Close calls in cyberspace: strengthening cybersecurity by learning from near-misses

Steen, T. van; Wolbers, J.; Real, C. del; Rizza, C.; Chauhan, A.; Matser, A.; ... ; Treurniet, W.

Close Calls in Cyberspace: Strengthening Cybersecurity by Learning from Near-Misses

Tommy van Steen
Leiden University
t.van.steen@fgga.leidenuniv.nl

Jeroen Wolbers
Leiden University
j.j.wolbers@fgga.leidenuniv.nl

Cristina Del-Real
Leiden University
c.del.real@fgga.leidenuniv.nl

ABSTRACT

Organizations can improve their cybersecurity by learning from both incidents and success stories. These successes can, in many cases, be defined as 'near-misses', when a potentially successful attack is stopped just in time. However, it is unclear what can be categorized as a near-miss in cybersecurity. To bridge this gap, we define four distinct categories of near-misses in cybersecurity, along two axes of interest, being the locus of intervention (human or system) and the nature of the intervention (active or passive). We argue that understanding the workings of these near-misses can support organizations in improving their cybersecurity practices to build stronger futures. To do so, we outline a research agenda of questions and topics that require further examination to fully leverage the insights on near-misses in organizational cybersecurity.

Keywords

Cybersecurity, near-misses, information systems, cyberspace, organizational cybersecurity

INTRODUCTION

Cyberattacks aimed at organizations are on the rise and the detection of these incidents is considered difficult (Albanese et al., 2011). For instance, on average, it takes 181 days for a data breach to be discovered and a further 60 days before the breach has been resolved (IBM, 2025). In timeframes as long as these, harm can easily materialize and affect the organization, its employees, and their clients. This harm can take many forms. One way of assessing harm in cybersecurity is by looking at the CIA-triad known in data protection (Goodman & Rowland, 2020), where the Confidentiality, Integrity, and Availability of data are key in securing data and systems, and harm can be categorized along one or more of these three categories. Confidentiality links to the data only being accessible by authorized parties, where harm could take the form of data leaks or public access to sensitive data. Integrity links to the trustworthiness of the data, that it has not been altered and no data is added or deleted. Availability links to the data being available to the users who need access, where harm can be experienced in the form of not being able to access the data due to the system being offline, or the data being encrypted by cybercriminals who demand a ransom in exchange for the decryption key.

Organizations are well aware of the potential harm that can be caused by cybercriminals to their organization, its employees, and their clients. While they might take many precautions to avoid these harms, no organization is immune and successful attacks do occur. As a result, organizations can learn from incidents by assessing what went wrong and by adopting best practices as set out by other organizations and cybersecurity standards such as the NIST framework or the ISO 27001 standard. These efforts can help improving organizational cybersecurity after an incident. However, not all attacks are successful and learning should not only occur based on incidents. Understanding what is going well in organizations, for instance when attacks are successfully prevented, or the impact minimalized, is another avenue to explore so that the toolbox of cybersecurity solutions can be expanded (van Steen et al., 2024; Wolbers et al. 2025).

Learning from attacks that did not lead to a harmful incident can help organizations to improve their cybersecurity. However, that an attack fails to cause harm, does not mean that everything went as expected. Perhaps some defense mechanisms were overcome or some barriers were breached whereas others were not. Or perhaps it was only due to the intervention of the system under attack, or an attentive employee, that the attack was detected in time and neutralized before any harm could be done. These cases, where an attack did reach beyond some barriers, but was stopped by others, could be seen as near-misses, a safety science concept that suggests that there is not simply 'an incident' or 'no incident', but that there are cases that warrant attention, simply by only just going right instead of terribly wrong (Gnoci & Saleh, 2017). Learning from near misses is less costly than learning from incidents, so they provide an important opportunity for organizations to find out why things went 'just right'. Organizations can examine whether ensuring safety or security was due to the system, employees, and processes working as intended, or whether changes need to be made to ensure these successes in the future. In cybersecurity, the ecosystem of digital infrastructure and ever-changing threats similarly produces precursors and weak signals of incidents. Although regulators have begun to recognize the relevance of near-misses –for example, the European Union defines in the NIS2 Directive near-misses as an event that could have compromised data or services but was prevented or did not materialize– organizational practice remains focused on reactive incident management.

This paper proposes an analytical framework for transferring the lessons of near-miss management from safety science to organizational cybersecurity. Our analytical framework defines four categories of near-misses that can occur, based on active or passive human or system action. These four categories all warrant a different approach by organizations that want to learn from these near-misses to build stronger cybersecurity for the future. After discussing the categories and their implications for organizational strategies, we put forward a research agenda to advance the field of near-misses in cybersecurity.

