended MCAS operation on which Boeing based its hazard classifications.”[40]

Airbus Incidents

Perhaps an even more powerful case study concerning the consequences of technology design choices is the case of Air France Flight 447, which crashed in the Atlantic on June 1, 2009. Nearly three years later, the French Civil Aviation Safety Investigation Authority released its final report detailing how technical issues caused by ice on parts of the plane led to inconsistent speed measures and the shutting off of autopilot. This shutoff caused the crew to make choices that stalled the plane—an uncommon occurrence thanks to onboard automated systems—and eventually led to the crash. [41]

Post-accident reporting and subsequent analysis raised the question that even if one conceded the design flaw that led to the initial autopilot shutoff, "How could the pilots have a computer yelling "stall" at them and not realize they were in a stall?"[42] Ultimately, it was a confluence of human error and poor system design. The system design issue was with the flight management system, which presented a flurry of alerts and warnings to the pilot that "made it overwhelmingly difficult to recognize what was happening."[43]

In addition to the alerts, it was clear that automation itself played some role in the crash. In particular, there was a concern that approaches like Airbus', which emphasized protecting the pilot, actually went too far and were eroding pilot capabilities and skills by making them too dependent on the automated systems.[44] Ironically, as automation has made air travel much safer, it has also reduced the instances where a human pilot must take control of the plane in more complicated situations. This may in turn degrade the pilot's ability to properly control the plane when it is most needed.

Case Study 2: Key Takeaways from Technical Design Level Case Study

- Even with highly trained users, system design flaws can induce more harms. Neglecting human factors in system design can undermine users' ability to operate technology effectively and safely.
- Maintaining a "human-in-the-loop" is insufficient to preventing accidents or errors. There must be clear communication between the human user and the system, as well as sufficient training for the user, such that the "handover" from the system to the human does not become a weak point.
- Different design philosophies have different risks; they are not necessarily inherently better or worse. Any approach requires clear, consistent communication and application.

Both Boeing's and Airbus' past incidents underscore the complexity and risks associated with human-machine interaction. The interface design, physical layout, and functionality of controls directly influence user behavior and decision-making processes. In essence, design can induce user biases, including automation bias.

Both design approaches—whether prioritizing human control or protection—can be successful when communicated effectively, consistently, and purposefully. Human factors design choices should not be an afterthought. The rationale behind technical design choices should be aligned with organizational goals, priorities, and preferences. In these cases, users can better anticipate system behavior, respond promptly to changing circumstances, and more rapidly identify and explain any deviations from the norm, hopefully before accidents occur. That said, no system is 100% error-proof.

Case Study 3: How Organizations Can Institutionalize Automation Bias

While the Airbus incident with Air France 447 is a case study in human factors design choices, the after-action report also explained that "the behavior observed at the time of an event is often consistent with, or an extension of, a specific culture and work organization."[45] Organizational factors influencing automation bias include formal guidance documents, institutional processes, procurement guidelines, audits or inspections, incentive programs, and stated priorities, as well as informal norms or training expectations. These factors should be appreciated as both a source of risk and a hedge against errors by humans or technologies.[46] Organizational policies and processes for risk reduction are widely practiced in areas such as occupational safety and cybersecurity. The healthcare field has extensively studied the factors that make for “high reliability organizations,” a term that was first studied in the context of aircraft carrier operations.[47] These organizational controls take as a premise that if “we cannot change the human condition, we can change the conditions under which humans work.”[48]

Divergent Organizational Approaches to Automation: Army vs. Navy

The U.S. military provides an insightful case study of how an organization can shape automation bias. The military is able to exercise significant control over its users through organizational policies and nearly a century’s worth of experience deploying highly automated defensive systems to service members. Within the military, the Army and Navy deploy a very similar automated missile defense system with two very different approaches.

The Navy’s AEGIS weapons system and the Army’s Patriot system are tiered autonomy systems that scan for incoming air threats (missiles or aircraft), track them with highly capable radars, and guide missiles (for AEGIS the “Standard Missile” or SM; for Patriot, the Patriot Advanced Capability or PAC) to strike an incoming threat.[49] The systems are capable of supervised autonomous operations up to and including launching defensive missiles without human input, in comparable ways (see Table 3). They have been widely viewed as successful defensive systems since the late 1980s, though there have been notable disasters associated with both.[50]

