Risk Analysis, Vol. 42, No. 9, 2022 DOI: 10.1111/risa.13850

Learning from the Failure of Autonomous and Intelligent Systems: Accidents, Safety, and Sociotechnical Sources of Risk

Carl Macrae *

Efforts to develop autonomous and intelligent systems (AIS) have exploded across a range of settings in recent years, from self-driving cars to medical diagnostic chatbots. These have the potential to bring enormous benefits to society but also have the potential to introduce new—or amplify existing—risks. As these emerging technologies become more widespread, one of the most critical risk management challenges is to ensure that failures of AIS can be rigorously analyzed and understood so that the safety of these systems can be effectively governed and improved. AIS are necessarily developed and deployed within complex human, social, and organizational systems, but to date there has been little systematic examination of the sociotechnical sources of risk and failure in AIS. Accordingly, this article develops a conceptual framework that characterizes key sociotechnical sources of risk in AIS by reanalyzing one of the most publicly reported failures to date: the 2018 fatal crash of Uber's self-driving car. Publicly available investigative reports were systematically analyzed using constant comparative analysis to identify key sources and patterns of sociotechnical risk. Five fundamental domains of sociotechnical risk were conceptualized—structural, organizational, technological, epistemic, and cultural—each indicated by particular patterns of sociotechnical failure. The resulting SOTEC framework of sociotechnical risk in AIS extends existing theories of risk in complex systems and highlights important practical and theoretical implications for managing risk and developing infrastructures of learning in AIS.

KEY WORDS: Artificial intelligence; autonomous systems; organizational risk; system accident

1. INTRODUCTION

"I became tired of the ever-repeated robot plot. I didn't see robots that way. I saw them as machines—advanced machines—but machines. They might be dangerous but surely safety factors would be built in. The safety factors might be faulty or inadequate or might fail under unexpected types of stresses, but such failures could always yield experience that could be used to improve the models. After all, all devices have their dangers."

—Isaac Asimov, 1990

"I think people should be really concerned about it… I keep sounding the alarm bell but, you know, until people see robots going down the street, killing people, they don't know how to react because it seems so ethereal. I think we should be really concerned about AI."

—Elon Musk, 2017

People have started to see robots going down the street, killing people. In March 2018, a computer-driven autonomous vehicle being tested by Uber on

Centre for Health Innovation, Leadership and Learning, Nottingham University Business School, University of Nottingham, Nottingham, UK.
*Address correspondence to Carl Macrae, Centre for Health Innovation, Leadership and Learning, Nottingham University Business School, University of Nottingham, Jubilee Campus, Wollaton Road, Nottingham, NG8 1BB, UK; carlmacrae@mac.com

the streets of Tempe, Arizona, hit and killed a pedestrian who was crossing the road pushing a bicycle (NTSB, 2019a). A few days later a Tesla Model X driving under the guidance of its "Autopilot" automated driver assistance system drifted out of its lane near a junction and hit a barrier, killing the driver (NTSB, 2020a). These events followed a fatal collision when a Tesla Model S was driven under the side of a truck by its Autopilot system in 2016 (NTSB, 2017)—a sequence repeated by a Tesla Model 3 in 2019 (NTSB, 2020b). These accidents all involved computer-guided cars employing artificial intelligence to perform tasks that would previously have been the sole responsibility of a human driver. As such, they represent some of the most visible and lethal failures that have emerged from recent efforts to develop autonomous and intelligent systems.

Autonomous and intelligent systems (AIS) encompass a wide range of technologies that are able to instantiate various aspects of intelligent and self-directed behavior (IEEE, 2019). This can involve perceiving, predicting, planning, and performing in some particular environment—such as driving on public roads—or in relation to some particular problem, such as determining whether a skin lesion is cancerous or not (Tschandl et al., 2020). The field is currently undergoing something of an explosion. Progress has dramatically accelerated in recent years in part due to advances in key areas of artificial intelligence—particularly deep learning artificial neural networks (Russell, 2019)—together with the ready availability of enormous quantities of data and huge increases in computational power (Stoica et al., 2017). High profile public demonstrations have shown intelligent machines performing at superhuman levels in complex games such as Go (Silver, Schrittwieser, & Simonyan, 2017) and Quake III (Jaderberg, Czarnecki, & Dunning, 2019), and enormous investments have been made in attempts to develop and deploy large fleets of autonomous vehicles (Davies, 2021). Meanwhile, efforts are underway to incorporate AIS into many of the critical systems that modern society depends upon, from medical diagnosis (Fauw et al., 2018), to public transport (Mouratidis & Serrano, 2021), to financial trading (FCA, 2019), to care of the elderly (Abdi, Al-Hindawi, & Ng, 2018).

These rapid and widespread developments in AIS have the potential to bring enormous benefits to society but they also have the potential to introduce new—or amplify existing—risks to the safety of critical systems like healthcare and transportation. To maximize the benefits of AIS it will be necessary to ensure that these risks are robustly analyzed, governed, and regulated. One of the most critical risk management challenges is to ensure that the failures of AIS can be reliably identified, rigorously analyzed, and widely learnt from, so that the risks associated with these emerging technologies can be better understood and safety can be continuously improved (Macrae, 2019a). Learning from the failures of AIS is particularly urgent because many of these new technologies remain at early stages of experimentation and development—and some of those experiments are being conducted in the public arena and placing citizens at direct risk, such as self-driving cars learning by trial and error on public streets (Stilgoe, 2020) and medical triage chatbots being deployed and updated after encounters with real patients and concerned doctors (Lintern, 2021).

AIS technologies are developed and deployed within complex human, social, and organizational systems and the risks associated with AIS are therefore thoroughly sociotechnical (Salmon, Carden, & Hancock, 2020). However, the focus of much work to date has primarily been on defining broad ethical principles (Jobin, Ienca, & Vayena, 2019; IEEE, 2019), exploring issues of accountability and liability (Morley et al., 2020; Pöllänen, Read, Lane, Thompson, & Salmon, 2020), and addressing technical aspects of assurance (Brundage, Avin, & Wang, 2020; Hawkins et al., 2021). There has been little systematic consideration of the complex sociotechnical sources of risk and failure in AIS, or the organizational infrastructures required to learn from those failures (Bryson & Winfield, 2017; Elish, 2019; Stilgoe, 2018; Winfield & Jirotka, 2017). Accordingly, this article aims to explore and characterize the nature of sociotechnical risk in autonomous and intelligent systems by reanalyzing Uber's 2018 fatal autonomous vehicle accident (NTSB, 2019a), drawing on sociotechnical theories of risk and safety in complex systems. The overarching objective of this analysis is to develop an empirically grounded and theoretically informed framework that defines fundamental sources and patterns of sociotechnical risk that can threaten the safety of AIS, and which may be used to better understand the infrastructures needed to manage those risks and learn from AIS failure. The article concludes by exploring proposals for research, policy, and practice that may help move toward Asimov's (1990) aspirational future in which the failures of autonomous and intelligent systems will always yield experience that can be used to improve safety.

2. THE AUTONOMY EXPLOSION AND THE EVOLUTION OF SOCIOTECHNICAL RISK

Efforts to develop AIS technologies have exploded across a range of settings in recent years, from software-based agents that crawl the internet (Russell & Norvig, 2016) to robotic "embodied artificial intelligences" that walk the world (Winfield, 2012). These technologies are designed to act with some degree of autonomy and intelligence: that is, to be capable of engaging in some form of independent reasoning to produce behaviors that solve some specific objective without real-time human input. In healthcare, AIS are being developed to diagnose cancer (Ardila, Kiraly, & Bharadwaj, 2019), predict the onset of serious illness (Tomašev, Glorot, & Rae, 2019), and provide advice to patients (Babic, Gerke, & Evgeniou, 2021). In transport, self-driving cars (Krafcik, 2020), buses (Yu, 2021), and delivery robots (Heinla, 2021) already navigate public roads. In infrastructure, artificial intelligence manages energy networks (Evans & Gao, 2016) and in finance, AIS support algorithmic trading, credit scoring, and fraud detection (Chan et al., 2019). This autonomy explosion is striking in terms of the speed, scale, and variety of efforts underway to develop and deploy AIS. It is also striking because many of the target applications are within critical systems, in which failures pose significant risks to safety (Cummings, 2021). For instance, failures in healthcare advice chatbots may cause considerable harm if serious conditions are missed (Fraser, Coiera, & Wong, 2018). Failures in the ability of automated cars to recognize and avoid pedestrians or other hazards have already been fatal (NTSB, 2017, 2019a). Even momentary failures of small delivery robots have posed serious threats to public safety when they struggle to cross busy roads and trap wheelchair users in the path of oncoming vehicles (Ackerman, 2019).

Managing the safety of AIS is essential to building trust in these emerging technologies, but the safety implications of AIS are complex and remain poorly understood. Simplistic claims that AIS will reduce risk by removing fallible humans have been used to justify the rapid deployment of experimental systems with limited regulatory oversight (Dixon, 2020; Levin, 2016a; Mider, 2019; Ross, 2018). This ignores the inherently sociotechnical nature of AIS: that all technologies are designed, developed, built, deployed, maintained, supervised, operated, and governed by people (Reason, 1997); and those people necessarily work within, and are shaped by, complex social, cultural, and organizational processes (Hopkins, 2005; Pettersen Gould, 2021; Weick, 1987). Rather than removing the risks of human fallibility, AIS will instead transform and relocate the work of humans and their role as sources of both risk and safety in complex systems (Hancock, 2017; Murphy & Woods, 2009). The broad-based emergence of AIS therefore points to a new phase in the evolution of sociotechnical risk that will further complicate, rather than eliminate, human and social entanglements in technological safety (Bainbridge, 1983; Bradshaw, Bradshaw, Hoffman, & Woods, 2013; Salmon et al., 2020; Sarter, Woods, & Billings, 1997). A range of theories seek to explain the social, organizational, and technological processes involved in the failure of complex systems, encompassing organizational structure (Perrow, 1984), human activity (Reason, 1990), technological control (Leveson, 2011), and cultural incubation (Turner, 1976; Weick & Sutcliffe, 2003). These theories offer a rich and diverse conceptual foundation for understanding the emergence of failure and the nature of sociotechnical risk in AIS.

2.1. Robots Going Down the Street

One of the most visible and safety-critical applications of AIS is the development of self-driving cars, or autonomous vehicles (AV), designed to navigate from one place to another using a network of sensors, processors, and actuators that perform functions such as object detection, path prediction, motion planning, and hazard avoidance (NTSB, 2019b; Uber ATG, 2018a). At 9:58 p.m. on March 18th, 2018 in Tempe, Arizona, an AV being developed and tested on public roads by Uber's Advanced Technologies Group (ATG) fatally failed. Uber's AV collided with and killed Elaine Herzberg, who was pushing her bicycle across the street outside of a designated crossing area (NTSB, 2018). The AV made no attempt to brake or avoid Herzberg. The vehicle operator, responsible for monitoring the vehicle and intervening when necessary, did not apply the brakes until after Herzberg had been hit (NTSB, 2019a). The severity of the event—and its novelty and importance as the first pedestrian death involving an AV (Niedermeyer, 2019)—resulted in extensive media reporting, a lengthy federal accident investigation by the National Transportation Safety Board (NTSB, 2019a), and two safety reviews commissioned by Uber (Uber ATG, 2018b). Together, these revealed a range of contributory factors that arose not just in

Table I. Summary of Uber's Fatal Self-driving Collision with a Pedestrian, March 18th, 2018

Uber's Fatal Self-driving Crash: a Descriptive Summary
The crash At 9:58 p.m. on March 18th, 2018 in Tempe, Arizona, one of Uber's autonomous test vehicles collided with Elaine Herzberg as she pushed her bicycle across a road, causing fatal injuries. The car, a modified Volvo XC90, was being driven autonomously by a self-driving system (SDS) and monitored by a single vehicle operator. The car was traveling at 39 mph at the point of collision and made no attempt to stop. The vehicle operator did not apply the brakes until 0.7 seconds after impact (NTSB, 2019a). This was nearly a close call: police investigators calculated that Elaine Herzberg only needed time to walk an additional 2.1 feet to cross safely in front of the car (Stern, 2018).
Automated driving The SDS consisted of an array of sensors, software, and hardware that gathered and processed data about the environment to understand the car's location, identify objects and track and predict their movement, and plan and control the vehicle's motion and route (Uber, 2018a). Detected objects were classified and multiple paths were predicted based on tracking history and assumed goals. If a hazard was identified (e.g., an object in the vehicle's path) the motion plan was altered or emergency hazard avoidance initiated (NTSB, 2019a). The SDS detected the pedestrian 5.6 seconds before the impact but failed to correctly classify or predict her path, switching between classifications of "vehicle," "bicycle," and "other" (NTSB, 2019b). When an object was reclassified, historical path data became unavailable for path prediction. The system was also not designed to assign a goal to "other" objects or jaywalking pedestrians, relying on continuous tracking data to predict a path. So, when the pedestrian was detected as an "other" object in the vehicle's lane 1.5 seconds before impact, the SDS considered this to be a static object and just slightly altered the motion plan (NTSB, 2019a). At 1.2 seconds before impact the SDS reclassified the object as a bicycle and determined this was an emergency. However, an "action suppression" function delayed any emergency actions by 1 second, to check the hazard was not a false alarm and reduce the frequency of extreme braking events (NTSB, 2019b). The vehicle operator was not alerted that an emergency had been detected. Action suppression ended 0.2 seconds before impact, at which point the operator was alerted by an auditory alarm. The SDS was not designed to activate emergency braking if a crash was determined to be imminent—which it now was—as it was assumed that vehicle operators would intervene to avoid or mitigate the impact of a collision (NTSB, 2019a). Volvo's own in-built emergency braking system was disabled when the SDS was active (NTSB, 2019b).
Vehicle development and testing Uber had 40 test vehicles in Tempe which had driven this particular test circuit on public roads around 50,000 times (NTSB, 2019a). Human operators were relied on to monitor for failures and intervene in emergencies. Vehicle operators received three weeks of classroom and practical training that emphasized vehicle-handling skills and scanning for hazards such as jaywalking pedestrians, which operators encountered regularly during testing (NTSB, 2019a). The number of operators per vehicle had been reduced from two to one in October 2017 (NTSB, 2019a), and a team of specialist vehicle operators who tested the cars in difficult situations was disbanded (Wakabayashi, 2018). The operator involved in the collision had 152 hours experience in autonomous mode (NTSB, 2019a). The in-car camera showed the operator was not constantly monitoring the road. She was looking down, away from the road, for 34% of the 31.5 minutes the vehicle was moving and 23 times in the 3 minutes prior to the collision (NTSB, 2019a; 2019f). Records indicate her phone was streaming a TV show (NTSB, 2019a). No formalized or automated systems were in place to monitor the vigilance of vehicle operators (NTSB, 2019a). Vehicle operator shifts were around 8 hours with a requirement to take a 20–40 minutes break after 4.5 hours of continuous driving (NTSB, 2019a). No fatigue risk management system was in place (NTSB, 2019e) and there were reports of drivers feeling pressured to maximize miles driven (Bort, 2018d). Uber had no safety management system, no formal safety plans, and no dedicated safety manager to oversee the operational safety of its testing activities (NTSB, 2019a). New software could be deployed for testing on public roads without passing a defined set of track tests or formal safety requirements (NTSB, 2019e; Uber, 2018b). Vehicle operators and other staff were not able to halt testing if they identified a safety concern, and there were reports that safety incidents and concerns were not always rapidly acted on (Bort, 2018d; Efrati, 2018c). Uber had chosen Arizona for AV testing in part due to its limited regulation proudly proclaimed by its Governor, after California revoked Uber's licenses for flouting its regulations (Harris, 2018; Levin, 2016b, 2016b; Randazzo, 2019; Wong, 2016a, 2016b).

the functioning of the underlying technologies but in the social, cultural, and organizational practices of designing, supervising, managing, and developing those technologies (Table I). The extensive scrutiny of the Uber crash, coupled with the breadth and variety of safety weaknesses identified (Niedermeyer, 2019; Stanton, Salmon, Walker, & Stanton, 2019), has produced one of most detailed public analyses

of AIS failure to date, providing a unique case for developing an extensive analysis of the sociotechnical sources of risk in AIS.