NEAR-MISSES AND THE NATURE OF ACCIDENTS IN SAFETY SCIENCE

To understand the concept of a near-miss, we first need to broaden our scope to discuss accident models in safety science, of which near-misses are a specific category. Accident models have evolved significantly over the past decades, reflecting a broader conceptual shift from linear, failure-oriented explanations toward more dynamic and resilience-based understandings of how accidents occur (Wong & Pawlicki, 2025). Early ideas about how accidents occur in industrial settings relied on linear, chain-of-events reasoning, in which accidents were viewed as the inevitable outcome of a sequential breakdown (Reason, 1990a). Such 'domino-style' models assumed that removing or controlling one link in the causal chain could prevent an accident (Heinrich, 1941). It provided a good fit to analyze and explain mostly technical failures in organizations. However, as industries such as aviation, healthcare, nuclear power, and cyber-physical systems exhibited more complex, non-linear event chains, the limitation of such linear reasoning became apparent. This paved the way for more comprehensive accident models.

A major conceptual turning point can be traced back to James Reason's Swiss Cheese Model (Reason, 1990b). In this model, defenses that prevent an accident are visualized as slices of Swiss cheese that inhibit different gaps. An accident occurs when these weaknesses align to allow a hazard to pass through all layers and 'holes in the cheese', creating a fatal stream of events. In this line of thinking, Reason (1990a) argued that accidents do not solely arise from operator errors, but from the alignment of latent conditions embedded within organizational structures, processes, and cultures. This idea was pushed by a set of consecutive disasters in the 1980s, including the Bhopal chemical explosion (1984), the Chernobyl nuclear disaster (1986), NASA's Space shuttle Challenger explosion (1986), the Exxon Valdez fire (1987), the Herald of Free Enterprise sinking (1988) and the Piper Alpha oil platform fire (1988) (Larouzee & Le Coze, 2020). In the wake of these events, extensive accident reports began to highlight the organizational dimension that indicated the limited role of human error in the chain of events leading up to these accidents (Reason, 1990a). Reason reframed accident causation as a system-level phenomenon and highlighted how flawed supervision, inadequate training, and design shortcomings interact to create unsafe conditions long before a triggering event occurs. His framework became widely adopted across industries, further institutionalizing a systems-based view of human error.

While Reason's model emphasized latent failures, Rasmussen (1997) introduced a more dynamic view of accident causation based on cognitive science concepts. He conceptualized socio-technical systems as constantly adapting under competing pressures, such as economic constraints, workload demands, and performance expectations. Under these pressures, organizational behavior naturally 'drifts' toward the boundaries of acceptable performance as actors seek efficiency and try to cope with these demands and constrains. Instead of discrete failures in an event chain, Rasmussen (1997) argued that systems gradually migrate into hazardous conditions and people do not have a clear view of the risks involved as they are occupied with other demands. This thinking illustrated how decisions made at various hierarchical levels and points in the organization converge across time. Similar processes of gradual migration towards unsafety became highly influential in later research on practical drift (Snook, 2011) and normalization of deviance (Vaughan, 1999).

Once the shift had been made from event chains to organizational systems, Leveson (2004) proposed the System-Theoretic Accident Model and Processes (STAMP). STAMP represents a shift away from event-chain metaphors entirely, instead treating safety as an emergent property of a system. According to Leveson (2004), accidents result from inadequate enforcement of safety constraints, rather than from isolated component failures. STAMP emphasizes feedback loops, control actions, and system-level interactions that play a vital role in modern high-tech, tightly coupled systems. In parallel, Hollnagel (2004) introduced the Functional Resonance Analysis Method (FRAM), which also challenged traditional assumptions about causality. Instead, FRAM assumes that everyday performance is variable and that this variability can produce both successes and failures (Hollnagel, 2017). This conceptualization of performance variability provides a bridge to Hollnagel's later work on Safety-I and Safety-II.

Safety-I represents the conventional view of safety with a focus on learning from mistakes to prevent future incidents (Hollnagel et al., 2006). Roots of this approach lie in industrial systems of the 1970s and the principles of Scientific Management (Taylor, 1911). Here we go back to the linear, failure-oriented explanations we introduced at the beginning of this section. In the 70s, substantial efficiency gains were achieved in manufacturing firms by studies of task optimalization. The foundational idea was that safety could be achieved if work processes were carefully designed and codified into procedures, and personnel was adequately trained to recognize and anticipate potential issues (Provan et al., 2020). The Safety-I approach builds on this idea of 'task decomposability', which means that one can break down a system into subtasks to see if it is functioning or malfunctioning, and identify and isolate failures.