Table 3: Comparison of AEGIS and Patriot Weapons Systems Autonomous Functions

AEGIS[51]	**Patriot**[52]
Manual Identification The **user must evaluate a detected radar track and assign an identity** (e.g., friend, unknown, hostile) based on the track's location, speed, the Identification Friend or Foe System (IFF), and electronic emissions.	Manual Identification The **user must evaluate a detected radar track and assign an identity** (e.g., friend, unknown, hostile) based on the track's location, speed, IFF, and electronic emissions.
IFF, Identification, Drop-Track Doctrine These three separate doctrines can be individually or collectively activated to perform track identification tasks. IFF doctrine **automatically performs an IFF query** within a certain geographic area. Identification doctrine **automatically identifies a detected track and assigns an identity** (e.g., friend, unknown, or hostile) based on location, speed, IFF and course. Drop-Track will **automatically remove tracks** from a user's display if they meet predefined criteria for being incorrect tracks (e.g., weather-related clutter).	Automatic Identification Mode The **system will automatically identify a detected track and assign an identity** (e.g., friend, unknown, hostile) based on the track's location, speed, IFF, and electronic emissions.
Auto SM Doctrine The **system automatically identifies** threatening targets and notifies **users to manually engage**.	Semiautomatic Engagement Mode The **system automatically identifies** and prioritizes threatening targets for **users to manually engage**.
Auto-Special Doctrine The **system will automatically engage** and fire against threats that meet set parameters without human user action required. A human user can halt the engagement.	Automatic Engagement Mode The **system will automatically engage** and fire against threats that meet set parameters without human user action required. A human user can halt the engagement.

While the two systems function very similarly, the U.S. Army and Navy employ different approaches in how they are governed and provide a useful study of how organizations can shape user interactions.

Patriot: A Bias Towards the System

Both the Army and Navy employ detailed, specific instructions and processes to govern deployed weapons. Among these are rules of engagement (ROE), weapon control status orders, self-defense engagement criteria, and airspace control orders which are among the controls "developed specifically for the theater, and put into operation quickly to reduce the possibility of fratricide . . ."[53]

Despite these controls and consistent success in Operation Desert Storm, in 2003 the Patriot system was involved in three separate friendly fire incidents during Operation Iraqi Freedom: one that mistook a Patriot battery for an enemy surface-to-air missile system, and two that misclassified coalition aircraft. The latter incidents resulted in three fatalities.[54]

In 2005, the Defense Science Board conducted a review of the overall performance of the Patriot system in Operation Iraqi Freedom and found that these incidents followed the "Swiss cheese" model of safety incidents, a result of a series of failures—"some human and some machine"— that all contributed to the unfortunate outcomes. Among their conclusions as to the source of the fratricides, they included fault with the Patriot system operating philosophy, protocols, displays, and software, which they found inappropriately tailored for the mission.[55]

On this point, the report elaborated that the Army preferred to use the system in the "automatic" mode so it could operate faster.[56] Official Army guidance from 2002 does instruct users that the "default" mode for Patriot is to fight in the "automatic engagement mode" as opposed to manual or automatic identification mode (see Table 3). In the case of theater ballistic missile (TBM), for example, the instruction states:

> *"When the system has classified a target as a TBM, engagement decisions and the time in which the user has to make those decisions are very limited."*[57]

In addition to this documented guidance, Air Defense Artillery training was criticized as a factor contributing to automation bias and cognitive off-loading by users because it emphasized "rote drills versus the exercise of high-level judgment."[58]

Between doctrine (guidance to operate in an automatic mode), training (which was "rote"), and the success of Patriot 10 years earlier in Operation Desert Storm, Patriot operators became biased toward "reacting quickly, engaging early, and trusting the system without question."[59] The bias was such that in some of the incidents, Patriot was operating only in semi-automatic engagement mode and a human user confirmed an engagement on an incorrectly identified track. As one researcher later put it, "Patriot operators, while nominally in control, exhibited automation bias: an unwarranted and uncritical trust in automation. In essence, control responsibility is ceded to the machine."[60] It was put more bluntly by a later researcher: "A semi-automatic system in the hands of an inadequately trained staff is de facto a fully automated system."[61]

AEGIS: A Bias Towards the Human

Balancing the dynamic roles between human and machine is complicated. Moreover, as the AEGIS system demonstrates, weighting decision-making towards the human will not eliminate all risks from autonomous systems.