3. METHODS AND APPROACH

The overarching aim of this analysis was to develop a theoretical framework that defined and characterized fundamental sources of sociotechnical risk in AIS, using the Uber event as a focused case study (Eisenhardt, 1989). Accordingly, the analytical approach adopted was qualitative and oriented to conceptual development. Public investigative reports on the Uber event were analyzed through constant comparative analysis (Glaser & Strauss, 1967; Locke, 2001; Strauss & Corbin, 1998)—an iterative coding process in which qualitative data on the event was systematically reviewed, categorized, and conceptualized to develop a higher-order theoretical account of the key contributory processes and sociotechnical patterns involved in the Uber crash (Turner, 1981, 1983). This analytical coding process was purposefully sensitive to existing sociotechnical theories of risk, accidents, and safety (e.g., Perrow, 1984; Turner, 1976, 1978; Vaughan, 1996), and sought to build a theoretical framework that integrated and extended current accounts of sociotechnical risk (Glaser, 1978). The analysis proceeded in three interrelated phases. First, key investigative reports on the Uber event were identified through searches of the NTSB's Accident Docket database (NTSB, 2021), Uber corporate websites, and reputable media outlets. A total of 48 reports were identified, including official investigation report materials and corporate documentation, supplemented by a set of in-depth news reports that provided detailed information and context, interviews with current and former staff, and public disclosure of some additional corporate materials (see Table II). Second, all reports and materials were analyzed to identify and code key factors and processes in the development of the Uber event. This qualitative coding proceeded in several cyclical stages. It began with "initial" coding of each report to identify, label, and define all relevant aspects of the Uber event. It then moved to a process of "core" coding that compared, combined, and organized these initial codes into a smaller number of higher-order concepts of AIS failure. Analysis moved between these stages of initial and core coding until all investigative reports had been analyzed and a conceptually coherent coding structure emerged that could account for and explain all the key patterns of failure described in the data. Finally, the resulting coding framework was reviewed and compared with foundational concepts and theories in the extant literature, and further refined and organized into an overarching theoretical framework structured around five core sources of sociotechnical risk with each illustrated by a range of indicative patterns of sociotechnical failure. Given the contemporaneous and contested nature of some of the public reporting on the Uber event, the emerging nature of the field of AIS, and the analytical focus on a single case study, the theoretical framework developed here is necessarily provisional and represents an initial process of theorizing, rather than a single settled and final theory (Weick, 1995).

4. SOCIOTECHNICAL SOURCES OF RISK IN AUTONOMOUS AND INTELLIGENT SYSTEMS

Analyzing publicly available reporting on the 2018 Uber AV crash allows the characterization of five fundamental sources of sociotechnical risk in AIS, each indicated by a range of particular patterns of sociotechnical failure. *Structural* sources of risk arise from interdependencies and interactions between different parts of the technical and social structures that constitute AIS. *Organizational* sources of risk arise from the social processes, organizing activities, and human and contextual factors that underpin AIS. *Technological* sources of risk arise from the capabilities, affordances, and constraints inscribed into and produced by the material technologies of AIS. *Epistemic* sources of risk arise from the ways that knowledge and ignorance are constructed in relation to, and within, AIS. *Cultural* sources of risk arise from the collective values, beliefs, norms, and practices that surround and shape AIS. Taken together, these domains of structural, organizational, technological, epistemic, and cultural risk form a SOTEC framework of sociotechnical risk in AIS (Fig. 1). Each of these five intersecting domains draws on a distinct theoretical lineage and provides a broad conceptual lens through which to identify and understand specific sources and patterns of sociotechnical risk in AIS.

4.1. Structural Sources of Risk

The structure of a sociotechnical system determines how different parts of the system interact with each other, and shapes how failures evolve.

Table II. Public Sources of Data and Evidence Drawn on to Analyze the Uber Crash

Public Reporting on Uber's Fatal Self-driving Crash	
Federal investigation materials	
The National Transportation Safety Board (NTSB) produced a final major investigation report and alongside this published 43 analytical documents and data sheets in the public "accident docket," as well as a set of documents presented at to the public Board meeting that concluded the investigation. Eighteen core investigative documents were analyzed:	
Final and summary accident reports (2 reports)	Final accident report (NTSB, 2019a); Preliminary accident report (NTSB, 2018)
Specialist factual reports (6 reports)	Vehicle Automation Report (NTSB, 2019b); Highway Factors Report (NTSB, 2019c); Vehicle Factors Report (NTSB, 2019d); Operations Factors Report (NTSB, 2019e); Human Performance Report (NTSB, 2019f); Onboard Image and Data Recorder Report (NTSB, 2019g)
Party submissions to investigation (3 reports)	Uber ATG submission (NTSB, 2019h); Volvo Cars submission (NTSB, 2019i); Thatcham safety test submission (NTSB, 2019j)
Board meeting presentations (7 reports)	Board Meeting Summary (NTSB, 2019k); Opening Statement (NTSB, 2019l); Crash overview (NTSB, 2019m); Pedestrian and vehicle operator (NTSB, 2019n); Managing risk of ADS testing (NTSB, 2019o); Uber ATG operations (NTSB, 2019p); Testing automated vehicles (NTSB, 2019q)
Corporate safety reviews	
Uber conducted an internal safety review and commissioned an external safety review from a team led by a former leader of the NTSB. The findings and recommendations of these reviews were published alongside Uber's annual public safety report. Further background information was provided in related company blog posts.	
Uber safety reviews (3 reports)	Annual public safety report (Uber ATG, 2018a); Internal and external reviews (Uber ATG, 2018b); Independent safety review (Hart, Dombroff, & Tochen, 2018)
Uber corporate blog posts (4 reports)	Self-driving launch (Uber ATG, 2016); Learning from the past (Uber ATG, 2018c); Principled approach to safety (Uber ATG, 2018d); Groundwork for self-driving safety (Uber ATG, 2019)
Investigative press reporting	
A set of press reports presented detailed, substantive, and new information on the background to the crash and the organizational context and culture at Uber ATG not explored by the NTSB investigation or corporate reviews. These press reports were well-sourced, with many based on multiple company sources. Several reports publicly disclosed the details of emails, presentations, and other relevant materials and company data.	
Historical reporting on Uber safety (6 reports)	Carson (2016); Levin (2016a); Levin (2016b); Wong (2016a); Wong (2016b); Nguyen (2017)
Reporting on reasons for the crash (9 reports)	Wakabayashi (2018); Harris (2018); Efrati (2018a); Efrati (2018b); Stern (2018); Bort (2018a); Efrati (2018c); Randazzo (2019); Bort (2018b)
Reporting on context and response (8 reports)	Bort (2018c); Bort (2018d); Bort (2019); Efrati (2019a); Efrati (2019b); Efrati (2020); Marshall (2018); Wakabayashi and Conger (2018)

Structural arrangements can act as sources of risk in AIS by amplifying or transmitting local failures in ways that disable the entire system. The Uber AV accident emerged from structurally interlinked failures that interacted across different parts and at different scales of the system, encompassing the design of the self-driving system (SDS), the role of vehicle operators, the decisions of engineers and managers, the processes of vehicle testing, and the actions of other road users and regulators. The structural characteristics of this sociotechnical system allowed failures in one area to rapidly degrade or impact on other parts of the system. Prior theoretical accounts of structural risks in sociotechnical systems have focused on two key properties: complexity and coupling (Perrow, 1999, 2011; Sagan, 1995). High levels of interactive complexity—where many activities or components interact in unpredictable ways—means that disruptions in one part of a system can dramatically enlarge by impacting the performance of many

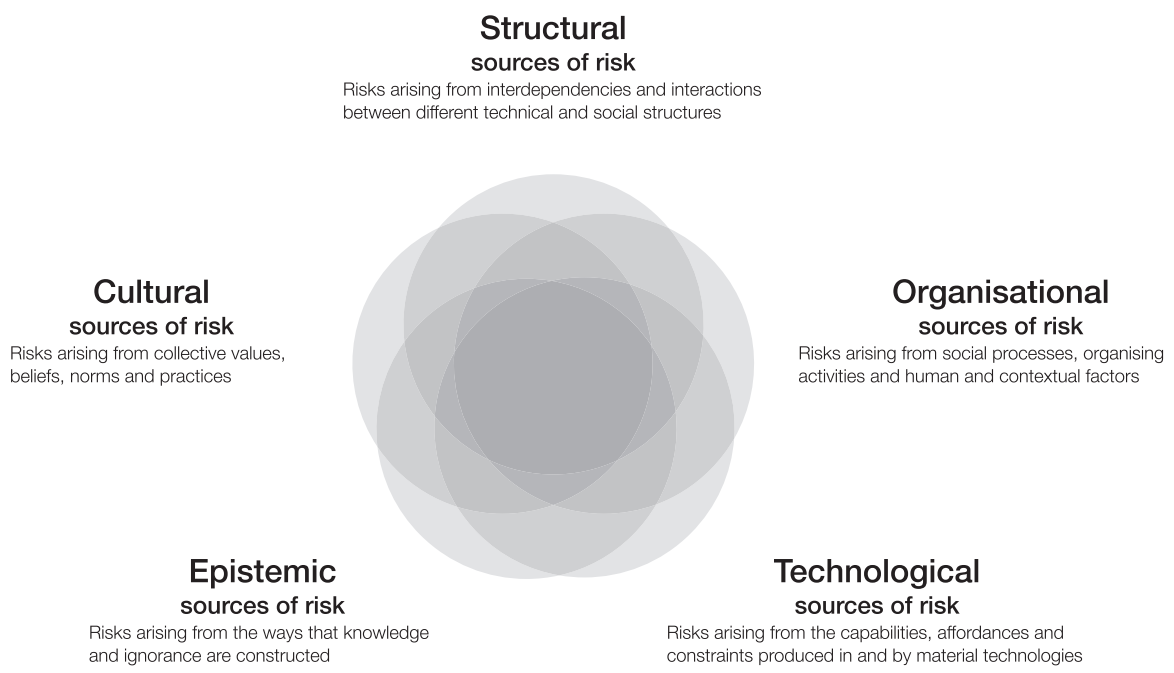

Fig 1. SOTEC framework of sociotechnical sources of risk

other parts in ways that are hard to predict (Hulme, Stanton, Walker, Waterson, & Salmon, 2021; Weick, 2004). High levels of coupling—where components or activities in one part of a system are heavily dependent on those in other parts—means that failures can rapidly cascade through systems in ways that are difficult to avert (Downer, 2009, Hulme et al., 2021). The inherent potential for rapid and unpredictable escalations of failure means that catastrophic system breakdowns may be inevitable—or "normal"—in interactively complex and tightly coupled systems (Gephart, 2004; Perrow, 1999).

A range of structural characteristics contributed to the Uber AV accident, illustrating a number of structural patterns of risk that can emerge in AIS (Table III). An action suppression function that delayed the vehicle responding to perceived hazards (NTSB, 2019a) allowed a minor disruption—a pedestrian crossing the road—to quickly amplify and grow into a critical situation ("disruption amplifiers"). The functional structure of the SDS meant that repeated failures of the perception system to correctly categorize the pedestrian then prevented the prediction system from calculating a predicted path for the hazard, and once the pedestrian was in front of the vehicle the motion planning system was unable to apply emergency brakes (NTSB, 2019b) as that function was disabled in situations when collisions were deemed imminent ("failure cascades"). The functional safety of the vehicle and the role of vehicle operators were heavily predicated on capacities for ongoing, real-time—and inherently unreliable—human vigilance (NTSB, 2019a) and intervention ("vigilance dependencies"). And the structure of the vehicle development and testing program was highly permissive, allowing the rapid and regular deployment of new iterations of the autonomy software on public roads with few controls (NTSB, 2019e; Uber ATG, 2018b), meaning that weaknesses in software development were tightly coupled to on-road vehicle behavior ("test permeabilities"). These structural properties meant that algorithmic failures could cascade directly onto the street and escalate rapidly into catastrophic operational failure.