In reaction to the failure-centric Safety-I approach, and the previously described emergence of system level explanations of safety, scholars began to recognize the importance of learning from how people adapt in complex technological environments (Provan et al., 2020), whereby near misses form opportunities to adapt. This cumulated into the Safety-II approach that analyses how actors perform their work to manage variable circumstances. Various theories describe how industries operating with highly complex technologies can achieve high safety records, such as High Reliability Theory and Resilience Engineering (Hollnagel et al., 2006). In these theories variability in performance is considered both normal and necessary, as adjusting performance to changing conditions is essential for maintaining system functionality. In the literature, we thus see that Safety-II studies place a strong emphasis on how professionals at the 'sharp end' of the organization adapt and align their actions to ensure safety during operations (Reason, 1990b; Schulman, 1993). This idea, where professionals play a critical role to anticipate risks and intervene to prevent incidents, is known as guided adaptability (Provan et al., 2020).

This conceptualization of near-misses as opportunities for adaptation has not always been around in this way, as previous studies often treat near-misses in relation to a Safety-I notion that centers human error. Similar to accident theories, early ideas around near-misses also find their roots in linear safety models (Heinrich, 1931). Bird and Germain (1996) point out their value as an effective analysis of root causes to reduce the number and/or impact of accidents. In these Safety-I oriented studies, accidents are regarded as a chain of events that have a specific sequence, in which certain conditions and events occur in a chain that results in a failure to prevent harm. Such a sequence starts off with a nominal initiating event, followed by increasingly more hazardous events or system states that lead up to an accident with adverse consequences (Gnoni & Saleh, 2017). In this sequential reasoning, a near-miss is very similar to an accident sequence except for a few missing elements in the accident sequence which prevent further escalation (Saleh et al, 2013). Such safety analysis provided a lot of insights into accidents in technically oriented, linear systems.

Safety-II studies often argue that organizations need to look at both successful and unsuccessful adjustments, recognizing that everyday performance contains the seeds of both safety and risk (Hollnagel et al., 2006). Near-miss data exposes these patterns of functional resonance that appear through mismatches between "work-as-imagined" and "work-as-done." Because near-misses capture real-world adaptations before they lead to adverse outcomes, they provide empirically grounded opportunities for organizational learning without the social, financial, or psychological costs associated with actual accidents. Instead of treating near-misses as simple precursors to failure, a Safety-II approach encourages organizations to study them as byproducts of system that are stretching their resilience, revealing how people navigate complexity, try to compensate for design limitations, and maintain performance despite variable or unexpected conditions.

Still, research on near misses also consistently shows that organizations struggle to convert early warnings into meaningful safety improvements (Gnoni et al., 2022). The research on near-miss management systems is somewhat fragmented and unevenly implemented across sectors. There are only a few baseline architectures of near-miss management systems in place (van der Schaaf, 1995). Phimister et al. (2003) and Gnoni et al. (2022) furthermore argue that we need to clearly define what is regarded as a near-miss, so that detection and reporting systems can be fine-tuned to identify the origins of the event so that near miss data can be transformed into useful information for safety improvement and prevention of future occurrences (Gnoni & Saleh, 2017). Efforts are required to strengthen the role of near misses in preventing accidents and increasing resilience.

NEAR-MISSES IN CYBERSECURITY

The cybersecurity field has started to recognize near-misses as a category distinct from incidents. Bair et al. (2017) defines a cyber near-miss as an event short of a full incident because some controls function as intended and contain the damage. The authors propose a Cyber Safety Reporting System analogous to the Aviation Safety Report System. They note that research into near-misses can reveal trends in attacks, control effectiveness and avoidable mistakes; for example, if a user clicks a phishing link but the phishing site has been taken down, the event should be treated as a near-miss. Sbriz (2023) similarly describes a near-miss incident, in the context of cybersecurity, as an unplanned event with potential impact that does not cause significant consequences, and discusses the role of antivirus detection. These events are important risk indicators and should be analyzed even if remediation is automatic (Sbriz, 2023). Others have suggested that near-misses are important, as business decisions can be affected by situations where attacks did not result in large incidents, but where the situation could also not be categorized as business as usual. In a study by Schuetz and colleagues (2025), investors showed positive responses to near-misses in ransomware attacks, valuing stock prices higher than when no attack took place.