The AEGIS weapons systems are central pillars of air defense for the U.S. Navy. Despite its capabilities and centrality to naval defense, Navy training and doctrine show a preference towards decisions by users rather than determinations by autonomous systems. These biases are visible in the staffing, doctrine, and training for AEGIS. An AEGIS air engagement, for example, will involve several qualified sailors (officers and enlisted) in the task of identifying a radar track, tasks they can manually perform even when the system is in an autonomous mode. Furthermore, Navy training documentation makes clear that elements of AEGIS are prone to failure, saying, for example in a 1991 training manual:

> *"It is quite possible that the IFF equipment may be functioning improperly. The only reasonable recourse in the event of no IFF return is to get as many [air] units as possible on the contact . . . If time is short, and we cannot receive the correct IFF response, we must assume that the contact is an enemy."*[62]

Unclassified documents also make clear that the onus of responsibility is placed on the AEGIS watch-standers; for example, the same training document concludes an overview of the AEGIS system with the following paragraph:

> *"Your training in combat systems is a never-ending process which you must approach with an aggressive and unremitting attitude until actions become almost second nature. Your duties are many and complex; to be effective requires your total commitment."*[63]

Despite the sophistication of the AEGIS system and the emphasis on human control, in 1988 during the Iran-Iraq war, the *USS Vincennes*, one of the first ships to employ AEGIS, inadvertently shot down civilian aircraft Iran Air Flight 655, having mistaken it for an Iranian fighter aircraft. The incident occurred within the context of extreme stress: The ship was concurrently engaging Iranian ships, and intelligence and warnings suggested an assault that particular weekend. Furthermore, the *USS Stark* had been struck by an Iraqi jet a year earlier.[64] The shootdown resulted in the deaths of nearly 300 people and further tension between the United States and Iran that has continued to today.[65]

Analysis of the *USS Vincennes* incident found that AEGIS worked correctly. It identified the aircraft in question as a civilian aircraft ascending from launch. However, the *Vincennes* crew did not seem to recognize the information, instead reporting that the aircraft was descending and a military aircraft, thus justifying defensive weapons.[66]

The *USS Vincennes* incident shows that even when humans are taught and trained to be skeptical of a system, users can fail to correctly interpret the system's output or appropriately trust the technology, particularly under situations of extreme stress.[67]

Case Study 3: Key Takeaways from Organizational Level Case Study

- Organizations can employ the same tools and technologies in very different ways, based on protocols, operations, doctrine, training, and certification. Choices in each of these areas of governance can embed automation biases.
- Organizational efforts to mitigate automation bias can be successful, but mishaps are still possible, especially when human users are under stress.

The AEGIS and Patriot weapons systems show how organizational policies play a significant role in shaping automation bias. In the case of AEGIS, the Navy organized itself to preference human decisions. In the case of Patriot, the Army made decisions that preference automated system decisions.

The 2003 Patriot fratricides and the 1988 *USS Vincennes* incident further highlight that regardless of approach, there are risks of mishap. The Army has successfully employed Patriot for decades and it is a coveted defensive weapon, despite tragic past errors. The Navy has also successfully employed AEGIS under different rules and assumptions, but it has also experienced at least one lethal failure when sailors were under extreme stress. Therefore, organizational decisions that shape automation bias are not fail-safe against risk, and they must be carefully considered in light of technology capabilities, user understanding, and context of deployment.

Conclusion

Unaddressed automation bias has already culminated in catastrophic accidents. From these case studies of past mishaps, we identify three important factors affecting a user's automation bias: those intrinsic to the human user, such as personal biases, experience, and confidence in using the system; those inherent to the AI system, such as how it can be operated or how it presents information; and those created by organizational factors such as standard processes and procedures.

Addressing these factors affecting risk in the application of AI, particularly in safety-critical contexts, requires focused attention during the design and deployment of AI systems. With the lessons learned from the three case studies, we recommend as a starting point the following mitigations:

- **Create and maintain qualification standards for user understanding.** In each of our case studies, we learned that misunderstandings by users often contributed to the incident, either generally or due to a specific recent system change or upgrade. Since user understanding is critical to safe operation, **system developers and vendors must invest in clear communications about their systems** and organizations and governments may need to **create qualification or re-qualification regimes** appropriate to the technology and its use.

- **Value and enforce consistent design and design philosophies, especially for systems likely to be upgraded.** When necessary, justify departures from a design philosophy and make choices well-known to legacy users. **Where possible, develop common design criteria, standards, and expectations, and consistently communicate them (either through organizational policy or industry standards**) to reduce the risk of confusion and automation bias.

- **Where autonomous systems are used by organizations, design and regularly review organizational policies to be consistent with technical capabilities and organizational priorities**. Update policies and processes as technologies change to best account for new capabilities and mitigate novel risks. Look for opportunities to implement principles of high-reliability organizations around the management of frontline AI deployment.

The risk of accidents or misuse of AI-enabled systems will evolve alongside technology, the design of human-machine interactions, and user understanding. The successful, safe, and ethical deployment of AI relies not only on its capacity to work seamlessly with human users but also on the competence and accountability of the humans overseeing, monitoring, managing, using, and ultimately relying on these systems. If humans “in-the-loop” are to be effective, they must learn when and how to cognitively offload tasks to AI systems.