4.2. Organizational Sources of Risk

Organizational activities can act as sources of risk in AIS, particularly when organizational contexts are insensitive to human performance characteristics and organizational processes are ineffective at detecting and managing errors and disruptions. A network of organizational, contextual, and human factors contributed to the Uber AV accident, including weaknesses in supervisory systems, gaps in safety expertise and leadership, poor human–machine

Table III. Definitions of Structural Patterns of Risk in Autonomous and Intelligent Systems and Illustrative Examples from Public Reporting on the Uber Autonomous Vehicle Crash

Structural Patterns of Sociotechnical Risk
Disruption amplifiers
System features which cause disruptions or failures to enlarge, expand or develop into more critical situations which are harder to deal with or recover from.
Example: Action suppression functions in the Uber self-driving system (SDS) prevented the vehicle responding to any perceived hazard for 1 second, in case the perception system had generated a false alarm or a hazard resolved itself (NTSB, 2019a, 2019b). However, when a persistent hazard did exist, action suppression could allow a situation to grow more critical and make a collision harder to avoid.
Failure cascades
Structural characteristics that allow interlinked failures to cascade rapidly through interdependent functions of a system with few opportunities for identification or intervention.
Example: The Uber SDS failed multiple times to correctly classify a pedestrian crossing the road (NTSB, 2019b), which prevented a predicted path being assigned, which precluded the vehicle taking early avoiding action, while action suppression rules prevented an immediate response once a hazard was detected, exacerbated by the prevention of emergency braking when a collision was deemed imminent (NTSB, 2019a).
Vigilance dependencies
System functions that rely on active, real-time, moment-by-moment human vigilance to monitor automated behavior and detect and address failures.
Example: Uber confirmed to accident investigators that the developmental SDS depended on an attentive vehicle operator to monitor and intervene if the system fails (NTSB, 2019a), and publicly claimed that Californian self-driving regulations were not appropriate as their vehicles required a human operator at the controls at all times (Uber ATG, 2016).
Test permeabilities
New iterations of developmental or operational autonomous systems are released for testing into the public domain with few safety controls, criteria, or assurance processes.
Example: Processes for requesting and conducting testing of the Uber SDS on public roads did not incorporate integrated safety assessments or requirements to meet a set of performance standards in on-track testing prior to on-road testing (NTSB, 2019e; Uber ATG, 2018b), with little formal guidance or safety criteria defining levels of acceptable system performance prior to release onto public roads (Efrati, 2018c).

interfaces, and the absence of formalized safety management systems or regulatory requirements. These factors reduced or removed organizational capacities to prevent, catch, or correct failures. Prior theoretical approaches to organizational risk in sociotechnical systems consider errors, failures, and fluctuations to be inherent to all organized activity, and therefore focus on the organizational mechanisms needed to prevent or recover from disruption (Pettersen Gould, 2021; Rasmussen, 1990; Reason, 1997; Roe & Schulman, 2008). These mechanisms are typically conceptualized as safety defenses or barriers (Hollnagel, 2004; Reason, 1990; Svenson, 1991)—ranging from "hard" defense such as physical barriers or multiple back-up systems to "soft" defense such as procedural controls or training programs (Reason, Hollnagel, & Paries, 2006)—or as capacities for resilience that enable rapid identification and flexible adaptation to unexpected events (Hollnagel, Paries, Woods, & Wreathall, 2012; Macrae, 2014a; Wiig et al., 2020). Defenses and adaptations will themselves always be partial or fallible, with weaknesses arising from latent organizational factors such as poorly designed equipment or inadequate resourcing. Organizational safety therefore depends on multiple layers of defense or diverse capabilities for resilience (Rasmussen, 1997; Reason, 1997), and catastrophic system failures occur when a range of factors combine to overwhelm these safety defenses and adaptive capacities.

Various organizational weaknesses and human and contextual factors contributed to the Uber AV accident and illustrate a set of organizational

Table IV. Definitions of Organizational Patterns of Risk in Autonomous and Intelligent Systems and Illustrative Examples from Public Reporting on the Uber Autonomous Vehicle Crash

Organizational Patterns of Sociotechnical Risk
Invisible automation Weaknesses or gaps in processes that maintain awareness, provide insight and issue alerts regarding the status, activities, and decisions of automated systems. Example: No alerting process existed to inform vehicle operators that the Uber SDS had detected a potential hazard in the driving environment but was initiating action suppression and delaying any automated response (NTSB, 2019a, 2019o).
Governance gaps Gaps in organizational processes and systems that set standards for safety, monitor safety performance, and initiate action to address safety deficiencies. Example: Uber had no safety management system to govern and assure the operational safety of its vehicles, with no overarching safety plan or standard operating procedures setting out roles, responsibilities, and processes for the analysis and management of risk in its driving operations (NTSB, 2019a, 2019e).
Regulatory voids Absence of regulatory requirements, performance standards and associated oversight activities to assure the safe development, testing, deployment, and operation of automated systems. Example: The on-road testing of Uber's SDS was subject to little regulatory oversight with no regulatory systems for State monitoring of safety performance in Arizona (NTSB, 2019a, 2019e; Wakabayashi, 2018), and only a voluntary and underspecified process for safety self-assessment at the Federal level, with self-assessments not subject to formal regulatory review (NTSB, 2019a, 2019e, 2019q).
Supervisory degradation Reduced or inadequate organizational arrangements to support, monitor, and assure the activities of supervising automated systems. Example: Uber instituted no routine processes or automated system to monitor vehicle operator vigilance and reduced the number of operators per vehicle from two to one (NTSB, 2019a), reducing the cognitive capacity available in the vehicle to monitor both the SDS and the driving environment while also removing the opportunity for mutual social reinforcement of appropriate in-car operator norms and behavior.
Competency limits Limitations or gaps in the roles, expertise, and experience available for analyzing and managing all aspects of safety across the development and deployment of an automated system. Example: Uber did not have a dedicated safety division, team, or personnel with responsibility for managing the operational safety of its vehicles (NTSB, 2019a, 2019e). General safety responsibilities were combined with operational leadership responsibilities (NTSB, 2019a) and relevant safety competencies were not defined for key roles (Hart et al., 2018).

patterns of risk that can emerge in AIS (Table IV). Perhaps most fundamentally, Uber ATG had no formalized safety management systems in place (NTSB, 2019e, 2019a) to analyze and manage the operational risks of its development and testing activities ("governance gaps"), and had no dedicated roles with responsibility for—or expertise in—the management of operational risk and safety ("competency limits") (Hart et al., 2018; NTSB, 2019a; Uber ATG, 2018b). This allowed fundamental operational risks, such as automation complacency in vehicle operators (Parasuraman & Manzey, 2010) and hazards associated with pedestrians crossing the street, to be largely overlooked (NTSB, 2019a, 2019b). Vehicle operators were provided with little insight into the ongoing activities or decisions of the automated systems they were responsible for monitoring ("invisible automation"), with no alerts provided to the operator when the SDS detected a potential hazard but initiated action suppression (NTSB, 2019a, 2019o). And, while active monitoring and supervision of the SDS by an operator was a critical safety function, these activities were not themselves routinely monitored within the organization (NTSB, 2019a,

2019f), and the capacity for monitoring vehicle behavior had been degraded by reducing the number of operators in each vehicle (Bort, 2018a; NTSB, 2019a) from two to one a few months before the collision ("supervisory degradation"). More broadly, development and testing of Uber's SDS on public roads was subject to little regulatory oversight ("regulatory voids") at both the State (NTSB, 2019a, 2019e; Wakabayashi, 2018) and Federal levels, which focused instead primarily on voluntary safety self-assessment (NTSB, 2019a, 2019e).

4.3. Technological Sources of Risk

AIS are built on technologies with specific capabilities and constraints, and which provide different affordances for interaction and use—and for failure (Barley, 2020; Beane & Orlikowski, 2015). These technological properties can act as sources of risk when they result in AIS failing to perform as intended, behaving in unexpected ways or becoming challenging to control. The Uber AV accident resulted from a complex web of technical failures and limitations including problems with the car's ability to perceive objects and predict the path of pedestrians, action suppression rules that delayed the car's responses to perceived hazards, and constraints on emergency braking functions. These technological weaknesses created major gaps in the car's ability to operate safely. Theories of technological risk in sociotechnical systems focus on problems that emerge in the design, use, and control of technical objects (Collingridge, 1996; Leveson, 2011). Technological weaknesses can emerge during design and development: coding errors or mistaken assumptions may become embedded in technical objects (Amodei et al., 2016; Leveson & Turner, 1993), technological reliability and robustness may be poorly engineered or not well understood (Cummings, 2021; Downer, 2009; Kletz, 1994), and design features may complicate or preclude effective human–machine interaction (Carroll, 2003; Norman, 2013). Similarly, failures can emerge in the monitoring and control of technologies: performance standards may be poorly defined or misapplied, and monitoring processes may be degraded or focus on inappropriate aspects of performance (Leveson, 2004, 2011), allowing technical systems to move beyond the boundaries of safe operation (Dekker, 2011; Rasmussen, 1997).

Limitations in technological capabilities and control defined key aspects of the Uber AV accident and indicate a set of technological patterns of risk that can emerge in AIS (Table V). Uber's developmental vehicles were reported to regularly encounter on-road situations they found challenging (Bort, 2018a; Nguyen, 2017; Wakabayashi, 2018; Wakabayashi & Conger, 2018), with relatively high levels of damage and frequent safety events (Bort, 2019; Efrati, 2018c), suggesting a degree of immaturity in some elements of the SDS that needed further development before operating on public roads ("automation immaturity"). Key safety capabilities of the vehicles were constrained to optimize overall performance ("capability constraints"), with the SDS designed to be unable to apply emergency brakes when it calculated a crash was imminent, to reduce the frequency of sudden braking events (Bort, 2018a; NTSB, 2019a). The design of the perception and prediction systems inadvertently disguised or complicated the presence of potential hazards ("hazard masking"): when the SDS changed the classification of an object the historical path data for that object was no longer available for predicting its trajectory (NTSB, 2019b); and the SDS was not designed to assume pedestrians in the road might be trying to cross it, and instead attempted to predict the path of "jaywalking" pedestrians by continually tracking their movement (NTSB, 2019a, 2019b). The SDS was also designed to temporarily ignore perceived hazards so as to smooth the car's motion ("sensitivity smoothing"), delaying avoidance action for 1 second in an attempt to reduce excessive braking (Efrati, 2018a; NTSB, 2019a). And the vehicle relied on its experimental autonomy system to provide all technical collision avoidance functionality; the simpler on-board Volvo emergency braking system was disabled ("autonomy reliance") (NTSB, 2019a).

4.4. Epistemic Sources of Risk

Developing and operating AIS depends upon building and maintaining knowledge of how a system is working—and how and why it might fail. Epistemic challenges can act as sources of risk when this knowledge is partial, incorrect, or out of date, allowing pockets of ignorance to hide unexpected threats. A variety of epistemic processes shaped the Uber accident, which involved an experimental vehicle that was being used to test and understand system performance—activities which were reportedly constrained by delays in reviewing surprising incidents, gaps in the exploration of on-road experiences, and limitations on accessing safety-relevant data. Theoretical perspectives on epistemic sources

Table V. Definitions of Technological Patterns of Risk in Autonomous and Intelligent Systems and Illustrative Examples from Public Reporting on the Uber Autonomous Vehicle Crash

Technological Patterns of Sociotechnical Risk
Automation immaturity Automated systems regularly encounter situations, objects or hazards that are not recognized or are beyond the system's capabilities, leading to frequent failures. Example: The Uber automated driving system reportedly experienced high levels of disengagements and required regular human intervention during on-road testing (Bort, 2018a; Wakabayashi, 2018). "Raw miles per intervention" in the three weeks before the accident was reported as one disengagement every 1–3 miles and a vehicle was reportedly being damaged on average almost every other day in the month before the accident (Efrati, 2018c). Frequent braking and swerving by the SDS reportedly gave a vehicle operator mild concussion in late 2017 (Bort, 2018d).
Capability constraints Technical features that reduce or constrain the safety capabilities of a system in order to optimize other aspects of system performance, efficiency, or experience. Example: The Uber SDS was designed to be unable to activate emergency braking in circumstances where a collision was determined to be imminent, to reduce the frequency of sudden braking (NTSB, 2019a, 2019b). This decision reportedly coincided with the development of new "rider-experience metrics" that targeted no more than one "bad experience" per ride (Bort, 2018a). Uber was reported to have initially disabled both emergency swerving and braking after these metrics were announced but soon reinstated the former (Bort, 2018a).
Hazard masking Technical processes or features that result in hazards being inadvertently hidden, disguised, or rendered ambiguous. Example: When the Uber SDS changed the classification of an object it was unable to access historical path data for the object, meaning that a reclassified object was perceived as a new nonpersistent object with no movement history available to predict a trajectory (NTSB, 2019a, 2019b). The SDS was also not designed to assign a goal to pedestrians walking in the road outside an official crossing area (jaywalkers), so the system did not assume that a pedestrian in the road might be trying to cross it, and could only attempt to predict the path of a pedestrian based on tracking movement (NTSB, 2019a, 2019b).
Sensitivity smoothing Technical features that attenuate warning signals or reduce responses to perceived hazards to smooth the behavior of an automated system. Example: When the SDS detected a hazard, an action suppression rule delayed any planned motion for 1 second to reduce unnecessary extreme maneuvers in case it was a false alarm or the hazard resolved itself (NTSB, 2019a, 2019b). If the hazard was still deemed to be present after 1 second then the action would be carried out and an audible alert would be provided to the vehicle operator that action was being taken.
Autonomy reliance Dependance on autonomous functionality for technical safety protections and controls, to the exclusion of other lower-technology components that could provide safety redundancy. Example: The Uber vehicle relied solely on its SDS to provide all hazard detection and avoidance functions, and the in-built and independent Volvo emergency braking functions were disabled (NTSB, 2019a). It was established that, if active, this relatively simple and reliable system would likely have prevented a fatal collision (NTSB, 2019a).

of risk in sociotechnical systems assume that surprising failures emerge from gaps or errors in existing knowledge of how systems work (Downer, 2011, 2019; Macrae, 2009; Smithson, 1989, 1990). These may be deep and fundamental, where foundational scientific theories or technical models are erroneous or entirely absent (Downer, 2011, 2019). Or they may be more localized, where working knowledge or practical models within a particular context are wrong or incomplete (Macrae, 2009, 2014; Turner, 1978; Weick & Sutcliffe, 2001). Epistemic challenges are amplified by complex, innovative

Table VI. Definitions of Epistemic Patterns of Risk in Autonomous and Intelligent Systems and Illustrative Examples from Public Reporting on the Uber Autonomous Vehicle Crash