Regulation has begun to codify the concept. As mentioned above, the NIS2 Directive (EU Directive 2022/2555) defines a near miss as an event that could have compromised the availability, authenticity, integrity or confidentiality of network data or services but was successfully prevented or did not materialize. This can be seen as an indicator of the growing importance of near-misses analysis in cybersecurity. Unlike physical systems, cyber systems face intelligent adversaries who adapt their tactics, making early indicators crucial. Knake et al. (2021) argue that many factors that make incident investigation complex (such as liability concerns and fear of reputational damage) are reduced when studying near-misses. Because at least one control has succeeded, victims can share information more openly and attackers know there is a chance they have been detected. Researchers have therefore suggested creating confidential, anonymized near-miss reporting systems and rewarding reporting to encourage participation. For this reason, the authors propose liability protection to incentivize near-miss reporting. Ebert and colleagues (2025) have argued in favor of integrating near-misses into incident reporting systems at the design stage. They conducted a review into potential factors to be included in incident reporting systems, outlining several factors they consider vital in improving these systems to support the reporting of cybersecurity incidents. The authors consider near-misses an integrated part of such a system, but also conclude that clear definitions of near-misses in cybersecurity have not yet been established beyond simply providing examples of what a near-miss in cybersecurity could look like.

We propose an analytical typology to make sense of near-misses. Drawing on both safety science and cyber incident case studies, we classify near-misses in four quadrants along two axes: the locus of intervention (human or system) and the nature of the intervention (active or passive). The typology recognizes that both human actors and technical controls can either act or refrain from acting to prevent an incident. Each category is defined below with examples.

Human Active Near-Misses

A human active near-miss is a type of near-miss in which a person detects a malicious activity and deliberately intervenes to prevent harm. This intervention could involve applying expertise, exercising vigilance or making a discretionary decision. Many near-misses can be the result of people actively stopping an attack. For instance, an employee in a Security Operations Centre (SOC) might notice a change in network traffic that does not fit with the everyday workings of the organization and decide to block access to the network from certain devices to investigate the anomaly without harming the system. In these cases, it is the 'human in the loop' that ensures the security of the system by actively interfering in the process that is carried out by the system. Other forms of human active near-misses can be found in phishing attacks. If an employee receives an email and correctly suspects that it is a phishing attempt, they might decide to delete the email from their system, or report it to a security specialist so that they can investigate its origins and potential harms.

Examples of real-world cases include the WannaCry ransomware kill switch in 2017. During the global WannaCry outbreak, security researcher Marcus Hutchins noticed that the malware queried an unregistered domain and registered the domain, which acted as a kill switch and halted the ransomware's propagation. Hutchins's deliberate action prevented further infections and allowed organizations that had not yet been hit to avoid serious disruption (Jonsson & Modig, 2023). Another example occurred during the Bangladesh Bank heist in 2016. Attackers attempted to steal nearly one billion US dollars via fraudulent SWIFT messages. A Deutsche Bank staff member observed that the word *foundation* was misspelled as *fandation* in a payment instruction and sought clarification; this led to the discovery of the fraud and prevented additional losses of around US$850–870 million (Reuters, 2016).

Human Passive Near-Misses

A human passive near-miss is a type of near miss in which an individual does not take an action that would have facilitated an attack. At times, the *not* acting of people can result in a near-miss. In this case, their continuation with everyday work allows the threat to fizzle out. Such non-action should not be understood as mere negligence, but as the product of contextual constraints, busy work schedules, or organizational structures that shape everyday work. From a Safety-II perspective, human passivity often reflects how real work is organized under time pressure, cognitive load, and prioritization demands (Hollnagel et al., 2006). Employees continuously triage tasks, defer actions, and suspend decisions when signals are ambiguous or workloads peak (Rader & Wash, 2015). In this sense, not acting may reflect adaptive behavior to competing demands rather than inattentiveness. Near-misses arise when such postponement inadvertently prevents attack exploitation. For example, this can be the case when an employee receives a phishing email and is inclined to believe it is a genuine message, but does not reply to it in any way. When a phishing email is considered believable but the receiver is distracted, has other tasks to complete first, or is just about to start a meeting or end a workday, they might not reply to the message right away. When they next open their mailbox, new, more pressing, messages might be received that need to be dealt with first and the older phishing email is forgotten (Frank et al., 2022). In finance, policies often require a second sign-off for wire transfers or changes to vendor details. When an employee refuses to bypass these controls despite pressure, fraudulent transfers are averted. Such inaction constitutes a near-miss because the absence of a risky action prevents harm. In those cases, it is not the skill, expertise or knowledge of the employee that prevents the incident from happening, but merely the passive act of not interacting with the email even though they were deceived into believing that the email was genuine.

Importantly, these events raise an analytical question: to what extent should organizations learn from outcomes that were shaped by non-deliberate adaptation? While passive near-misses may appear less informative than active interventions, they highlight how workload management contributes to system resilience. Studying these events can reveal whether organizational safety margins are accidental or intentionally designed, and whether reliance on passivity introduces hidden fragilities. Treating human passivity as analytically meaningful thus aligns with Safety-II principles by recognizing that successful outcomes may also depend on how work is not done, postponed, or deprioritized.