Authors

Lauren Kahn completed her contributions to this research while she was a senior research analyst at CSET, and she is currently on assignment to the Office of the Deputy Assistant Secretary of Defense for Force Development and Emerging Capabilities under an Intergovernmental Personnel Act agreement with CSET.

Emelia S. Probasco is a senior fellow at CSET.

Ronnie Kinoshita is the deputy director of data science and research at CSET.

The views expressed herein are the authors' and do not necessarily reflect those of the U.S. government.

Acknowledgments

The authors would like to thank Cat Aiken, John Bansemer, Quinn Baumgartner, Nathan Bos, Mitchell Brown, Daniella Castillo, James Dunham, Shelton Fitch, Bob Lennox, Matt Mahoney, Igor Mikolic-Torreira, Chris Murphy, Mina Narayanan, Michael Partida, Ioana Puscas, Helen Toner, and Vikram Venkatram for feedback and assistance.

Document Identifier: doi: 10.51593/20230057

Endnotes

[1] Kate Goddard, Abdul Roudsari, Jeremy C Wyatt, "Automation bias: a systematic review of frequency, effect mediators, and mitigators," *Journal of the American Medical Informatics Association*, Volume 19, Issue 1 (2012): 121–27, https://doi.org/10.1136/amiajnl-2011-000089.

[2] Mary Cummings, "Automation Bias in Intelligence Time Critical Decision Support Systems," AIAA 1st Intelligent Systems Technical Conference (AIAA), Chicago, IL (2012), https://doi.org/10.2514/6.2004-6313.

[3] N.A., Incidents, AI Incident Database, accessed April 2, 2024, https://incidentdatabase.ai/apps/incidents/.

[4] Grace Augustine, Jan Lodge, Mislav Radic, "Mr. Bates vs The Post Office depicts one of the UK's worst miscarriages of justice: here's why so many victims didn't speak out," *The Conversation*, January 4, 2024, https://theconversation.com/mr-bates-vs-the-post-office-depicts-one-of-the-uks-worst-miscarriages-of-justice-heres-why-so-many-victims-didnt-speak-out-220513.

[5] N.A., "Victims of UK Post Office IT scandal faced four main barriers to speaking out – new research," University of Bath, January 8, 2024, https://www.bath.ac.uk/announcements/victims-of-uk-post-office-it-scandal-faced-four-main-barriers-to-speaking-out-new-research/.

[6] S. W. A. Dekker, D. D. Woods, "MABA-MABA or Abracadabra? Progress on Human-Automation Co-ordination," *Cognition, Technology & Work*, Volume 4 (2002): 240–244, https://doi.org/10.1007/s101110200022.

[7] Sara E. McBride, Wendy A. Rogers, Arthur D. Fisk, "Understanding human management of automation errors," *Theoretical Issues in Ergonomics Science*, Volume 15, Issue 6 (2013): 545–577, https://doi.org/10.1080/1463922X.2013.817625, and Kate Goddard, Abdul Roudsari, Jeremy C. Wyatt, "Automation bias: a systematic review of frequency, effect mediators, and mitigators," *Journal of the American Medical Informatics Association*, Volume 19, Issue 1 (2012): 121–127, https://doi.org/10.1136/amiajnl-2011-000089.

[8] Kate Goddard, Abdul Roudsari, Jeremy C. Wyatt, "Automation bias: Empirical results assessing influencing factors," *International Journal of Medical Informatics*, Volume 83, Issue 5 (2014): 368–375, https://doi.org/10.1016/j.ijmedinf.2014.01.001.

[9] David M. Sanbonmatsu, David L. Strayer, Zhenghui Yu, Francesco Biondi, Joel M. Cooper, "Cognitive underpinnings of beliefs and confidence in beliefs about fully automated vehicles," *Transportation Research Part F: Traffic Psychology and Behaviour*, Volume 55 (2018): 114–122, https://psycnet.apa.org/doi/10.1016/j.trf.2018.02.029; Michael C. Horowitz, Lauren Kahn, Julia Macdonald, Jacquelyn Schneider, "Adopting AI: how familiarity breeds both trust and contempt," *AI & Society*, Volume 39 (2023): 1721–1735, https://doi.org/10.1007/s00146-023-01666-5; Robert E.

Burnkrant, Alain Cousineau, "Informational and Normative Social Influence in Buyer Behavior," *Journal of Consumer Research*, Volume 2, Issue 3 (1975): 206–215, https://psycnet.apa.org/doi/10.1086/208633.