Epistemic Patterns of Sociotechnical Risk
Learning lag Operational and developmental activities exceed the capacity, systems, and resources available to analyze and learn from those activities and the surprises that they generate. Example: Several significant near-miss events, including one test vehicle driving on the pavement and another nearly causing a collision with a vehicle in another lane, were reportedly not reviewed for several days after they occurred, until they were pursued by an experienced manager who raised concerns that the fleet size was larger than the resources available to review and analyze the events being generated (Efrati, 2018c).
Operational disengagement Limited or reduced efforts to gather data on, engage with, and make use of insights drawn from the operational experiences of people interacting with an autonomous system. Example: Uber was reported to have disbanded a small group of drivers dedicated to stressing the cars in particularly challenging situations (Wakabayashi, 2018), appears not to have built strong information sharing processes between vehicle operators and system developers (Bort, 2018a; Efrati, 2020; Hart et al., 2018; NTSB, 2019e), and moved from two to one operator per vehicle (NTSB, 2019a), reducing in-car capacity to identify and record interesting or surprising vehicle behavior.
Insensitivity to experience Failure to anticipate, notice, or effectively explore the safety implications of events experienced during developmental, testing, or operational activities. Example: Vehicle operators regularly reported encountering pedestrians jaywalking during testing operations (NTSB, 2019a) and scanning for pedestrians and jaywalkers was a key part of operator training (NTSB, 2019a), clearly demonstrating that "Pedestrians crossing a road midblock should be an anticipated safety risk when testing in urban environments" (NTSB, 2019a, p.39), but the SDS capabilities and protections around handling pedestrians crossing the road outside of crossing areas remained underdeveloped.
Simulatory inattention Peripheral or limited use of different forms of simulation to explore, test, train, and improve the behavior of an autonomous system. Example: Developing and using software to simulate and test car behavior was reportedly not a central part of the Uber AV development and testing program (Bort, 2018a; NTSB, 2019e): the development of vehicle self-driving software was reportedly prioritized over the development of simulation software, with simulation engineers paid less than colleagues working in other areas and incompatibilities between simulation software and self-driving software complicating efforts to run simulation tests (Efrati, 2018b, 2019b, 2020).
Competitive secrecy Reluctance to create or share safety data due to fears it may disclose commercially sensitive information or reveal performance weaknesses in a competitive arena. Example: Sharing of safety data appears to have been limited within Uber, with an experienced manager recommending a few days before the accident that details on safety events should be circulated within the organization and access to the safety incident database should be expanded (Efrati, 2018c), and staff were reportedly on occasion dissuaded from seeking out data related to limitations and gaps in safety performance and described that suggestions to compile details of safety requirements and tests were not well received (Bort, 2018d; Efrati, 2018c).

technologies that create regular opportunities for surprise (Downer, 2011; Perrow, 1999), and by social dynamics that promote the differential distribution of ignorance—such as professional boundaries or organizational silos—or that purposefully produce secrecy, such as the withholding of confidential information (Burrell, 2016; Costas & Grey, 2016; McGoey, 2019; Turner, 1978; Vaughan, 1996).

Epistemic challenges and limitations characterized key aspects of the Uber AV accident, and indicate important epistemic patterns of risk that can emerge in AIS (Table VI). The development of AIS such as an autonomous vehicle involves learning

about the functioning, behavior and limitations of complex new technologies that are inherently challenging to understand (Ford, 2018; NTSB, 2019a). These processes of learning appear to have been hampered by delays and limitations in the capacity to analyze and review unexpected operational events ("learning lag"), with on-road testing activities reportedly exceeding the analytical capacity to triage and review events (Efrati, 2018c). Opportunities to gather rich information and insights into the on-road behavior of vehicles were seemingly not fully engaged with ("operational disengagement"), with gaps and weaknesses identified in communication channels between vehicle operators and system developers (Bort, 2018a; Efrati, 2020; Hart et al., 2018; NTSB, 2019e), a specialist team of vehicle stress-testers being disbanded (Wakabayashi, 2018), and each car having only one operator to notice and record interesting vehicle behavior while also monitoring the SDS and road environment (NTSB, 2019a). There was an apparent insensitivity to events and experiences encountered during vehicle testing, which should have provided valuable opportunities to learn about risks and improve safety ("insensitivity to experience"). For example, safety drivers reported that jaywalking pedestrians were regularly encountered by test vehicles (NTSB, 2019a), but these events do not appear to have received close scrutiny within the organization or triggered focused efforts to improve the way the SDS recognized or handled this fundamental hazard of urban driving. The development and testing of the SDS relied heavily on driving cars on public roads and simulation techniques were reportedly not valued as a fundamental part of the development program ("simulatory inattention"). As a result, vehicle and software behavior appears not to have been systematically or extensively explored through computer simulation prior to running cars on public roads (Efrati, 2018b, 2019b, 2020; NTSB, 2019e). Finally, there were reports of reluctance to openly share safety-relevant information widely across the organization, particularly when that information might be perceived to question vehicle or organizational performance ("competitive secrecy"), with some staff reportedly dissuaded from developing safety requirements or requesting data (Bort, 2018d; Efrati, 2018c).

4.5. Cultural Sources of Risk

Cultural patterns of thinking and acting influence how AIS are developed and operated within organizations—and how failures emerge and are interpreted. Cultural characteristics can act as sources of risk by supporting behaviors and beliefs that move a system toward the limits of safe operation, and which lead to warning signs being misinterpreted, minimized, missed, or ignored. A range of cultural factors were reportedly associated with the Uber accident, including a focus on inappropriate performance metrics, production pressures and a perceived race for corporate survival, the disempowerment of operational staff, and faulty assumptions regarding the efficacy of human vigilance and the risks of on-road testing. Theoretical accounts of cultural risk in sociotechnical systems explain how collective practices, norms, values, and assumptions deeply shape organizational behavior and what organizations pay attention to—and what they ignore. Culturally shaped patterns of communication can disempower certain groups and disincentivize people from raising concerns (Morrison & Milliken, 2000; Turner, 1978; Weick & Sutcliffe, 2003). Collective values and beliefs can focus organizational attention on particular data and metrics while systematically discounting others (Hopkins, 1999a, 2005; Turner & Pidgeon, 1997). Shared norms can gradually shift to accommodate and normalize deviance from acceptable standards (Vaughan, 1996) and shared practices can drift away from established and understood ways of working (Snook, 2000). And cultural assumptions about how systems work can generate unwarranted optimism and misplaced trust (Turner, 1979; Turner & Pidgeon, 1997), and create disjunctions between what people believe is happening in a sociotechnical system and what actually is (Hopkins, 1999b; Pidgeon & O'Leary, 2000; Turner, 1994).

A range of cultural processes were associated with the Uber AV accident, and indicate several cultural patterns of risk that can emerge in AIS (Table VII). Considerable leadership attention was focused on improving performance against a widely reported metric of the number of autonomous miles driven (Bort, 2018a; Uber, 2018a)—a metric largely unrelated to progress in developing underlying self-driving technologies (Efrati, 2019a)—and the value placed on amassing autonomous miles appears to have influenced a range of key operational decisions ("performative production") (Bort, 2018d). Uber ATG was working to meet extremely ambitious targets to bring a consumer-ready driverless taxi to market (Efrati, 2018a; Wakabayashi, 2018), an ambition linked to publicly espoused beliefs that this was necessary for corporate survival (Carson, 2016) and fears

Table VII. Definitions of Cultural Patterns of Risk in Autonomous and Intelligent Systems and Illustrative Examples from Public Reporting on the Uber Autonomous Vehicle Crash

Cultural Patterns of Sociotechnical Risk
Performative production Organizational attention and activities focus on maximizing performance on narrow public metrics which do not represent underlying quality, safety, or improvement of the system. Example: There was reportedly a significant focus within Uber on increasing on-road autonomous miles driven, a metric which was widely reported and publicly focused on in the AV industry at the time (Marshall, 2018). Uber later acknowledged this metric may create "perverse incentives" (Uber, 2018a, p66) and this focus was reported to have influenced organizational activities including a shift to using only one operator per vehicles and reported pressures on vehicle operators to maximize mileage (Bort, 2018a, 2018d).
Existential pressure Overambitious targets and production pressures arising from fears that the existence of the organization is at stake in a competitive race to market. Example: Uber ATG was reportedly aiming to meet an ambitious internal goal of launching a driverless taxi service by the end of 2018 (Efrati, 2018a; Wakabayashi, 2018). In 2016 the former Uber CEO had described autonomous vehicles as an existential threat for the company, and that if Uber was not first or one of the first to market then "Uber is no longer a thing" (Carson, 2016). The new CEO was reportedly considering shutting down the AV unit (Bort, 2018a), and the AV unit's leaders were reportedly focused on improving the vehicle to impress the new CEO on a test drive (Bort, 2018a) alongside considerable attention on improving passenger experience, with new "rider-experience metrics" reportedly issued in November 2017 that allowed only one "bad experience" per ride (Bort, 2018a). A few days later emergency braking capabilities were removed from the SDS to reduce unnecessary extreme maneuvers and justified on the basis of safety to avoid surprising other road users (Bort, 2018a).
Concern quashing Staff disempowered or discouraged in raising safety concerns, speaking up, or challenging assumptions due to fears of punitive responses or lack of support. Example: Uber ATG staff were not empowered or authorized to "stop the line" and ground the fleet of test vehicles if they were concerned that the vehicles were not safe (Efrati, 2018c), though this was viewed as standard practice at other AV developers. There were reports that staff were dissuaded from or fearful of raising concerns and that speaking up to question decisions was not encouraged by some leaders (Bort, 2018a, 2018b, 2018d), and nonpunitive channels for reporting safety concerns had not been implemented before the accident (Hart et al., 2018; Uber, 2018a).
Developmental disintegration Norms and values that privilege rapid, widely distributed development and public-domain testing and deprioritize integrated safety oversight and assurance. Example: Uber ATG had not developed a set of clear requirements or an integrated system to review, approve, and monitor the use of test vehicles on public roads (Efrati, 2018c; NTSB, 2019b; Uber, 2019e) and reportedly had multiple teams working on different aspects of the vehicle with limited coordination between different testing activities (Bort, 2018a).
Presumptive reliability Assumptions and beliefs that human vigilance is effective for monitoring complex automated systems for long periods of time and requires little support or monitoring. Example: The Uber SDS was critically dependent on the vigilance and attentiveness of a human operator to detect and intervene when it failed (NTSB, 2018). At the same time, the vigilance, behavior, and attentiveness of operators was not routinely monitored via the vehicle's inward-facing cameras, which were only reviewed on an *ad hoc* basis with no records kept of these reviews, and vehicle operators were expected to work on their own monitoring the SDS for relatively long periods (NTSB, 2019a; 2019e).

that the self-driving unit may be shut down (Bort, 2018a). This appears to have created considerable production pressures (Bort, 2018a; Efrati, 2018c) and a premature focus on optimizing passenger experience (Efrati, 2018a) ("existential pressure"). Cultural norms and communication patterns seemingly minimized the impact of safety concerns: engineers and operational staff did not have authority to "stop

the line" and pause testing operations if they identified problems (Efrati, 2018c), serious safety incidents could reportedly take days to review (Efrati, 2018c), and reports indicated that some leaders were reluctant to engage with dissenting voices and concerns (Bort, 2018a, 2018b, 2018d) ("concern quashing"). Corporate values and norms prioritized rapid development and real-world testing to find and fix problems in the SDS, and deemphasized the importance of integrated, proactive and centralised risk management (Efrati, 2018c; NTSB, 2019a, 2019e; Uber, 2018b) ("developmental disintegration"). And the use of on-road testing appears to have been heavily predicated on unchallenged assumptions about the reliability and safety of human vigilance over prolonged periods of time (NTSB, 2018, 2019a), resulting in little monitoring of or support for operator vigilance (NTSB, 2019e) ("presumptive reliability").

4.6. A Note on Sociotechnical Interactivity

While the primary purpose of the analysis developed here is to conceptually unravel and differentiate these five domains of sociotechnical risk, the SOTEC framework also provides a foundation for understanding the complex interactions that can occur between different sociotechnical processes in the emergence of AIS failures. Each of these five domains of sociotechnical risk is deeply interrelated to and constitutive of the others, with the patterns of risk identified here amplifying, reinforcing, interacting, and overlapping with one another (Fig. 1). For instance, in the Uber AV accident, the cultural patterns of *existential pressure* and *performative production* seem to have reinforced one another, as concerns about corporate survival were aligned with a focus on increasing autonomous mileage. These cultural patterns, in turn, appear to have been closely associated with the organizational pattern of *supervisory degradation* and the technological pattern of *sensitivity smoothing*: an eagerness to increase the quantity of autonomous miles driven was reportedly associated with the company's decision to reduce the number of operators required per vehicle from two to one (Bort, 2018a, 2018d); and the focus on creating a more comfortable passenger experience that would be ready to launch as a service was reportedly associated with the implementation of action suppression functions in an attempt to reduce the frequency of uncomfortable excessive braking events. Similarly, important self-reinforcing interactions appear to have unfolded between the cultural pattern of *presumptive reliability*, which allowed the persistence of a general belief in the sufficiency of human vigilance, and the structural pattern of *vigilance dependencies*, which built a reliance on sustained human vigilance deep into the architecture of the vehicle development and testing program. In turn, the emergence of these mutually reinforcing patterns appears to have been enabled by the organizational pattern of *governance gaps*, one expression of which was a lack of sufficient operational safety expertise or oversight which meant that basic human reliability risks associated with the work of safety drivers were overlooked. Governance gaps appear to have further contributed to and interacted with the emergence and persistence of the epistemic pattern of *insensitivity to experience*, in which regularly encountered and easily anticipated hazards—such as pedestrians crossing the road outside a crossing area—were not recognized as a significant risk worthy of specific risk mitigations. A full exploration of these many sociotechnical interactions underlying the Uber AV accident is beyond the immediate scope of this article, but the SOTEC framework provides an initial conceptual architecture and language within which these complex sociotechnical interactions can be characterized and analyzed across structural, organizational, technological, epistemic, and cultural domains of sociotechnical risk.