System Active Near-Misses

A system active near-miss is a near-miss in which technical controls detect and respond to a threat by taking corrective action. The system might block, quarantine or remediate malicious activity without human intervention. This type of near-miss works in similar ways to the active near-misses in humans, with the difference that the decision to act is made by the system, be it through rules set earlier by the developer of the system, or through (reinforcement) learning as a result of AI implemented solutions or broader machine learning methods. In these cases, similar to the SOC employee in the human active near misses, a system could block access by a suspicious account based on monitoring and logging of data that suggests something unusual is taking place in that account. This can be in the form of too many data sources being accessed in a short period of time, but also the types of data being accessed, copied, stored, moved or deleted. If the system actively takes measures to ensure the workings of a system after such an event, this could be categorized as a near-miss, assuming that the system is the final or near final barrier in the protective process.

Sbriz (2023) notes that near-misses include events where antivirus software detects and isolates a virus or where an intrusion detection system blocks a connection attempt. These automated responses prevent damage while signaling weaknesses. The role of monitoring in near-misses is also evident in the example of a manufacturer that detected an unauthorized Raspberry Pi device on its industrial control network using Darktrace's OT monitoring. Through continuous network monitoring and passive asset identification, the system identified a "near-miss" supply-chain threat before it could disrupt operations (Wong, 2025).

System Passive Near-Misses

A system passive near-miss is a type of near-miss in which the absence of a system response inadvertently prevents harm. A system passive near-miss is probably the least likely to occur out of the four types of near-misses. In these cases, it is a system that does not actively stop an attack from happening, but that does so by not acting. This could be the case in circumstances where a system might already malfunction as a result of a bug or failure of physical hardware, leading to the attack being stopped simply by shutting down a system because of other factors than the actual attack. Such near-misses are particularly relevant in tightly coupled socio-technical systems where attackers rely on specific assumptions about system availability, configuration, or interoperability. When these assumptions are violated due to system downtime, version mismatches, or architectural heterogeneity the attack may stall or collapse. From a system perspective, this reflects a form of latent resilience that emerges because

cyberspace and its information systems often develop in a rhizomatic manner that result in an unpredictable maze of system-level interdependencies (Vieira & Ferasso, 2010).

A second example that could be categorized as a passive system near-miss, is when an attacker has not configured their malware correctly. If the malware is not suited for the hardware and software used by the victim organization, or if there is a bug in the malware, this might lead to a first breach of the system, without causing the harm that was intended by the attacker. For instance, if the attacker configured their malware for a newly updated version of software present in the system, but the system did not successfully update yet, the attack might be stopped without active interference from the system.

Although these outcomes may appear accidental, system-level safety theories argue that successful functioning often depends on such negative feedback effects. Complex systems rarely fail, or succeed, solely due to a single action. Instead, safety also emerges from misalignments, delays, and frictions across interacting components. One might argue that system passive near-misses expose these emergent properties by revealing where attacker models diverge from real system behavior. However, these near-misses also pose a dilemma for organizational learning. Because they do not reflect intentional defensive capability, over-reliance on such outcomes may mask vulnerabilities and create a false sense of security. A system passive near-miss therefore demands careful interpretation: organizations must distinguish whether the outcome reflects exploitable structural diversity or merely temporary luck.

While passive near-misses may appear less informative than active interventions, safety science has repeatedly shown that successful outcomes often depend on non-events, deferrals, and mismatches between assumptions and reality (Hollnagel, 2017). Excluding passive near-misses would bias learning toward visible actions and obscure other critical system properties that shape resilience.

ORGANIZATIONAL STRATEGIES FOR MANAGING NEAR-MISSES IN CYBERSECURITY

While the distinction between active and passive near-misses is useful for understanding the types of near-misses that an organization can experience, this distinction should also serve as a starting point for organizations to decide which steps to take to strengthen barriers and reduce the near-misses to non-events. To do so, organizations must develop separate strategies for the four categories of near-misses as defined in this paper. As a starting point, organizations should likely focus on the active near-misses before turning to the passive quadrants. Both systems and employees can be strengthened based on the active near-misses that are found in the everyday working of the organization. The reason to focus on active near-misses first, lies in that a human or system taking action is likely more easy to monitor, encourage, and sustain than the absence of action.