[10] Michael C. Horowitz, Lauren Kahn, "Bending the Automation Bias Curve: A Study of Human and AI-Based Decision Making in National Security Contexts," *International Studies Quarterly*, Volume 68, Issue 2 (2024), https://doi.org/10.1093/isq/sqae020; Justin Kruger, David Dunning, "Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments," *Journal of Personality and Social Psychology*, Volume 77, no. 6 (1999): 1121–1134, https://psycnet.apa.org/doi/10.1037/0022-3514.77.6.1121; Carmen Sanchez, David Dunning, "Overconfidence among beginners: Is a little learning a dangerous thing?" *Journal of Personality and Social Psychology*, Volume 114, no. 1 (2018): 10–28, https://doi.org/10.1037/pspa0000102.

[11] National Highway Traffic Safety Administration, "Automated Vehicles for Safety," U.S. Department of Transportation, accessed April 2, 2024, https://www.nhtsa.gov/vehicle-safety/automated-vehicles-safety#:~:text=In%20some%20circumstances%2C%20automated%20technologies,%2C%20injuries%2C%20and%20economic%20tolls.

[12] N.A., "Despite warnings, many people treat partially automated vehicles as self-driving," Insurance Institute for Highway Safety (IIHS)/Highway Loss Data Institute (HLDI), October 11, 2022, https://www.iihs.org/news/detail/despite-warnings-many-people-treat-partially-automated-vehicles-as-self-driving.

[13] Moritz Körber, Eva Baseler, Klaus Bengler, "Introduction matters: Manipulating trust in automation and reliance in automated driving," *Applied Ergonomics*, Volume 66 (2018): 18–31. https://doi.org/10.1016/j.apergo.2017.07.006.

[14] Chris Schwarz, John Gaspar, Timothy Brown, "The effect of reliability on drivers' trust and behavior in conditional automation," *Cognition, Technology & Work*, Volume 21 (2019): 41–54. https://link.springer.com/article/10.1007/s10111-018-0522-y.

[15] Apoorva P. Hungund, Ganesh Pai, Anuj K. Pradhan, "Systematic Review of Research on Driver Distraction in the Context of Advanced Driver Assistance Systems," *Transportation Research Record*, Volume 2675, Issue 9 (2021): 756–765, https://doi.org/10.1177/03611981211004129; Moritz Körber, Eva Baseler, Klaus Bengler, "Introduction matters: Manipulating trust in automation and reliance in automated driving," *Applied Ergonomics*, Volume 66 (2018): 18–31. https://doi.org/10.1016/j.apergo.2017.07.006.

[16] Danny Yadron, Dan Tynan, "Tesla driver dies in first fatal crash while using autopilot mode," *The Guardian*, June 30, 2016, https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk; Dan Levine and Hyunjoo Jin, "Next Autopilot trial to test Tesla's blame-the-driver defense," *Reuters*, March 11, 2024, https://www.reuters.com/business/autos-transportation/next-autopilot-trial-test-teslas-blame-the-driver-defense-2024-03-11/.

[17] Russ Mitchell, "DMV probing whether Tesla violates state regulations with self-driving claims," *Los Angeles Times*, May 17, 2021, https://www.latimes.com/business/story/2021-05-17/dmv-tesla-

california-fsd-autopilot-safety; "Tesla's Autopilot misleading because humans still in control: Pete Buttigieg," *New York Post*, May 11, 2023, https://nypost.com/2023/05/11/tesla-shouldnt-call-driving-system-autopilot-pete-buttigieg-says/; Dan Mihalascu, "Tesla FSD Might Reach Level 4 Or Level 5 Autonomy This Year: Musk," *Inside EVs*, July 7, 2023, https://insideevs.com/news/675701/tesla-fsd-might-reach-level-4-level-5-autonomy-this-year-musk/.

[18] N.A., "Incident 52: Tesla on Autopilot Killed Driver in Crash in Florida While Watching Movie," AI Incident Database, accessed April 2, 2024, https://incidentdatabase.ai/cite/52/.

[19] N.A., "NHTSA ACTION NUMBER: PE21020 Autopilot & First Responder Scenes," August 13, 2021; N.A., "NHTSA ACTION NUMBER: EA22002 Autopilot System Driver Controls," June 8, 2022.

[20] N.A. "NHTSA ACTION NUMBER: RQ24009 OPEN INVESTIGATION, Recall 23V838 Remedy Effectiveness," April 25, 2024.

[21] N.A., "Additional Information Regarding EA22002," National Highway Traffic Safety Administration, April 25, 2024, https://static.nhtsa.gov/odi/inv/2022/INCR-EA22002-14496.pdf.

[22] David Shepardson, "US probes Tesla recall of 2 million vehicles over Autopilot," *Reuters*, April 26, 2024, https://www.reuters.com/business/autos-transportation/us-probes-tesla-recall-2-million-vehicles-over-autopilot-citing-concerns-2024-04-26/.