5. DISCUSSION: MANAGING SOCIOTECHNICAL RISK AND BUILDING INFRASTRUCTURES OF LEARNING IN AUTONOMOUS AND INTELLIGENT SYSTEMS

Recent innovations have created an urgent need to better understand the sociotechnical sources of risk that can lead to catastrophic failures in autonomous and intelligent systems, in order to develop models that can inform risk analysis and underpin practical systems of risk management, governance, and learning. This analysis of Uber's fatal self-driving accident explores and defines five interrelated domains of sociotechnical risk in AIS—structural, organizational, technological, epistemic, and cultural—forming an overarching SOTEC framework that offers a complementary set of lenses through which particular patterns of sociotechnical risk can be identified and analyzed. Risk analysis methods and risk management processes applied to AIS will need to be able to accommodate this diverse spectrum of sociotechnical risk, and the particular patterns of so-

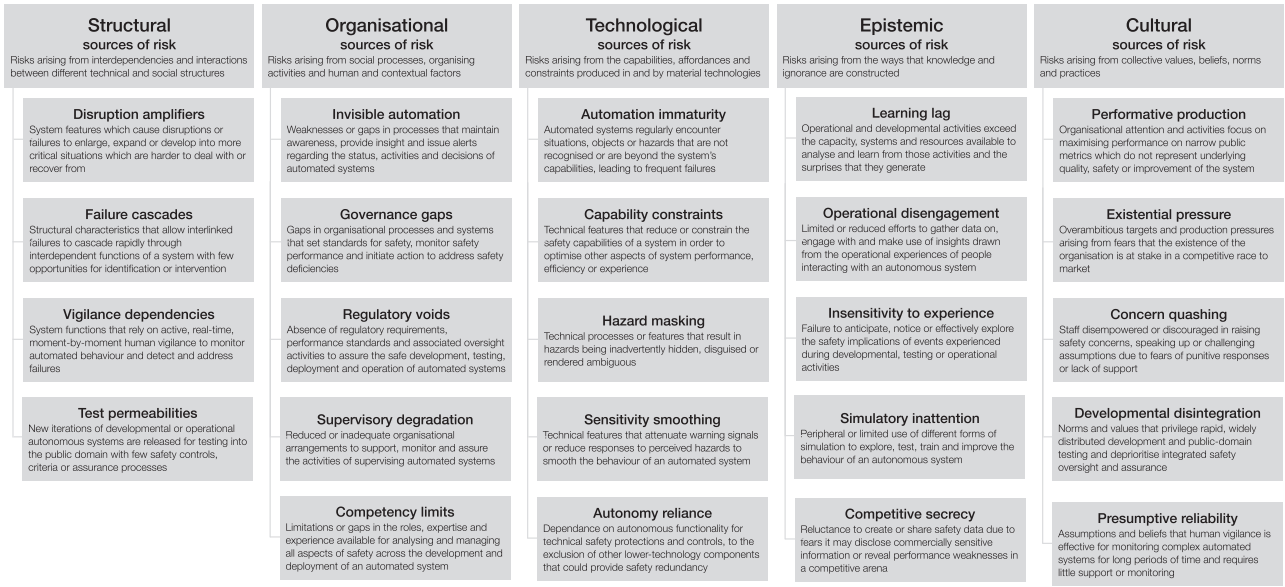

Fig 2. Indicative Patterns of Sociotechnical Risk in Autonomous and Intelligent Systems Across the Five Domains of the SOTEC Risk Framework

ciotechnical risk identified in this analysis offer an initial set of indicators that may be useful in identifying areas of emerging risk in current AIS development and deployment activities (Fig. 2).

This SOTEC framework has been built from an empirically grounded and theoretically informed analysis of the widely reported crash of a single developmental autonomous vehicle operating on public roads. Nonetheless, this is intended to provide an initial overarching and more generally applicable framework that defines the fundamental sources and patterns of sociotechnical risk that may threaten safety in a range of autonomous and intelligent systems, spanning the entire AIS lifecycle from design and development to implementation and operation. As stated earlier, this framework represents an early phase in a process of theorizing, rather than providing a full and final theory (Weick, 1995), and a range of further empirical and analytical work will be needed to examine the extent to which this initial framework and its constituent patterns can explain AIS risk and failure in contexts such as healthcare, transport, or finance. While the Uber self-driving accident is currently unique in terms of the extent and depth of public reporting on the event and the conditions that surrounded it, a number of illustrative examples indicate how aspects of the SOTEC framework may be applicable and useful in a range of different settings. Structural sources of risk such as disruption amplifiers and failure cascades may help explain the 2010 "flash crash" in US equity markets, when the rapid actions and unexpected interactions of automated trading algorithms caused a trillion-dollar stock market crash that lasted barely 36 minutes (CFTC & SEC, 2010). Organizational sources of risk such as governance gaps and regulatory voids were evident in the findings of a regulatory sandbox on the use of artificial intelligence in healthcare, particularly in the initial implementation and verification of such systems in clinical settings (CQC, 2020). Technological sources of risk such as hazard masking and automation immaturity may also be applicable to machine learning failures in healthcare, such as when a predictive model developed to make recommendations about whether to admit pneumonia patients to hospital deemed patients who also had a history of asthma to be of lower risk (when in fact the opposite is true). In this case the machine learning algorithm had correctly identified and learnt a pattern in the data—that patients with asthma and pneumonia had better outcomes than those with just pneumonia—but on close investigation this was due to localized hospital practices in which patients with asthma and pneumonia were always proactively admitted straight to intensive care to prevent risky complications (Cabitza, Rasoini, & Gensini, 2017). Epistemic sources of risk such as competitive secrecy may help to explain the reported reluctance of

autonomous vehicle developers to share detailed data across the industry (Marshall, 2018); and insensitivity to experience may help explain the repetition of a seemingly near-identical fatal accident sequence by two Autopilot-guided Teslas driving under the side of a tractor-trailer in 2016 and 2019 (NTSB, 2017, 2020b). And cultural sources of risk such as concern quashing and performative production may be evident in attempts by healthcare chatbot developers to claim equivalent performance to human clinicians and to publicly dismiss concerns raised by concerned doctors about diagnostic errors (Iacobucci, 2020). Taken together, these brief examples point to some of the ways that the SOTEC framework may be applied, refined, elaborated, and revised in future across a diverse range of contexts and the whole lifecycle of AIS development and use. More immediately, the analysis developed here points to a number of theoretical implications regarding how risk and safety are understood in AIS, as well as a set of practical implications for developing infrastructures for managing and learning from AIS failures.

5.1. Theoretical Implications for Understanding Risk and Safety in AIS

The analysis developed in this article has five key theoretical implications for understanding, analyzing, and managing risk in AIS.

5.1.1. The Development of Epistemic Capacity

First, AIS safety is contingent on organizational capacities for learning. Surprising and harmful failures occur when AIS developmental or operational activities exceed an organization's capacity to understand and learn from those activities (Downer, 2011). As illustrated by the analysis developed here, a focus on the rapid development and deployment of AIS can generate a learning lag, where operational and developmental activities outstrip the organizational systems and resources that are available to analyze and learn from those activities. These weaknesses in epistemic capacity can be amplified by tendencies toward competitive secrecy which reduce the sharing or creation of sensitive safety data, and by failures to properly value or engage with a variety of sources of knowledge such as simulation and practical insights from front-line system operators—who may have a relatively low status in an organization (Weick, Sutcliffe, & Obstfeld, 1999). Taken together, this implies that the speed and scale of AIS development and deployment should be determined by the learning capacity of the sociotechnical system that an AIS is situated within: operational activity should be scaled to epistemic capacity. Or, as one concerned Uber manager bluntly recommended a few days before the fatal accident, "do not drive the cars more than is necessary" (Efrati, 2018c). Managing risk in AIS will therefore depend on developing more sophisticated ways of assessing and managing the mechanisms that underpin social and organizational learning (Macrae, 2014a, 2014b; Waterson, 2020), just as much as it depends on developing more sophisticated mechanisms of machine learning (Stilgoe, 2018).

5.1.2. The Management of Sociotechnical Complexity

Second, AIS safety depends on the systematic management of sociotechnical complexity. Many AIS technologies are inherently complex, and the practical activities involved in developing them are too. This complexity creates ideal conditions for the emergence of unexpected failures that are hard to understand and difficult to contain (Perrow, 1999; Weick & Sutcliffe, 2001). As the analysis developed here indicates, the highly complex sociotechnical structures of AIS can act as disruption amplifiers that encourage small disruptions to enlarge rapidly, can facilitate rapid failure cascades that are hard to identify or shut down, can build in vigilance dependencies that require impossibly perfect human performance, and can create test permeabilities that allow new or updated systems to be rapidly and regularly released into the public domain. Risk management may therefore need to focus on identifying the most problematic thickets of sociotechnical complexity in AIS, and develop mechanisms to decouple, modularize, or otherwise simplify the technological and organizational sources of that complexity (Perrow, 2011)—such as separating functional stages of deep learning diagnostic processes (Fauw et al., 2018), creating institutional gatekeepers to review and approve change requests, or instituting "circuit breakers" to prevent failures cascading through a system (Macrae, 2019b; SEC, 2010). Identifying and understanding these complex sociotechnical interactions may be supported by further elaboration of the SOTEC framework, with a particular focus on exploring the interlinkages between different patterns of sociotechnical failure, such as those highlighted previously. A valuable focus for future research may therefore involve developing better

models of how different patterns of failure in AIS can reinforce and amplify one another, alongside the development of more theoretically sophisticated explanations of how sociotechnical patterns of failure become entangled in different contexts, and under what conditions these entanglements may represent higher-order or more systemic patterns of risk.

5.1.3. *The Integrated Governance of Development and Operation*

Third, AIS safety requires risk management strategies that integrate the governance of both the development and operation of AIS. Experimental technologies can pose unique risk management challenges by generating new and surprising system behaviors that are beyond the bounds of prior experience, and which therefore require continual exploration and renegotiation of acceptable limits of performance (Vaughan, 1996). AIS are likely to present particular challenges for risk management because the experimental and developmental phases of AIS can often involve prolonged periods of "real-world" operational testing, verification, and optimization to explore how systems actually perform *in situ*—which therefore combines some aspects of system development and system operation. For self-driving cars this may involve extensive phases of on-road testing in the public domain. For clinical diagnostic systems it may involve long periods of testing or verification in different clinical settings. Moreover, one of the proposed benefits of many AIS is that they can continually learn or be updated so as to adapt to changing conditions and iteratively improve performance over time. This similarly means that the processes of developing, training, and optimizing AIS can become permanently intermixed with the ongoing operation of these systems. This blending of development activities (such as training machine learning algorithms to accurately identify pedestrians) with operational activities (such as running self-driving cars for thousands of miles on public roads) creates a range of risk management challenges. In particular, this integration of developmental and operational AIS activity involves taking some sort of incomplete experimental technology (Stilgoe, 2020) that routinely produces novel, unexpected, and surprising behaviors and regularly operating it in complex, dynamic, and inherently hazardous real-world settings in close proximity to vulnerable members of the public. As illustrated by the analysis developed in this article, integrating AIS developmental and operational activities—such as intensively operating developmental self-driving cars on public roads—can create significant governance gaps, regulatory voids, and competency limits. Safety governance processes set up initially for technical experimentation may be unable to handle the increased complexity and societal risks created by large-scale operational activities in the public domain. Regulatory requirements may be inappropriate or entirely absent in these new hybrid phases of intensive AIS "developmental-operations" that are beginning to occur in various public spaces. Similarly, organizations and regulators may not have the safety expertise and oversight roles that are needed to properly engage with the broad spectrum of risks associated with the intermixing of AIS development, deployment, and operation. Accordingly, the blending of developmental and operational activities that is a characteristic of the AIS lifecycle points to the importance of developing more integrated, adaptive and reflexive approaches to the management and regulation of risk in AIS. Adaptive and reflexive systems of risk governance are based on processes that allow governance processes themselves to be regularly reexamined and adapted by learning from experience (Brass & Sowell, 2020; McCray, Oye, & Petersen, 2010). It will therefore be important to explore how hybrid and integrated models of risk governance—that span the entire spectrum of AIS development and operation—can be adaptively developed and implemented.

5.1.4. *The Creation of Cultures of Openness and Inquiry*

Fourth, AIS safety depends on people noticing, generating, discussing, interrogating, and acting on weak signals of emerging risk. Most AIS remain in the early stages of development and adoption and have not yet experienced many of the failure modes or harmful outcomes that these systems have the potential to produce. As such, neither collective memory nor formal models are likely to capture the full array of warning signs that may signal the development of catastrophic failure (Schulman, 1993). It is therefore important to understand how organizational cultures can be built that encourage people to speak up and challenge assumptions, enable open and collective inquiry into early signals of risk, and develop well-specified fears of specific forms of failure (Macrae, 2014b; Weick & Sutcliffe, 2001). As the analysis developed here indicates, it will be particularly important to understand what these cultures

might look like in highly competitive and technology-oriented AIS organizations, in which developmental disintegration may emerge through the cultural prioritization of rapid development and deployment at the expense of carefully integrated safety analysis; and where there may be broad-based presumptions of human reliability that preclude more critical exploration and management of the risks of human–machine interaction. This analysis also points to the importance of better understanding the precursors and consequences of cultures in which existential pressures and fears of corporate survival overtake other concerns, in which patterns of concern quashing limit the ability of individuals to speak up, and in which organizational attention and effort becomes focused on performative production to meet external metrics of progress.

5.1.5. *The Cultural Construction of Technological Systems*

Fifth, AIS safety is shaped by the ways that cultural assumptions and values become inscribed into technological objects, which can in turn determine fundamental system capabilities, constraints, and affordances. The cultural characteristics of the organizations that develop and implement AIS cannot be neatly separated from the technological characteristics of the resulting systems (Leonardi & Barley, 2010). This is starkly illustrated by the Uber accident, where an organizational culture that appears to have systematically overlooked signals of emerging danger produced a self-driving system that was itself tuned to discount warning signs through sensitivity smoothing, had capability constraints which minimized the initiation of emergency responses, and incorporated design features that resulted in hazard masking of vulnerable road users. Understanding the safety of autonomous and intelligent systems will therefore depend on building more sophisticated theories of how technological affordances, cultural values, and assumptions about risk critically shape the evolution, design, functionality, and performance of AIS (Crawford & Calo, 2016; Leonardi & Barley, 2008).

5.2. Practical Implications for Governing Risk and Learning from Failure in AIS

This analysis offers a range of practical implications for managing risk and learning from AIS failure. These can be framed as seven interdependent principles that define key functional building blocks of a learning infrastructure that support the governance of safety and the management of risk across the lifecycle of AIS and that are relevant both within individual organizations and across entire industries (Fig. 3).

5.2.1. *System Transparency*

Transparency is a core principle in AIS safety (Jobin et al., 2019; Winfield & Jirotka, 2018): it must be possible to understand what an AIS is doing and why, both to safely interact with and supervise these technologies (Sarter et al., 1997; Wortham, Theodorou, & Bryson, 2017) and to retrospectively investigate failures (Bryson & Winfield, 2017; Winfield et al., 2021). Prior discussions of transparency have focused on the importance of making the working of intelligent technologies transparent, interpretable, and explainable but the analysis developed here extends this principle, emphasizing the importance of building mechanisms that can help to render entire sociotechnical systems transparent—encompassing core technologies as well as the human activities and organizational processes that surround them. For instance, organizational processes and technological methods need to be incorporated into AIS that can identify pockets of invisible automation, highlight the potential for failure cascades, flag the possibility of hazard masking, reveal areas of supervisory degradation and acknowledge sources of existential pressure. As such, transparency is not purely a technical requirement but a sociotechnical one: to manage risk and learn from failure, the network of organizational decisions, cultural values, and human interactions that AIS are embedded within must also be made legible and open to scrutiny (Kroll, 2018).