In terms of active system near-misses, the system can be evaluated and strengthened to carry out automated checks and security responses in case of suspicious activity, such as extreme levels of CPU use, or changes made to databases. This can be strengthened not only by investigating near-misses from the past, but also the ‘work as done’ situations that employees find themselves in. What is considered reasonable behavior of an employee that should not be cause for concern, and where does an employee perform unusual actions that warrant further investigation? By adjusting the system expectations of normal and abnormal situations, the system can be improved to flag issues and take preventative measures before an attack turns into a system breach. While this outlines the regular working of a security system, assessing near-misses from the past to update the ways of working creates an important feedback loop that helps improve the security of organizations before incidents occur.

For the human active near-misses, learning from the behavior of employees is vital. Organizations need to determine what forms of active near-misses exist on the employee level and strengthen those initiatives. This involves asking questions about the near-miss that usually cannot be easily taken from logs or monitoring system activity. Asking employees why they decided to report an email as a potential phishing message might lead to insights that are not clear from an analysis of the suspicious URL alone. Perhaps an employee based their assessment on earlier experiences with phishing in their job, realized the language was slightly different than could normally be expected, or felt that the email touched upon corporate topics that usually would not be discussed with them based on their job role or responsibility. Alternatively, perhaps they cannot explain their reasoning, but had a gut feeling that ‘something is off’ that cannot easily be pinpointed to a clue in the message. Furthermore, while message factors might play a role in an email being considered suspicious or not, research has shown that other factors such as context and personality play a role as well (Eftimie et al., 2022; Frank et al., 2022). If near-misses can be identified based on those aspects, organizations can work to strengthen the barrier of employee behavior by training end-users, or forming teams of people with specific skills or ways of working.

Apart from strengthening existing barriers, such as automated system responses, flagging behavior of SOC-employees, or the reporting behavior of end-users, organizations can also opt to add new barriers to reduce the

number of successful attacks. If the near-misses identified by an organization are all barriers that are a so-called 'last line of defense', new security measures might be required to reduce the number of attacks that reach this last line of defense, or even overcome those barriers and successfully breach into the system. While the specifics of what these barriers might look like is beyond the scope of this paper, adding new barriers alongside strengthening the existing ones might be warranted in some situations. However, it is important to note that adding too many new barriers might run the risk of increasing the difficulty that employees have in carrying out their day-to-day activities. Balancing security and productivity is key, and creating extra barriers might be considered unwanted from that perspective, whereas strengthening existing barriers might put less of an additional strain on employees and systems to work as expected.

TOWARDS A RESEARCH AGENDA ON NEAR-MISSES IN CYBERSECURITY

The study of near-misses in cybersecurity requires a dedicated and structured research approach that reflects the unique characteristics of cyber incidents. While safety science provides a rich foundation for understanding how organizations can learn from near misses, the cybersecurity field currently lacks the conceptual clarity to move forward scholarship. A first priority for future research is to develop shared definitions of this phenomenon. The lack of clarity is directly related to the nature of cybersecurity incidents. Near-misses in cybersecurity are difficult to delineate because the boundary between a normal event, an anomaly, a thwarted attack, and a genuine incident is often ambiguous. Determining whether a near-miss is defined by technical signals alone, by organizational exposure, or by the number of breached barriers remains an open question. Clarifying these definitional boundaries is essential, as the field cannot build comparable datasets or design interventions without conceptual consistency.

A second avenue for research concerns the detection and measurement of near-misses. Cyber events often unfold without leaving clear pre-determined traces, and many near-misses are only visible because a security analyst intervened or because a system behaved unexpectedly. Capturing these events thus requires methodological innovation. Future work could explore how log data, SOC workflows, user reports, and automated monitoring systems can be combined into detection frameworks. It is equally important to develop ways of observing how cybersecurity "work is done," so that tacit cues, human judgment, and informal practices that stop attacks can be systematically captured. Such methods must also take account of the long detection delays that characterize cybersecurity, raising the question of when a late detection can still be considered a near-miss rather than an incident.

Alongside detection, research could examine how organizations can build effective reporting and learning systems for near-misses. In existing cybersecurity practice, reporting is inconsistent and can be overshadowed by alert fatigue. While high-risk industries have long recognized that open reporting cultures support organizational learning, cybersecurity presents additional challenges: adversarial intent, confidentiality constraints, and a high degree of technical specialization. Future research could investigate how reporting systems can be designed to capture near-misses and store them in databases that can be shared with other organizations. The incentives, governance structures, and feedback loops that encourage consistent vulnerability disclosure require careful study.