[23] N.A., "Update Vehicle Firmware to Prevent Driver Misuse of Autosteer," Tesla, N.D., https://www.tesla.com/support/vehicle-firmware-prevent-autosteer-misuse.

[24] N.A., "Update Vehicle Firmware to Prevent Driver Misuse of Autosteer," Tesla, N.D., https://www.tesla.com/support/vehicle-firmware-prevent-autosteer-misuse.

[25] N.A., "Federal Regulators Investigating Tesla's Autopilot Recall Fix," *Consumer Reports*, April 26, 2024, https://www.consumerreports.org/cars/car-safety/tesla-autopilot-recall-fix-does-not-address-safety-problems-a5133751100/; N.A. "NHTSA ACTION NUMBER: RQ24009 OPEN INVESTIGATION, Recall 23V838 Remedy Effectiveness," April 25, 2024; Moritz Körber, Eva Baseler, Klaus Bengler "Introduction matters: Manipulating trust in automation and reliance in automated driving," *Applied Ergonomics*, Volume 66 (2018): 18–31. https://doi.org/10.1016/j.apergo.2017.07.006.

[26] Qiyuan Zhang, Christopher D. Wallbridge, Dylan M. Jones, Phillip L. Morgan, "Public perception of autonomous vehicle capability determines judgment of blame and trust in road traffic accidents," *Transportation Research Part A: Policy and Practice*, Volume 179 (2024), https://doi.org/10.1016/j.tra.2023.103887.

[27] David Zipper, "The Deadly Myth That Human Error Causes Most Car Crashes," *The Atlantic*, November 26, 2021, https://www.theatlantic.com/ideas/archive/2021/11/deadly-myth-human-error-causes-most-car-crashes/620808/.

[28] Kathleen L. Mosier, Linda J. Skitka, Susan Heers, Mark Burdick, "Automation Bias: Decision Making and Performance in High-Tech Cockpits," *The International Journal of Aviation Psychology*, Volume 8, Issue 1 (1998): 47–63, https://doi.org/10.1207/s15327108ijap0801_3.

[29] John Dowell, John Long, "Towards a conception for an engineering discipline of human factors," *Ergonomics*, Volume 32, no. 11 (1989): 1513–1535, https://doi.org/10.1080/00140138908966921.

[30] Amy B. Wang, "Hawaii missile alert: How one employee 'pushed the wrong button' and caused a wave of panic," *The Washington Post*, January 14, 2018, https://www.washingtonpost.com/news/post-nation/wp/2018/01/14/hawaii-missile-alert-how-one-employee-pushed-the-wrong-button-and-caused-a-wave-of-panic/; Alex Hern, "Hawaii missile false alarm due to badly designed user interface, reports say," *The Guardian*, January 15, 2018, https://www.theguardian.com/technology/2018/jan/15/hawaii-missile-false-alarm-design-user-interface; Kim Flaherty, "What the Erroneous Hawaiian Missile Alert Can Teach Us About Error Prevention," Nielsen Norman Group, January 16, 2018, https://www.nngroup.com/articles/error-prevention/.

[31] Tara Leggett, "The Evolution of Autopilot," *Key.Aero*, August 21, 2020, https://www.key.aero/article/evolution-autopilot.

[32] Lawrence J. Prinzel III, *Team-Centered Perspective for Adapted Automation Design*, (Hampton, VA: NASA Langley Research Center, 2003).

[33] Charles E. Billings, *Aviation Automation: The Search For A Human-Centered Approach*, (Mahwah, NJ: Lawrence Erlbaum Associates, Publishers, 1997); "Statistical Summary of Commercial Jet Airplane Accidents: Worldwide Operations 1959–2022," Boeing, September 2023, https://www.faa.gov/sites/faa.gov/files/2023-10/statsum_summary_2022.pdf.

[34] N.A., *Commercial Airplane Certification Process Study: An Evaluation of Selected Aircraft Certification, Operations, and Maintenance Processes*, (Washington, DC: U.S. Department of Transportation Federal Aviation Administration, 2002).

[35] Title 14 Code of Federal Regulations, Part 25, https://www.ecfr.gov/current/title-14/chapter-I/subchapter-C/part-25?toc=1; "Regulations," European Union Aviation Safety Agency, accessed April 4, 2024, https://www.easa.europa.eu/en/regulations; Alexander Z. Ibsen, "The politics of airplane production: The emergence of two technological frames in the competition between Boeing and Airbus," *Technology in Society*, Volume 31, Issue 4 (2009): 342–349, https://doi.org/10.1016/j.techsoc.2009.10.006.