5.2.2. *Event Recording*

To enable learning, AIS need robust mechanisms to record information about failure events and capture contemporaneous data about the functioning—and malfunctioning—of systems. A core component of this should be event recorders—like the "black boxes" used in the aviation industry—that capture rich, real-time information about AIS processes before and during accidents (Murphy & Woods, 2009; Winfield & Jirotka, 2017). The analysis here also emphasizes the importance of expanding the focus and mechanisms of event recording in AIS to capture more minor safety disruptions from a

Fig 3. Core components of an infrastructure for governing safety, managing risk and learning from autonomous and intelligent system failures

diverse range of sources—such as routine operational monitoring systems that identify and record deviations from predetermined safety standards, equivalent to the continuous safety monitoring programs used in airlines (FAA, 2004; O'Leary, Macrae, & Pidgeon, 2002), as well as nonpunitive incident reporting systems for professionals and the public to report safety events and "near-miss" incidents (Macrae, 2014a, 2016; McGregor, 2020). These sorts of event recording mechanisms can help to reveal the preconditions of sociotechnical failure, and form the foundational safety data infrastructure that is needed to underpin the epistemic processes of risk management and avoid the emergence of learning lag.

5.2.3. Data Access

Investigating and learning from AIS failures requires safety-relevant data—such as operational records, technical designs, and user experiences—to not only be collected but also to be readily accessible and shared, both within individual organizations and across entire industries. The fundamental principle that safety data should be easily accessible and widely usable by a range of different parties and actors depends on the creation of both a technical and social infrastructure for data sharing. Data access partly depends on developing technical standards and mechanisms to define, record, and share safety data (McGregor, 2020; NHTSA, 2017; NTSB, 2017) to help counterbalance tendencies toward competitive secrecy that this analysis indicates can emerge in AIS. The analysis developed here also highlights the importance of social and cultural processes of trust, fear, and blame—particularly, for example, as they emerge in processes of concern quashing. Fear of punishment can seriously impede the sharing of safety data—both by individuals and organizations (Dekker, 2016; Macrae, 2016; Reason, 1997). To address this, safety-critical industries such as aviation strictly separate processes that determine liability from those of learning: data shared to support safety improvement cannot be used for punitive purposes (EU, 2010; Michaelides-Mateou & Mateou, 2010). Similar social agreements and legal frameworks will be needed to support learning from AIS failures.

5.2.4. Risk Professionals

The work of managing risk and learning from failure depends on a range of specialist skills and expertise, and needs to be led by professionals

occupying dedicated, impartial, and authoritative roles. This analysis indicates that significant problems can arise when competency limits preclude the effective analysis and management of safety across developmental and operational phases of AIS. This, in turn, highlights that managing AIS safety will require a new cadre of multidisciplinary risk professionals with the expertise and authority to manage a broad array of sociotechnical risks. Dedicated safety teams and leadership roles (e.g., Shepardson, 2020) will need to be created with the power and remit to pause organizational activities, investigate and analyze systems, challenge design and operational decisions and meaningfully oversee the entire lifecycle of AIS systems (Lehr & Ohm, 2017). This professional community will need to draw on models, methods, and knowledge beyond the engineering disciplines (Cummings, 2017; Steinhardt, 2015) to encompass human factors, social psychology, organizational sociology, and a broad range of the safety sciences (Klinke, Renn, & Goble, 2021; Salmon et al., 2020; Stanton et al., 2019; Waterson, 2019).

5.2.5. Systemic Investigation

Learning from AIS failures depends on rigorously investigating and addressing the systemic factors that contribute to safety incidents. The analysis developed here highlights the importance of building investigative capacity both within individual organizations—as a key mechanism to mitigate and avoid learning lag in the development of particular AIS—and across entire sectors so that AIS incidents can be routinely and impartially investigated to improve safety across the entire field (BSI, 2021; Winfield & Jirotka, 2017; Winfield et al., 2021), just as they are in sectors like aviation (Macrae, 2014) and healthcare (Macrae & Vincent, 2014). AIS safety investigations will need to be led by expert investigators, conducted solely for the purposes of learning and draw on investigative methods that allow the entire sociotechnical system to be examined—from the technical design of machine learning models to the social characteristics of organizational cultures (Stanton et al., 2019; Waterson, Jenkins, Salmon, & Underwood, 2017), and spanning the full array of sociotechnical domains and processes that this analysis indicates can be involved in the failure of AIS. Critically, just as in existing safety-critical sectors, AIS incident investigations will need to be protected from production pressures and kept entirely independent from punitive processes that seek to allocate blame or liability, objectives which can undermine attempts to make systemic improvements (Civil Aviation Regulations, 1996; Macrae & Vincent, 2017a). If these investigative principles are not embedded within AIS safety investigations then the potential for learning is likely to be limited, because the processes, insights, and outputs of an investigation may be shaped by competitive secrecy, existential pressures, and competency limits that have been explored in this analysis.

5.2.6. Safety Governance

Risk management in AIS requires a systematic approach to overseeing, monitoring, and governing safety. The implications of the governance gaps, regulatory voids, and developmental disintegration identified in this analysis reinforces the importance of establishing Safety Management Systems (CAA, 2014; Hart et al., 2018; NTSB, 2019a) and formal safety governance and regulatory infrastructures for AIS. These should be organized around clear safety objectives, including precise definitions of the outcomes that organizations are seeking to avoid (Macrae, 2014b; Schulman, 1993) and the safety criteria applicable to different AIS activities (Cummings, 2021), and should enable organizations to build integrated pictures of AIS risk that encompass the full range of sociotechnical risks and span from design to development to deployment. Crucially, safety management systems and wider governance systems should provide an infrastructure to proactively engage with and act on the safety concerns of all stakeholders impacted by AIS, both inside and outside the organization (Klinke & Renn, 2021).

5.2.7. Learning Cultures

Learning from failure is only possible when people are able to routinely highlight and openly discuss potential safety problems without fear of being penalized (Edmondson, 2018; Reason, 2000). This analysis emphasizes the need to create organizational cultures around AIS in which people are supported to speak up with safety concerns and proactively act on safety issues—and are empowered to rapidly act on urgent problems (Weick & Sutcliffe, 2001). Patterns such as concern quashing and operational disengagement characterized in this analysis particularly point to the importance of building a "just culture" around AIS safety (Dekker, 2016; Reason, 1997), in which people are able to share safety concerns and

information secure in the knowledge they will not be unfairly blamed or punished. Building a just culture will involve establishing norms, policies, rules, and structures that entirely separate safety improvement responsibilities from processes that seek to apportion liability or blame (Braithwaite, 2011; EU, 2010; Macrae & Vincent, 2017b).

6. CONCLUSION

The fatal 2018 Uber self-driving crash represents a watershed moment in AIS safety. Reanalyzing this event reveals the inherently sociotechnical nature of risk in AIS and allows the development of a theoretically informed and empirically grounded overarching framework that characterizes key sociotechnical sources and patterns of risk in AIS. This analysis highlights a set of theoretical challenges that will need to be addressed as AIS become more widespread, and indicates the importance of building infrastructures of learning that can support the analysis, management, and governance of risk in AIS. Perhaps most fundamentally, this analysis stresses the need for the analysis and management of risk in AIS to be approached as a fundamentally sociotechnical problem that encompasses the full range of human, social, cultural, and organizational entanglements that technologies of autonomy and intelligence necessarily emerge from and are embedded within.

Acknowledgments

I wish to thank John Downer and the anonymous reviewers for insightful comments on earlier versions of this article. This research was funded in whole, or in part, by the Wellcome Trust (Grant number /213632/Z/18/Z/). For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

References

Abdi, J., Al-Hindawi, A., Ng, T., & Vizcaychipi, M. P. (2018). Scoping review on the use of socially assistive robot technology in elderly care. *BMJ Open*, *8*, e018815.

Ackerman, E. (2019, November 19). My fight with a sidewalk robot. *Bloomberg CityLab*. Retrieved from https://www.bloomberg.com/news/articles/2019-11-19/why-tech-needs-more-designers-with-disabilities

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mane, D. (2016). Concrete problems in AI safety. arXiv:1606.06565v2 [cs.AI]

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., … Shetty, S. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*, *25*, 954–961.

Asimov, I. (1990). *Robot visions*. New York: Roc Books.

Babic, B., Gerke, S., Evgeniou, T., & Cohen, I. G. (2021). Direct-to-consumer medical machine learning and artificial intelligence applications. *Nature Machine Intelligence*, *3*, 283–287.

Bainbridge, L. (1983). Ironies of automation. *Automatica*, *19*(6), 775–779.

Barley, S. R. (2020). *Work and technological change*. Oxford, U: Oxford University Press.

Beane, M., & Orlikowski, W. J. (2015). What difference does a robot make? The material enactment of distributed coordination. *Organization Science*, *26*(6), 1553–1573.

Bort, J. (2018a, November 19). Uber insiders describe infighting and questionable decisions before its self-driving car killed a pedestrian. *Insider*. Retrieved from https://www.businessinsider.com/sources-describe-questionable-decisions-and-dysfunction-inside-ubers-self-driving-unit-before-one-of-its-cars-killed-a-pedestrian-2018-10?r=US%26IR=T

Bort, J. (2018b, November 28). 'We have screwed up': Uber CEO Dara Khosrowshahi says in an all-hands meeting that the company deserves some fault after its self-driving car killed a pedestrian. *Insider*. Retrieved from https://www.businessinsider.com/uber-ceo-tells-employees-self-driving-car-unit-screwed-up-2018-11?r=US%26IR=T

Bort, J. (2018c, December 21). Uber employees working on self-driving cars feel their cars are safer but their careers are stuck, according to leaked employee survey. *Insider*. Retrieved from https://www.businessinsider.com/uber-self-driving-car-unit-leaked-employee-survey-2018-12?r=US%26IR=T

Bort, J. (2018d, December 21). Uber employees describe a stressful and 'ridiculous' culture at the self-driving-car unit under its current leader, Eric Meyhofer. *Insider*. Retrieved from https://www.businessinsider.com/uber-atg-employees-describe-a-stressful-culture-2018-12

Bort, J. (2019, April 20). An engineer at Uber's self-driving car unit warns that it's more like 'a science experiment' than a real car capable of driving itself. *Insider*. Retrieved from https://www.businessinsider.com/uber-self-driving-car-is-like-science-experiment-insider-says-2019-4

Bradshaw, J. M., Bradshaw, J. M., Hoffman, R. R., & Woods, D. D. (2013). The seven deadly myths of "autonomous systems." *IEEE Intelligent Systems*, *28*(3), 54–61.

Braithwaite, J. (2011). The essence of responsive regulation. *UBC Law Review*, *44*(3), 475–520.

Brass, I., & Sowell, J. H. (2020). Adaptive governance for the Internet of Things: Coping with emerging security risks. *Regulation & Governance*. https://doi.org/10.1111/rego.12343

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., … Anderljung, A. (2020). Toward trustworthy AI development: Mechanisms for supporting verifiable claims. arXiv:2004.07213v2 [cs.CY]

Bryson, J., & Winfield, A. (2017). Standardizing ethical design for artificial intelligence and autonomous systems. *Computer*, *50*(5), 116–119.

BSI. (2021). *Data collection and management for automated vehicle trials for the purpose of incident investigation–Specification*. PAS1882:2021. London, UK: British Standards Institution.

Burrell, J. (2016). How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 1–12.

CAA. (2014). *Safety management systems (SMS): Guidance to organisations; CAP795*. London, UK: Civil Aviation Authority.

Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. *JAMA: The*

Journal of the American Medical Association, *318*, 517–518.
Carroll, J. M. (2003). Toward a multidsciplinary science of human-computer interaction. In J. M. Carroll (Ed.), *HCI models, theories and frameworks: Toward a multidisciplinary science* (pp. 1–10). London, UK: Morgan Kaufmann.
Carson, B. (2016, August 18). Travis Kalanick on Uber's bet on self-driving cars: 'I can't be wrong'. *Insider*. Retrieved from https://www.businessinsider.com/travis-kalanick-interview-on-self-driving-cars-future-driver-jobs-2016-8?r=US%26IR=T
CFTC and SEC. (2010). Findings regarding the market events of May 6, 2010. Report of the staffs of the CFTC and SEC to the joint advisory committee on emerging regulatory issues. Washington, DC: US Commodities and Futures Trading Commission and US Securities and Exchange Commission.
Chan, C., Chow, C., Wong, J., Dimakis, N., Nayler, D., Bermudes, J., … Baker, M. (2019). *Artificial intelligence applications in financial services: Asset management, banking and insurance*. London, UK: Marsh and McLennon Companies.
Civil Aviation Regulations. (1996). *The civil aviation (investigation of air accidents and incidents) regulations 1996*. Retrieved from http://www.legislation.gov.uk/uksi/1996/2798/contents/made
Collingridge, D. (1996). Resilience, flexibility, and diversity in managing the risks of technologies. In C. Hood & D. K. C., Jones (Eds.), *Accident and design: Contemporary debates on risk management* (pp. 40–45). London, UK: Taylor and Francis.
Costas, J., & Grey, C. (2016). *Secrecy at work: The hidden architecture of organizational life*. Palo Alto, CA: Stanford University Press.
CQC. (2020). *Using machine learning in diagnostic services: CQC's regulatory sandbox report*. London, UK: Care Quality Commission.
Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature*, *538*(7625), 311–313.
Cummings, M. (2017). The brave new world of driverless cars. *Transportation Research Board News*, *308*, 34–37.
Cummings, M. (2021). Rethinking the maturity of artificial intelligence in safety-critical settings. *AI Magazine*, *42*(1), 6–15.
Davies, A. (2021). *Driven: the race to create the autonomous car*. London, UK: Simon and Schuster.
Decker, S. (2016). *Just culture: Restoring trust and accountability in your organisation*. Abingdon, UK: CRC Press.
Dekker, S. (2011). *Drift into failure: From hunting broken components to understanding complex systems*. London, UK: Routledge.
Dixon, L. (2020). Autonowashing: The greenwashing of vehicle automation. *Transportation Research Interdisciplinary Perspectives*, *5*, 100113.
Downer, J. (2009). When failure is an option: Redundancy, reliability and regulation in complex technical systems. London School of Economics and Political Science Centre for Analysis of Risk and Regulation Discussion Paper 53. Retrieved from http://eprints.lse.ac.uk/36537/1/Disspaper53.pdf
Downer, J. (2011). "737-Cabriolet": The limits of knowledge and the sociology of inevitable failure. *American Journal of Sociology*, *117*(3), 725–762.
Downer, J. (2019). On ignorance and apocalypse: A brief introduction to 'epistemic accidents'. In J. Le Coze (Ed.), *Safety science research: Evolution, challenges and new directions* (pp. 75–85). London, UK: CRC Press.
Edmondson, A. C. (2018). *The fearless organisation: Creating psychologiical safety in the worksplace for learning, innovation and growth*. London, UK: Wiley.
Efrati, A. (2018a, May 7). Uber finds deadly accident likely caused by software set to ignore objects on road. *The Information*. Retrieved from https://www.theinformation.com/articles/uber-finds-deadly-accident-likely-caused-by-software-set-to-ignore-objects-on-road
Efrati, A. (2018b, June 19). Uber neglected simulation testing on self-driving cars, insiders say. *The Information*. Retrieved from https://www.theinformation.com/articles/uber-neglected-simulation-testing-on-self-driving-cars-insiders-say
Efrati, A. (2018c, December 10). How an Uber whistleblower tried to stop self-driving car disaster. *The Information*. Retrieved from https://www.theinformation.com/articles/how-an-uber-whistleblower-tried-to-stop-self-driving-car-disaster
Efrati, A. (2019a, June 7). How Uber wants Self driving cars to help ride-hailing business. *The Information*. Retrieved from https://www.theinformation.com/articles/how-uber-wants-self-driving-cars-to-help-ride-hailing-business
Efrati, A. (2019b, December 9). Uber nears deal for self-driving car simulation startup. *The Information*. Retrieved from https://www.theinformation.com/articles/uber-nears-deal-for-self-driving-car-simulation-startup
Efrati, A. (2020, September 28). infighting, 'busywork,' missed warnings: How uber wasted $2.5 billion on self-driving cars. *The Information*. Retrieved from https://www.theinformation.com/articles/infighting-busywork-missed-warnings-how-uber-wasted-2-5-billion-on-self-driving-cars
Eisenhardt, K. (1989). Building theories from case study research. *The Academy of Management Review*, *14*(4), 532–550.
Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society*, *5*, 40–60.
EU. (2010). Regulation No 996/2010 of the European Parliament and of the Council of 20 October 2010 on the investigation and prevention of accidents and incidents in civil aviation and repealing Directive 94/56/EC. *Official Journal of the European Union*, 12.11.2010.
Evans, R., & Gao, J. (2016, July 20). DeepMind AI reduces Google data centre cooling bill by 40%. *DeepMind Blog*. Retrieved from https://deepmind.com/blog/article/deepmind-ai-reduces-google-data-centre-cooling-bill-40
FAA. (2004). *Flight operational quality assurance: Advisory circular 120–82*. Washington, DC: Federal Aviation Administration.
Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., … Ronneberger, O. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, *24*(9), 1342–1350.
FCA. (2019). *Machine learning in UK financial services*. London, UK: Financial Conduct Authority and Bank of England.
Ford, M. (2018). *Architects of intelligence: The truth about AI from the people building it*. Birmingham, UK: Packt Publishing.
Fraser, H., Coiera, E., & Wong, D. (2018). Safety of patient-facing digital symptom checkers. *Lancet*, *392*(10161), 2263–2264.
Gephart, R. P. (2004). Normal risk: Technology, sense making, and environmental disasters. *Organization and Environment*, *17*, 20–26.
Glaser, B. G. (1978). *Theoretical sensitivity*. Mill Valley: Sociology Press.
Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory*. Chicago, IL: Aldine.
Hart, C., Dombroff, M. A., & Tochen, D. K. (2018). Independent review of the safety culture of Uber Technologies, Inc's Advanced Technologies Group. Final report. Richmond, Virginia: LeClairRyan PLLC.
Hancock, P. A. (2017). Imposing limits on autonomous systems. *Ergonomics*, *60*, 284–291.
Harris, M. (2018, March 28). Exclusive: Arizona governor and Uber kept self-driving program secret, emails reveal. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2018/mar/28/uber-arizona-secret-self-driving-program-governor-doug-ducey
Hawkins, R., Paterson, C., Picardi, C., Jia, Y., Calinescu, R., & Habli, I. (2021). *Guidance on the assurance of machine learning*

in autonomous systems (AMLAS). York, UK: Assuring Autonomy International Programme.
Heinla, A. (2021, January 27). Starship completes one million autonomous deliveries. *Starship Technologies Blog*. Retrieved from https://medium.com/starshiptechnologies/one-million-autonomous-deliveries-milestone-65fe56a41e4c
Hollnagel, E. (2004). *Barriers and accident prevention: Or how to improve safety by understanding the nature of accidents*. Aldershot, UK: Ashgate.
Hollnagel, E., Paries, J., Woods, D., & Wreathall, J. (2012). *Resilience engineering in practice: A guidebook*. Aldershot, UK: Ashgate.
Hopkins, A. (1999a). *Managing major hazards: The lessons of the Moura mine disaster*. St Leonards: Allen and Unwin.
Hopkins, A. (1999b). The limits of normal accident theory. *Safety Science*, *32*, 93–102.
Hopkins, A. (2005). *Safety, culture and risk: The organisational causes of disasters*. Sydney, Australia: CCH.
Hulme, A., Stanton, N. A., Walker, G. H., Waterson, P., & Salmon, P. M. (2021). Complexity theory in accident causation: Using AcciMap to identify the systems thinking tenets in 11 catastrophes, *Ergonomics*, *64*(7), 821–838. https://doi.org/10.1080/00140139.2020.1869321
Iacobucci, G. (2020). Row over Babylon's chatbot shows lack of regulation. *British Medical Journal*, *368*, m815.
IEEE. (2019). *Ethically aligned design: A vision for prioritising human well-being with autonomous and intelligent systems*. New York: IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems.
Jaderberg, M., Czarnecki, W. M., Dunning, I., Marris, L., Lever, G., Castaneda, A. G., … Graepel, T. (2019). Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science*, *364*(6443), 859–865.
Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*, 389–399.
Kletz, T. (1994). *Learning from accidents*. Oxford, UK: Butterworth-Heinneman.
Klinke, A., & Renn, O. (2021). The coming of age of risk governance. *Risk Analysis*, *41*(3), 544–557.
Klinke, A., Renn, O., & Goble, R. (2021). Prologue: The "brave new world" of social sciences in interdisciplinary risk research. *Risk Analysis*, *41*(3), 407–413.
Krafcik, J. (2020, October 8). Waymo is opening its fully driverless service to the general public in Phoenix. *Waypoint: The official Waymo blog*. Retrieved from https://blog.waymo.com/2020/10/waymo-is-opening-its-fully-driverless.html
Kroll, J. A. (2018). The fallacy of inscrutability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133), 20180084.
Lehr, D., & Ohm, P. (2017). Playing with the data: What legal scholars should learn about machine learning. *UC Davis Law Review*, *51*, 653–717.
Leonardi, P. M., & Barley, S. R. (2008). Materiality and change: Challenges to building better theory about technology and organizing. *Information and Organization*, *18*(3), 159–176.
Leonardi, P. M., & Barley, S. R. (2010). What's under construction here? Social action, materiality, and power in constructivist studies of technology and organizing. *The Academy of Management Annals*, *4*(1), 1–51.
Leveson, N. (2004). A new accident model for engineering safer systems. *Safety Science*, *42*, 237–270.
Leveson, N. (2011). *Engineering a safer world: Systems thinking applied to safety*. Cambridge, MA: MIT Press.
Leveson, N. G., & Turner, C. S. (1993). An investigation of the Therac-25 Accidents. *Computer*, *26*(7), 18–41.
Levin, S. (2016a, December 15). Uber blames humans for self-driving car traffic offenses as California orders halt. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2016/dec/14/uber-self-driving-cars-run-red-lights-san-francisco
Levin, S. (2016b, December 16). Self-driving cars: Uber's open defiance of California shines light on brazen tactics. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2016/dec/16/uber-self-driving-cars-california-illegal-unethical-tactics
Lintern, S. (2021, March 6). Regulator has concerns over symptom checker app. *The Independent*. Retrieved from https://www.independent.co.uk/news/health/nhs-symptom-checker-app-safety-complaints-b1813142.html
Locke, K. (2001). *Grounded theory in management research*. London, UK: SAGE.
Macrae, C. (2009). Making risks visible: Identifying and interpreting threats to airline flight safety, *Journal of Occupational and Organisational Psychology*, *82*, 273–293.
Macrae, C. (2014a). *Close calls: managing risk and resilience in airline flight safety*. London, UK: Palgrave Macmillan.
Macrae, C. (2014b). Early warnings, weak signals, and learning from healthcare disasters. *BMJ Quality and Safety*, *23*, 440–445.
Macrae, C. (2016). The Problem with Incident Reporting. *BMJ Quality and Safety*, *25*(2), 71–75.
Macrae, C. (2019a). Governing the safety of artificial intelligence in healthcare. *BMJ Quality and Safety*, *28*, 495–498.
Macrae, C. (2019b). Moments of resilience: Space, time and the organisation of safety in complex systems. In S. Wiig & B. Fahlbruch (Eds.), *Exploring resilience: A scientific journey from practice to theory* (pp. 15–23). London, UK: Springer.
Macrae, C., & Vincent, C. (2014). Learning from failure: The need for independent safety investigation in healthcare. *Journal of the Royal Society of Medicine*, *107*, 439–443.
Macrae, C., & Vincent, C. (2017a). *Investigating for improvement: Building a national safety investigator for healthcare*. London, UK: Clinical Human Factors Group. Retrieved from https://chfg.org/investigating-for-improvement-building-a-national-safety-investigator-for-healthcare/
Macrae, C., & Vincent, C. (2017b). A new national safety investigator for healthcare: The road ahead. *Journal of the Royal Society of Medicine*, *110*(3), 90–92.
Marshall, A. (2018, October 29). We've been talking about self-driving car safety all wrong. *Wired*. Retrieved from https://www.wired.com/story/self-driving-cars-safety-metrics-miles-disengagements/
McCray, L. E., Oye, K. A., & Petersen, A. C. (2010). Planned adaptation in risk regulation: An initial survey of US environmental, health, and safety regulation. *Technological Forecasting and Social Change*, *77*, 951–959.
McGoey, L. (2019). *The unknowers: How strategic ignorance rules the world*. London, UK: ZED Books.
McGregor, S. (2020). Preventing repeated real world AI failures by cataloging incidents: The AI incident database. arXiv:2011.08512v1 [cs.CY]
Michaelides-Mateou, S., & Mateou, A. (2010). *Flying in the face of criminalization: The safety implications of prosecuting aviation professionals for accidents*. Aldershot, UK: Ashgate.
Mider, Z. (2019, October 14). Tesla's Autopilot could save the lives of millions, but it will kill some people first. *Bloomberg Businessweek*. Retrieved from https://www.bloomberg.com/news/features/2019-10-09/tesla-s-autopilot-could-save-the-lives-of-millions-but-it-will-kill-some-people-first
Morley, J., Machado, C. C. V., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2020). The ethics of AI in health care: A mapping review. *Social Science and Medicine*, *260*, 113–172.
Morrison, E. W., & Milliken, F. J. (2000). Organizational silence: A barrier to change and development in a pluralistic world. *Academy of Management Review*, *25*, 706–725.

Mouratidis, K., & Serrano, V. C. (2021). Autonomous buses: Intentions to use, passenger experiences, and suggestions for improvement. *Transportation Research Part F: Traffic Psychology and Behaviour*, *76*, 321–335.

Murphy, R., & Woods, D. D. (2009). Beyond Asimov: The three laws of responsible robotics. *Intelligent Systems*, *24*(4), 14–20.

Musk, E. (2017). Elon Musk, National Governors Association, July 15 2017. Available online at: Retrieved from https://www.youtube.com/watch?v=b3lzEQANdHk%26t=1542s (quoted at 25min 15sec)

Nguyen, N. (2017, December 12). My not too fast, not too furious ride in a self-driving Uber. *Buzzfeed News*. Retrieved from https://www.buzzfeednews.com/article/nicolenguyen/not-too-fast-not-too-furious-self-driving-uber-atg

NHTSA. (2017). *Automated driving systems: A vision for safety*. Washington, DC: National Highway Traffic Safety Administration.

Niedermeyer, E. (2019, March 18). Ten lessons from Uber's fatal self-driving car crash. *The Drive*. Retrieved from https://www.thedrive.com/tech/27023/10-lessons-from-ubers-fatal-self-driving-car-crash

Norman, D. (2013). *The design of everyday things: Revised and expanded edition*. London, UK: Basic Books.

NTSB. (2017). *Collision between a car operating with automated vehicle control systems and a tractor-semitrailer truck near Williston, Florida, May 7 2016*. Washington, DC: National Transportation Safety Board.

NTSB. (2018). *Preliminary report: Highway HWY18MH010*. Washington, DC: National Transportation Safety Board.

NTSB. (2019a). *Collision between vehicle controlled by developmental automated driving system and edestrian, Tempe, Arizona, March 18, 2018: Accident report NTSB/HAR-19/03*. Washington, DC: National Transportation Safety Board.

NTSB. (2019b). *Vehicle automation report, Tempe, AZ, HWY18MH010*. Washington, DC: National Transportation Safety Board.

NTSB. (2019c). *Highway factors group chairman's factual report, HWY18MH010*. Washington, DC: National Transportation Safety Board.

NTSB. (2019d). *Vehicle factors group chairman's factual report, HWY18MH010*. Washington, DC: National Transportation Safety Board.

NTSB. (2019e). *Operations factors group chairman's factual report, HWY18MH010*. Washington, DC: National Transportation Safety Board.

NTSB. (2019f). *human performance group chairman's factual report, HWY18MH010*. Washington, DC: National Transportation Safety Board.

NTSB. (2019g). *Onboard image and data recorder group chairman's factual report, HWY18MH010*. Washington, DC: National Transportation Safety Board.

NTSB. (2019h). *Uber ATG party submission to the national transportation safety board, HWY18MH010*. San Francisco, CA: Uber Advanced Technologies Group.

NTSB. (2019i). *Volvo cars submission to the national transportation safety board, HWY18MH010*. Gothenburg, Sweden: Volvo Cars.

NTSB. (2019j). *Volvo XC90 testing by Thatcham research: ADAS 18-15 XC90 Thatcham Euro NCAP EPT/EBT Report*. Thatcham, Berkshire: Thatcham Research.

NTSB. (2019k). *National transportation safety board public meeting of November 19, 2019 – Abstract*. Washington, DC: National Transportation Safety Board.