A fourth research direction builds directly on the human/system and active/passive quadrants introduced in this paper. Each quadrant represents a distinct mechanism through which near-misses can occur, yet little is known about how these mechanisms work in practice. Human active near-misses, for example, depend on cognitive cues, contextual factors, and experience, and therefore require research that examines how employees recognize suspicious behavior and decide to intervene. Human passive near-misses raise different questions: to what extent should organizations learn from events that went right by coincidence? System active near-misses, such as automated blocking of anomalous behavior, challenge researchers to distinguish between expected system functioning and genuine near-miss events that reveal new vulnerabilities. System passive near-misses that might arise from attacker misconfigurations or system downtime, require analysis of how organizations should interpret such events that reflect luck more than resilience. Studying different categories in depth will help determine how different types of near-misses contribute to security performance and where investments in strengthening barriers might be most effective.

Future research could also link near-misses to broader organizational dynamics that are often identified in safety science studies, such as drift, resilience, and adaptive capacity. Near-misses in a Safety-II analysis reveal how work is actually performed, where boundaries are stretched, and how actors compensate for system limitations. They might provide early warning signs of practical drift between formal security policies and real-world behavior, and expose weaknesses in monitoring, response, and feedback mechanisms. Understanding how near-miss patterns evolve over time may enable researchers to identify when organizations are moving toward hazardous states, offering insights that are highly valuable for proactive cybersecurity management. This line of inquiry also invites integration with resilience engineering, as near-misses may illuminate how anticipation,

monitoring, response, and learning interact under real-world pressures.

Finally, advancing the field requires shared datasets and cross-sector benchmarking. Current cybersecurity datasets overwhelmingly focus on successful attacks or malware characteristics, leaving little empirical basis for studying near-misses. Creating anonymized, cross-organizational near-miss repositories would allow researchers to analyze patterns, compare sectors, and identify recurring vulnerabilities. Because near-miss data is sensitive, research is needed on how to anonymize and aggregate events without exposing operational details. Incorporating insights from red-team and purple-team exercises, simulated attacks, and controlled experiments would enrich these datasets and help overcome the scarcity of naturally occurring near-miss observations.

Taken together, these lines of inquiry outline a coherent and ambitious research agenda. There is a unique opportunity to transform near-misses from incidental observations into a systematic source of learning. Such an agenda has the potential not only to advance academic understanding of cybersecurity, but also to significantly strengthen the everyday cybersecurity practices of organizations.

FUNDING ACKNOWLEDGEMENT

This research was supported by a grant from the Dutch Research Council with grant number KICH1.VE05.23.005.

REFERENCES

Albanese, M., Jajodia, S., Pugliese, A., & Subrahmanian, V. S. (2011). Scalable Detection of Cyber Attacks. In Chaki, N., Cortesi, A. (Eds.), *Computer Information Systems–Analysis and Technologies: 10th International Conference, CISIM 2011, Kolkata, India, December 14-16, 2011. Proceedings* (pp. 9-18). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-27245-5_4

Bair, J., Bellovin, S. M., Manley, A., Reid, B., & Shostack, A. (2017). That was Close: Reward Reporting of Cybersecurity Near Misses. *Colorado Technology Law Journal, 16*, 327-364.

Bird, F.E., Germain, G.L. (1996). *Practical Loss Control Leadership*. Det Norske Veritas.

Ebert, N., Schaltegger, T., Ambuehl, B., Geppert, T., Trammell, A., Knieps, M., & Zimmermann, V. (2025). Learning From Safety Science: Designing Incident Reporting Systems in Cybersecurity. *Journal of Cybersecurity*, *11*(1), tyaf019.

Eftimie, S., Moinescu, R., & Răcuciu, C. (2022). Spear-Phishing Susceptibility Stemming From Personality Traits. *IEEE Access*, *10*, 73548-73561. https://doi.org/10.1109/access.2022.3190009

EU Directive 2022/2555. *NIS 2 Directive*. European Parliament and Council. https://eur-lex.europa.eu/eli/dir/2022/2555/oj/eng

Frank, M., Jaeger, L., & Ranft, L. M. (2022). Contextual Drivers of Employees' Phishing Susceptibility: Insights From a Field Study. *Decision Support Systems*, *160*, 113818.

Gnoni, M. G., Tornese, F., Guglielmi, A., Pellicci, M., Campo, G., & De Merich, D. (2022). Near Miss Management Systems in the Industrial Sector: A Literature Review. *Safety Science*, *150*, 105704.

Goodman, H. B., & Rowland, P. (2020). Deficiencies of Compliancy for Data and Storage: Isolating the CIA Triad Components to Identify Gaps to Security. In E. Imsand, K.-K. R. Choo, T. Morris, & G. L. Peterson (Eds.), *National Cyber Summit* (pp. 170-192). Springer International Publishing. https://doi.org/10.1007/978-3-030-58703-1_11

Heinrich, H.W. (1941). *Industrial Accident Prevention: A Scientific Approach*. Mc-Graw Hill.