[36] Joe Kunzler, "Alaska Airlines To Become All-Boeing Carrier By October," *Simple Flying*, May 7, 2023, https://simpleflying.com/alaska-airlines-all-boeing-carrier-october/; Jack Herstam, "How Do Pilots Retain Their Type Ratings?," *Simple Flying*, June 4, 2023, https://simpleflying.com/how-pilots-retain-type-ratings/#:~:text=Though%20airlines%20only%20allow%20pilots,to%20have%20many%20type%20ratings; N.A., "Etihad becomes one of the first to enable pilots to fly both A350 and A380 aircraft," Etihad Airways, February 14, 2024, https://www.etihad.com/en-za/news/etihad-becomes-one-of-the-first-to-enable-pilots-to-fly-both-a350-and-a380-aircraft.

[37] Sylvia Pfeifer, Philip Georgiadis, Steff Chávez, "How Boeing's troubles are upsetting the balance of power in aviation," *Financial Times*, January 28, 2024, https://www.ft.com/content/ddc28f31-e1af-4a81-8295-cbccf3141f49.

[38] Kathy Abbott, "Human Factors Engineering and Flight Deck Design," in *The Avionics Handbook*, edited by Cary Spitzer, CRC Press LLC, 2001.

[39] N.A., *Safety Recommendation Report: Assumptions Used in the Safety Assessment Process and the Effects of Multiple Alerts and Indications on Pilot Performance*, (Washington, DC: National Transportation Safety Board, 2019); Sinéad Baker, "Boeing shunned automation for decades. When the aviation giant finally embraced it, ad automated system in the 737 Max kicked off the biggest crisis in its history," *Business Insider*, April 4, 2020, https://www.businessinsider.com/737-max-crashes-boeing-usually-downplays-automation-mcas-biggest-crisis-2020-3; Hearing Before the Subcommittee on Aviation of the Committee on Transportation and Infrastructure, House of Representatives, 116th Congress, "Status of the Boeing 737 Max: Stakeholder Perspectives," June 19, 2019. https://www.congress.gov/event/116th-congress/house-event/LC64168/text.

[40] N.A., *Safety Recommendation Report: Assumptions Used in the Safety Assessment Process and the Effects of Multiple Alerts and Indications on Pilot Performance*, (Washington, DC: National Transportation Safety Board, 2019).

[41] N.A., "Final Report on the accident on 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France flight AF 447 Rio de Janeiro - Paris," (Le Bourget Cedex, France: Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile, 2012).

[42] Roman Mars, "Air France Flight 447 and the Safety Paradox of Automated Cockpits," *Slate*, June 25, 2015, https://www.slate.com/blogs/the_eye/2015/06/25/air_france_flight_447_and_the_safety_paradox_of_airline_automation_on_99.html; Gregory Polek, "Court Acquits Airbus, Air France in AF447 Manslaughter Trial," *Aviation International News*, April 17, 2023, https://www.ainonline.com/aviation-news/air-transport/2023-04-17/court-acquits-airbus-air-france-af447-manslaughter-trial; William Langewiesche, "The Human Factor," *Vanity Fair*, September 17, 2014, https://www.vanityfair.com/news/business/2014/10/air-france-flight-447-crash.

[43] Paulus A.J.M. de Wit, Roberto Moraes Cruz, "Learning from AF447: Human-machine interaction," *Safety Science*, Volume 112, (2019): 48–56, https://doi.org/10.1016/j.ssci.2018.10.009.

[44] Nick Oliver, Thomas Calvard, Kristina Potočnik, "The Tragic Crash of Flight AF447 Shows the Unlikely but Catastrophic Consequences of Automation," *Harvard Business Review*, September 15, 2017, https://hbr.org/2017/09/the-tragic-crash-of-flight-af447-shows-the-unlikely-but-catastrophic-consequences-of-automation.

[45] N.A. "Final Report on the accident on 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France flight AF 447 Rio de Janeiro - Paris," (Le Bourget Cedex, France: Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile, 2012).

[46] Paola Amaldi, Anthony Smoker, "An Organizational Study into the Concept of Automation in Safety Critical Socio-technical System," *IFIP Advances in Information and Communication Technology*, (Berling: Springer, 2013): 183–197, https://doi.org/10.1007/978-3-642-41145-8_16.

[47] See for example, Kathleen M. Sutcliffe, "High Reliability Organizations (HROs)," *Best Practice & Research Clinical Anaesthesiology, Safety in Anaesthesia*, Volume 25, no. 2 (June 1, 2011): 133–44; and K. H. Roberts, "New challenges in organizational research: high reliability organizations," *Industrial Crisis Quarterly*, 1989.

[48] James Reason, "Human error: models and management," *The BMJ* (Clinical research ed.), Volume 320 (2000): 768–770, https://doi.org/10.1136/bmj.320.7237.768.