NTSB. (2019l). *Opening statement – crash involving a pedestrian and an Uber test vehicle, by Sumwalt, R. L. National Transportation Safety Board Public Meeting of November 19, 2019*. Washington, DC: National Transportation Safety Board.

NTSB. (2019m). *Crash Overview, by Pereira, D. National Transportation Safety Board Public Meeting of November 19, 2019*. Washington, DC: National Transportation Safety Board.

NTSB. (2019n). *Pedestrian and Vehicle Operator, by Marshall, R. National Transportation Safety Board Public Meeting of November 19, 2019*. Washington, DC: National Transportation Safety Board.

NTSB. (2019o). *Managing Risk of ADS Testing, by Becic, E. National Transportation Safety Board Public Meeting of November 19, 2019*. Washington, DC: National Transportation Safety Board.

NTSB. (2019p). *Uber ATG Operations, by Fox, M. National Transportation Safety Board Public Meeting of November 19, 2019*. Washington, DC: National Transportation Safety Board.

NTSB. (2019q). *Testing of Automated Vehicles, by Becic, E. National Transportation Safety Board Public Meeting of November 19, 2019*. Washington, DC: National Transportation Safety Board.

NTSB. (2020a). *Collision between a sport utility vehicle operating with partial driving automation and a crash attenuator, Mountain View, California, March 23 2018*. Washington, DC: National Transportation Safety Board.

NTSB. (2020b). *Collision between car operating with partial driving automation and truck-tractor semitrailer Delray Beach, Florida, March 1, 2019*. Washington, DC: National Transportation Safety Board.

NTSB. (2021). *National transportation safety board accident docket management system*. Retrieved from https://data.ntsb.gov/Docket/Forms/searchdocket

O'Leary, M., Macrae, C., & Pidgeon, N. (2002). Safety data collection in British Airways flight operations. In C. Johnson, (Ed.), *Proceedings of a workshop on the investigation and reporting of incidents and accidents* (pp. 89–98). Glasgow, UK: University of Glasgow.

Parasuraman, R., & Manzey, D.H. (2010). Complacency and bias in human use of automation: an attentional integration. *Human Factors*, *52*(3), 381–410.

Perrow, C. (1984). *Normal accidents: Living with high-risk technologies*. New York: Basic Books.

Perrow, C. (1999). *Normal accidents: Living with high-risk technologies* (2nd ed.). Princeton, NJ: Princeton University Press.

Perrow, C. (2011). *The next catastrophe: Reducing our vulnerabilities to natural, industrial and terrorist disasters*. Princeton, NJ: Princeton University Press.

Pettersen Gould, K. (2021). Organizational risk: "Muddling through" 40 years of research. *Risk Analysis*, *41*(3), 456–465.

Pidgeon, N., & O'Leary, M. (2000). Man-made disasters: Why technology and organizations (sometimes) fail. *Safety Science*, *34*, 15–30.

Pöllänen, E., Read, G. J. M., Lane, B. R., Thompson, J., & Salmon, P. M. (2020). Who is to blame for crashes involving autonomous vehicles? Exploring blame attribution across the road transport system, *Ergonomics*, *63*(5), 525–537.

Randazzo, R. (2019, March 19). Who was really at fault in fatal Uber crash? Here's the whole story. *AZ Central*. Retrieved from https://eu.azcentral.com/story/news/local/tempe/2019/03/17/one-year-after-self-driving-uber-rafaela-vasquez-behind-wheel-crash-death-elaine-herzberg-tempe/1296676002/

Rasmussen, J. (1990). Human error and the problem of causality in analysis of accidents. *Philosophical Transactions of the Royal Society of London B*, *327*, 449–462.

Rasmussen, J. (1997). Risk management in a dynamic society: A modelling problem. *Safety Science*, *27*(2/3), 183–213.

Reason, J. (1990). *Human error*. Cambridge, UK: Cambridge University Press.

Reason, J. (1997). *Managing the risks of organizational accidents*. Aldershot, UK: Ashgate.

Reason, J. (2000). Safety paradoxes and safety culture. *Injury Control and Safety Promotion*, *7*(1), 3–14.
Reason, J., Hollnagel, E., & Paries, J. (2006). *Revisiting the Swiss cheese model of accidents*. Brussels, Belgium: Eurocontrol.
Roe, E., & Schulman, P. (2008). *High reliability management: Operating on the edge*. Palo Alto, CA: Stanford University Press.
Ross, A. (2018, May 15). Disrupting healthcare: Interview with Dr Ali Parsa CEO of Babylon health. *Information Age*. Retrieved from https://www.information-age.com/dr-ali-parsa-ceo-babylon-healthcare-123472484/
Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. London, UK: Allen Lane.
Russell, S., & Norvig, P. (2016). *Artificial intelligence: A modern approach* (3rd ed.). Harlow, UK: Pearson.
Sagan, S. D. (1995). *The limits of safety: Organizations, accidents, and nuclear weapons*. London, UK: Princeton University Press.
Salmon, P. M., Carden, T., & Hancock, P. A. (2020). Putting the humanity into inhuman systems: How human factors and ergonomics can be used to manage the risks associated with artificial general intelligence. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *17*(4), 1–14.
Sarter, N. B., Woods, D. D., & Billings, C. E. (1997). Automation surprises. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (2nd ed., pp. 1926—1943). London, UK: Wiley.
Schulman, P. R. (1993). The negotiated order of organizational reliability. *Administration and Society*, *25*(3), 353–372.
SEC. (2010). *Findings regarding the market events of May 6, 2010: Report of the staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues*. Washington, DC: Securities and Exchange Commission.
Shepardson, D. (2020, May 21). Top safety official at Waymo self-driving unit stepping down. *Reuters*. Retrieved from https://www.reuters.com/article/us-waymo-safety-idUSKBN22x2Q9
Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., … Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*, 354–359.
Smithson, M. (1989). *Ignorance and uncertainty: Emerging paradigms*. London, UK: Springer-Verlag.
Smithson, M. (1990). Ignorance and disasters. *International Journal of Mass Emergencies and Disasters*, *8*(3), 207–235.
Snook, S. A. (2000). *Friendly fire: The accidental shootdown of US black hawks over northern Iraq*. Oxford, UK: Princeton University Press.
Stanton, N. A., Salmon, P. M., Walker, G. H., & Stanton, M. (2019). Models and methods for collision analysis: A comparison study based on the Uber collision with a pedestrian. *Safety Science*, *120*, 117–128.
Steinhardt, J. (2015, June 24). Long-term and short-term challenges to ensuring the safety of AI systems. *Academically Interesting blog*. Retrieved from https://jsteinhardt.wordpress.com/2015/06/24/long-term-and-short-term-challenges-to-ensuring-the-safety-of-ai-systems/
Stern, R. (2018, June 21). Self-driving uber crash 'avoidable,' driver's phone playing video before woman struck. *Phoenix New Times*. Retrieved from https://www.phoenixnewtimes.com/news/self-driving-uber-crash-avoidable-drivers-phone-playing-video-before-woman-struck-10543284
Stilgoe, J. (2018). Machine learning, social learning and the governance of self-driving cars. *Social Studies of Science*, *48*(1), 25–56.
Stilgoe, J. (2020). *Who's driving innovation? New technologies and the collaborative state*. London, UK: Palgrave Macmillan.
Stoica, I., Song, D., Ada Popa, R., Patterson, D., Mahoney, M. W., Katz, R., … Abbeel, P. (2017). A Berkeley view of systems challenges for AI. arXiv:1712.05855 [cs.AI]
Strauss, A. L., & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. London, UK: Sage.
Svenson, O. (1991). The accident evolution and barrier function (AEB) model applied to incident analysis in the processing industries. *Risk Analysis*, *11*, 499–507.
Tomašev, N., Glorot, X., Rae, J. W., Zielinski, M., Askham, H., Saraiva, A., … Mohamed, S. (2019). A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, *572*, 116–119.
Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., … Kittler, H. (2020). Human–computer collaboration for skin cancer recognition. *Nature Medicine*, *26*(8), 1229–1243.
Turner, B. (1976). The organizational and interorganizational development of disasters. *Administrative Science Quarterly*, *21*(3), 378–397.
Turner, B. (1978). *Man-made disasters*. London, UK: Wykeham.
Turner, B. (1979). The social aetiology of disasters. *Disasters*, *3*, 53–59.
Turner, B. (1983). The use of grounded theory for the qualitative analysis of organizational behaviour. *Journal of Management Studies*, *20*(3), 333–348.
Turner, B. (1994). Causes of disaster: Sloppy management. *British Journal of Management*, *5*, 215–219.
Turner, B. A. (1981). Some practical aspects of qualitative data analysis: One way of organizing the cognitive processes associated with the generation of grounded theory. *Quality and Quantity*, *15*, 225–247.
Turner, B., & Pidgeon, N. (1997). *Man-made disasters* (2nd ed). Oxford, UK: Butterworth-Heinemann.
Uber A. T. G. (2016, December 14). San Francisco, your self-driving Uber is arriving now. Retrieved from https://www.uber.com/blog/san-francisco/san-francisco-your-self-driving-uber-is-arriving-now/
Uber A. T. G. (2018a). *A principled approach to safety*. San Francisco, CA: Uber Advanced Technologies Group. Retrieved from https://data.ntsb.gov/Docket/Document/docBLOB?ID=40477737%26FileExtension=.PDF%26FileName=Operations%20Attachment%20-%20Uber%20Advanced%20Technologies%20Group%20A%20Principled%20Approach%20to%20Safety%202018-Master.PDF
Uber A. T. G. (2018b). *Uber ATG Safety report supplement: Internal and external safety reviews*. San Francisco, CA: Uber Advanced Technologies Group. https://uber.app.box.com/v/UberATGSafetySupplement And: Retrieved from https://aurora-dev.cdn.prismic.io/aurora-dev/4f3e03fe-7d41-4bed-a998-7b2d918b9579_UberATGSupplementSafetyReview2018.pdf
Uber A. T. G. (2018c, December 20). *Learning from the past to move forward, by Meyhofer E*. Retrieved from https://medium.com/@UberATG/learning-from-the-past-to-move-forward-f4af566f2c3
Uber A. T. G. (2018d, November 2). *A principled approach to safety, by Khosrowshahi D*. Retrieved from https://medium.com/@UberATG/a-principled-approach-to-safety-30dd0386a97c
Uber A. T. G. (2019). *Laying the groundwork for self-driving vehicle safety, by Meyhofer E*. Retrieved from https://medium.com/@UberATG/trailblazing-a-safe-path-forward-e02f5f9ef0cc
Vaughan, D. (1996). *The challenger launch decision: Risky technology, culture and deviance at NASA*. London, UK: Chicago University Press.
Wakabayashi, D. (2018, March 23). Uber's self-driving cars were struggling before Arizona crash. *The New York Times*. Retrieved from https://www.nytimes.com/2018/03/23/technology/uber-self-driving-cars-arizona.html

Wakabayashi, D., & Conger, K. (2018, December 5). Uber's self-driving cars are set to return in a downsized test. *The New York Times*. Retrieved from https://www.nytimes.com/2018/12/05/technology/uber-self-driving-cars.html

Waterson, P. (2019). Autonomous vehicles and human factors/ergonomics – A challenge but not a threat. *Ergonomics*, *62*(4), 509–511.

Waterson, P. (2020). Causation, levels of analysis and explanation in systems ergonomics. A closer look at the UK NHS Morecambe Bay investigation. *Applied Ergonomics*, *84*, 103011.

Waterson, P., Jenkins, D. P., Salmon, P. M., & Underwood, P. (2017). 'Remixing Rasmussen': The evolution of Accimaps within systemic accident analysis. *Applied Ergonomics*, *59*(Part B), 483–503.

Weick, K. E. (1987). Organizational culture as a source of high reliability. *California Management Review*, *24*(2), 112–127.

Weick, K. E. (1995). What theory is *not*, theorizing *is*. *Administrative Science Quarterly*, *40*, 385–390.

Weick, K. E. (2004). Normal accident theory as frame, link, and provocation. *Organization and Environment*, *17*, 27–31.

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (1999). Organizing for high reliability: Processes of collective mindfulness. In R.S. Sutton & B.M. Shaw, (Eds.), *Research in organizational behaviour* (Vol. *1*, pp. 81–123). Stanford, CA: JAI Press.

Weick, K. E., & Sutcliffe, K. M. (2001). *Managing the unexpected: Assuring high performance in an age of complexity*. San Francisco, CA: Jossey Bass.

Weick, K. E., & Sutcliffe, K. M. (2003). Hospitals as cultures of entrapment: A re-analysis of the Bristol Royal Infirmary. *California Management Review*, *45*(2), 73–84.

Wiig, S., Aase, K., Billett, S., Canfield, C., Røise, O., Njå, O., … RiH-team. (2020). Defining the boundaries and operational concepts of resilience in the resilience in healthcare research program. *BMC Health Services Research*, *20*, 330.

Winfield, A., Winkle, K., Webb, H., Lyngs, U., Jirotka, M., & Macrae, C. (2021). Robot accident investigation: A case study in responsible robotics. In A. Cavalcanti, B. Dongol, R. Hierons, J. Timmis, & J. Woodcock (Eds.), *Software engineering for robotics* (pp. 165–187). London, UK: Springer.

Winfield, A.F.T. (2012). *Robotics: A very short introduction*. Oxford, UK: Oxford University Press.

Winfield, A.F.T., & Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *376*(2133), 20180085.

Winfield, A. F. T., & Jirotka, M. (2017). The case for an ethical black box. In A. Natraj, S. Cameron, C. Melhuish, & M. Witkowski (Eds.), *Towards autonomous robotic systems* (pp. 262–273). Cham, Switzerland: Springer International Publishing.

Wong, J. C. (2016a, December 17). California threatens legal action against Uber unless it halts self-driving cars. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2016/dec/16/uber-defies-california-self-driving-cars-san-francisco

Wong, J. C. (2016b, December 22). Uber packs up failed self-driving car trial in California and moves to Arizona. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2016/dec/22/uber-self-driving-car-san-francisco-arizona

Wortham, R. H., Theodorou, A., & Bryson, J. J. (2017). Robot transparency: Improving understanding of intelligent behaviour for designers and users. In Y. Gao, S. Fallah, Y. Jin, & C. Lekakou (Eds.), *Towards autonomous robotic systems: TAROS 2017* (Vol. 10454, pp. 274–289). London: Springer.

Yu, E. (2021, January 25). First commercial autonomous bus services hit Singapore roads. *ZDNet*. Retrieved from https://www.zdnet.com/article/first-commercial-autonomous-bus-services-hit-singapore-roads/