Hollnagel, E. (2004). *Barriers and Accident Prevention*. Routledge.

Hollnagel, E. (2017). *FRAM: The Functional Resonance Analysis Method: Modelling Complex Socio-Technical Systems*. CRC Press.

Hollnagel, E., Woods, D. D., & Leveson, N. (2006). *Resilience Engineering: Concepts and Precepts*. Ashgate Publishing, Ltd.

IBM. (2025). *Cost of a Data Breach Report 2025*. https://www.ibm.com/reports/data-breach

Jonsson, K., & Modig, K. (2023). Information and Cyber Security: The WannaCry Attack. *Societal Security Challenges: Drawing Lessons from Case Studies on Cyber and Information Security, Climate Change, Global Covid Pandemic, and Youth, Security & Trust*, *78*.

Larouzee, J., & Le Coze, J. C. (2020). Good and Bad Reasons: The Swiss Cheese Model and its Critics. *Safety science*, *126*, 104660.

Leveson, N. (2004). A New Accident Model for Engineering Safer Systems. *Safety Science*, *42*(4), 237-270.

Phimister, J. R., Oktem, U., Kleindorfer, P. R., & Kunreuther, H. (2003). Near-Miss Incident Management in the Chemical Process Industry. *Risk Analysis: An International Journal*, *23*(3), 445-459.

Provan, D. J., Woods, D. D., Dekker, S. W., & Rae, A. J. (2020). Safety II Professionals: How Resilience Engineering can Transform Safety Practice. *Reliability Engineering & System Safety*, *195*, 10674

Rader, E., & Wash, R. (2015). Identifying Patterns in Informal Sources of Security Information. *Journal of Cybersecurity*, *1*(1), 121-144.

Rasmussen, J. (1997). Risk Management in a Dynamic Society: A Modelling Problem. *Safety Science*, *27*(2-3), 183-213.

Reason, J. T. (1990a). *Human Error.* Cambridge University Press.

Reason, J. T. (1990b). The Contribution of Latent Human Failures to the Breakdown of Complex Systems. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *327*(1241), 475-484.

Reuters. (2016). *This $1 Billion Typo Prevented a Bigger Bangladesh Bank Heist.* Fortune. https://fortune.com/2016/03/10/typo-bangladesh-bank-heist/

Sbriz, L. (2023). Using Near Miss Incidents as Risk Indicators. *ISACA Journal*, *4*, 49-52.

Schuetz, S. W., Chen, Y., Forderer, J., & Ma, Y. (2025). Does Ransomware Make Investors "WannaCry"? On Investors' Divergent Reactions to Ransomware Hits and Near Misses. *MIS Quarterly*, *49*(3), 1153-1168.

Schulman, P. R. (1993). The Negotiated Order of Organizational Reliability. *Administration & Society*, *25*(3), 353-372.

Snook, S. A. (2011). *Friendly Fire: The Accidental Shootdown of US Black Hawks Over Northern Iraq*. Princeton University Press.

Taylor, F. W. (1911). Principles and Methods of Scientific Management. *Journal of Accountancy*, *12*(3), 3.

van der Schaaf, T. W. (1995). Near Miss Reporting in the Chemical Process Industry: An Overview. *Microelectronics Reliability*, *35*(9-10), 1233-1243.

van Steen, T., Del-Real, C., & van den Berg, B. (2024). What Works Well? A Safety-II Approach to Cybersecurity. In *International Conference on Human-Computer Interaction* (pp. 250-262). Cham: Springer Nature Switzerland.

Vaughan, D. (1999). The Dark Side of Organizations: Mistake, Misconduct, and Disaster. *Annual Review of Sociology*, *25*(1), 271-305.

Vieira, L. M. M., & Ferasso, M. (2010). The Rhizomatic Structure of Cyberspace: Virtuality and its Possibilities. *International Journal of Networking and Virtual Organizations*, *7*(6), 549-559.

Wolbers, J., van Steen, T., Del-Real, C., & van den Berg, B. (2025). Cyber Crisis Averted: Using Safety Science Principles to Learn From Success. In *Proceedings of the International ISCRAM Conference*.

Wong, L. M., & Pawlicki, T. (2025). A Review of Accident Models and Incident Analysis Techniques. *Journal of Applied Clinical Medical Physics*, *26*(3), e14623.

Wong, N. (2025). Proactive OT Security | Managing Supply Chain Risk & Rogue Devices. *Darktrace*. https://www.darktrace.com/blog/proactive-ot-security-lessons-on-supply-chain-risk-management-from-a-rogue-raspberry-pi