[49] N.A., "AEGIS Weapon System," U.S. Navy, 20 September 2021, https://www.navy.mil/Resources/Fact-Files/Display-FactFiles/Article/2166739/aegis-weapon-system/; Center for Strategic and International Security (CSIS) Missile Defense Project, "Patriot," August 23, 2023, https://missilethreat.csis.org/system/patriot/.

[50] Missile Defense Project, "Aegis Ballistic Missile Defense," Center for Strategic and International Studies, June 14, 2018, last modified August 4, 2021, https://missilethreat.csis.org/system/aegis/; Missile Defense Project, "Patriot," Center for Strategic and International Studies, June 14, 2018, last modified August 23, 2023, https://missilethreat.csis.org/system/patriot/.

[51] Sharif H. Calfee, "Autonomous Agent-Based Simulation of an AEGIS Cruiser Combat Information Center Performing Battle Group Air-Defense Commander Operations," Naval Postgraduate School, March, 2003, https://faculty.nps.edu/ncrowe/oldstudents/calfee-thesis.htm; R. Stephen Howard, "Combat systems and weapons department management," Naval Education and Training Program Management Support Activity, Pensacola, Fl. September 1991.

[52] Headquarters, Department of the Army, "FM 44-15-1: Operations and Training Patriot," Washington, DC, February 1987, https://www.bits.de/NRANEU/others/amd-us-archive/FM44-15-1Pt1%2887%29.pdf.

[53] N.A., "FM 44-85 Patriot Battalion and Battery Operations," Washington, DC: Headquarters Department of the Army, 21 February 1997, https://www.ausairpower.net/PDF-A/FM-44-85-Patriot-Battalion-and-Battery-Operations.pdf.

[54] N.A., *Report of the Defense Science Board Task Force on Patriot System Performance: Report Summary*, (Washington, DC: Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics, 2005).

[55] N.A., *Report of the Defense Science Board Task Force on Patriot System Performance: Report Summary*, (Washington, DC: Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics, 2005).

[56] N.A. *Report of the Defense Science Board Task Force on Patriot System Performance: Report Summary*, (Washington, DC: Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics, 2005).

[57] N.A., "FM 3-01.85, Patriot Battalion and Battery Operations," Washington, DC: Headquarters Department of the Army, 13 May 2002.

[58] John K. Hawley, Anna L. Mares, *Developing Effective Human Supervisory Control for Air and Missile Defense Systems* (Adelphi, MD: Army Research Laboratory, 2006).

[59] John K. Hawley, "Patriot Wars: Automation and the Patriot Air and Missile Defense System," Center for a New American Security, January 25, 2017, https://www.cnas.org/publications/reports/patriot-wars; N.A., *Military Aircraft Accident Summary: Aircraft Accident to Royal Air Force Tornado GR MK4A ZG710* (London, UK: Ministry of Defence Directorate of Air Staff, 2004).

[60] Mary Cummings, "Automation Bias in Intelligence Time Critical Decision Support Systems," AIAA 1st Intelligent Systems Technical Conference (AIAA), Chicago, IL (2012), https://doi.org/10.2514/6.2004-6313.

[61] John K. Hawley, "Patriot Wars: Automation and the Patriot Air and Missile Defense System," Center for a New American Security, January 25, 2017, https://www.cnas.org/publications/reports/patriot-wars.

[62] R. Stephen Howard, *Combat systems and weapons department management* (Pensacola, FL: Naval Education and Training Program Management Support Activity, September 1991).

[63] R. Stephen Howard, *Combat systems and weapons department management* (Pensacola, FL: Naval Education and Training Program Management Support Activity, September 1991).

[64] N.A., *Formal Investigation into the Circumstances Surrounding the Downing of Iran Air Flight 655 on 3 July 1988*, (Washington, DC: The Department of Defense, 1988).

[65] Jon Gambrell, "30 years later, US downing of Iran flight haunts relations," *The Associated Press*, July 3, 2018, https://apnews.com/article/de425ed9833b4bdaa03306f37182995f.

[66] Anthony Tingle, "Human-Machine Team Failed Vincennes," *Proceedings*, Volume 144, no. 7 (2018), https://www.usni.org/magazines/proceedings/2018/july/human-machine-team-failed-vincennes.

[67] Study of the *USS Vincennes* incident under the Navy's Tactical Decision Making Under Stress program point to other sources of error besides automation bias, such as the design of the AEGIS user interface and its relationship to other human cognitive biases like framing errors and confirmation bias. For more see, Jeffrey G. Morrison, Richard T. Kelly, Ronald A. Moore, Susan G. Hutchins, "Tactical decision making under stress (TADMUS) decision support system," Calhoun: The NPS Institutional Archive (1996), https://hdl.handle.net/10945/41225.