Comparison of risk analysis approaches for analyzing emergent misbehavior in autonomous systems

Nektaria Kaloudi

Department of Computer Science, Norwegian University of Science and Technology, Norway.
E-mail: nektaria.kaloudi@ntnu.no

Jingyue Li

Department of Computer Science, Norwegian University of Science and Technology, Norway.
E-mail: jingyue.li@ntnu.no

The evolution of autonomous systems depends on their constituent parts' ability to act, seemingly independently, so that their collective behavior, termed emergent behavior, results in novel properties that appear at a higher level. Although these emergent behaviors can be beneficial, systems can also exhibit unintentionally and intentionally malicious emergent misbehaviors. As systems are becoming more complex and sophisticated, their emergence characteristics may result in a new type of risk, called *emergent risk*, which would affect both the systems and society. Although there have been several studies on achieving positive desirable emergent behavior, little attention has been given to the risk of undesirable emergence from either the safety or the security perspective. The main objective of this paper is to provide a structured approach to understanding emergent risks in the context of autonomous systems. This approach has been analyzed based on an emergent risk application example – a swarm of drones. We explore different security and safety risk co-analysis methods with a causal interpretation, and provide a comparative analysis based on theoretical factors that are important for assessing the emergence of various threats. The study results reveal each method's strengths and weaknesses for addressing emergent risks, by providing insights into the need for the development of an emergent risk analysis framework.

Keywords: Emergence, emergent behavior, risk analysis, emergent risks, cyber security, safety, autonomous systems.

1. Introduction

As the world becomes more complex and interconnected, we need a deeper understanding of how autonomy and interdependencies interact to deliver new capabilities without unintended emergent behaviors MITRE (2018). To some extent, their emergence is a natural consequence of the advances in the increasingly ubiquitous information and communication technologies (ICTs). As in the "butterfly effect", small changes in a system can lead to dramatically unexpected outcomes. The notion of emergent behavior has arisen mainly in the context of networked smart systems, where many elements can interact with each other, and an individual study on a limited number of components may not predict the behavior of the whole system.

However, the emergent behavior can be undesirable, leading to significant safety problems. An example of such a safety-related misbehavior is the closure of the London Millennium Footbridge after failing to anticipate the emergence of laterally-induced pedestrian forces Dallard et al. (2001). Another example is the potential for traffic jams in traffic management systems, where analysis of individual human drivers and cars cannot explain the emergence of traffic congestion. Another example, related to security, is the Stuxnet worm targeting Iran's nuclear industrial control systems (ICSs) Kushner (2013), which also had widespread effects on several other countries. Given the potentially serious consequences of such behavior, it is crucial to identify the causes of emergent misbehaviors in systems. These behaviors must be analyzed to find the potential risks of undesirable emergent properties so as to improve system resilience and reinforce its safety and security envelope. Therefore, predicting the risks most likely to emerge would not only ensure better risk mitigation, but could also prevent an undesirable emergent property from arising in the first place.

The objective of this study is to propose a structured approach for studying emergent risks – problems caused by the emergent misbehaviors. First we identify the contributing factors, which are the sources of emergence that need to be assessed by risk analysis methods. Then we discuss feasibility of several risk causal analysis approaches for assessing the risks of emergent misbehaviors. The results suggest that the existing

Proceedings of the 31st European Safety and Reliability Conference
Edited by Bruno Castanier, Marko Cepin, David Bigaud, and Christophe Berenguer
Published by Research Publishing, Singapore.
ISBN: 978-981-18-2016-8; doi:10.3850/978-981-18-2016-8_213-cd

approaches should be improved, and that perhaps a new emergent risk analysis framework should be developed in order to create an automated approach, which would consider both individual components and systemic factors to prevent undesirable emergent misbehaviors.

The paper is organized as follows. Section 2 provides the background of this study. Section 3 analyzes the literature on risk analysis of emergent behavior. Section 4 explains the research method. In Section 5, we present a taxonomy and a structured approach for analyzing emergent misbehavior in autonomous systems. We discuss our approach in a hypothetical scenario, and compare risk analysis approaches based on the factors that affect emergence. Section 6 discusses our results. The conclusions are presented in Section 7.

2. Background

2.1. *Autonomous systems*

An autonomous system can be defined as a *"machine, whether hardware or software, that, once activated, performs some task or function on its own"* Williams and Scharre (2015). Features of autonomous systems can be introduced when parts of the system begin to exhibit autonomous-like functioning behaviors as a result of the interactions among the system's parts and between the system and the external environment. Therefore, the first key concept is autonomy, defined as a system's capacity for *"integrated sensing, perceiving, analyzing, communicating, planning, decision-making, and acting, to achieve its goals as assigned by its human operator(s)"* Huang (2004). The degree of the autonomous capabilities can be represented by the level of autonomy (LoA), which is determined by the system's ability to sense and react to the environment in different ways. A six-level autonomy scale has been developed by the US Navy's Office of Naval Research Williams and Scharre (2015), with higher levels corresponding to increased complexity. From least to most autonomous, the levels are human-operated, human-assisted, human-delegated, human-supervised, mixed-initiative, and fully-autonomous. The second key feature is that autonomy may facilitate more efficient interconnectivity and distribution of functions between various parts of the system and between an autonomous system and its environment.

2.2. *Emergent behavior*

The term "emergent behavior" or "emergence" refers to the phenomenon of new collective properties arising unexpectedly from the behavior of the components in a system. Emergent behavior can be beneficial in artificial systems, such as the robotic systems inspired by the impressive capabilities of social insect colonies in nature. However, many intentional and unintentional events can also trigger undesirable emergent misbehavior, putting the system at risk. Thus, emergence can be considered from many different angles. For instance, Husted and Myers (2014) discuss how emergent phenomena could be studied from either an attack or a defense perspective.

Fromm (2005) classified emergence phenomena into four distinct types based on the spectrum of emergent behaviors: (i) simple emergence without top-down feedback, characterized by feedforward interactions between the components of a system; (ii) weak emergence, which includes top-down feedback and which deals with feedback relations through independent direct and indirect interactions at the low microscopic level; (iii) multiple emergence, which includes many types of feedbacks and deals with many feedback loops on different time scales; and (iv) strong emergence, which involves emergent properties and feedback relations on a higher macroscopic level of complexity, such as between systems of systems (SoS). For our risk analysis purposes, we focus on the first two types of emergence, which can be studied and predictable, at least in principle.

3. Related Work

The issue of unintended emergent behavior in the field of distributed networked systems has been raised by the US National Research Council. The resultant research report Council (2001) about networked systems of embedded computers emphasized that unintended emergent behaviors often emerge when multiple components are combined, and the outcomes are not immediately apparent from the individual components. Mogul (2006) discusses examples of emergent misbehavior in complex software systems, while Ferreira et al. (2013) provided a taxonomy for characterizing emergent properties. Husted and Myers (2014) emphasize the importance of risk analysis to describe non-linear security risks from emergent attacks. The risk posed by emergent security phenomena requires a suitable methodological basis. Allan et al. (2013) review eleven well-established complexity science tools and techniques that can be used to identify emerging risks in complex systems. However, we have reduced our scope to one part of complex systems: autonomous systems. Due to the smartness of autonomous systems Kalluri et al. (2020), we need to consider how their unique features influence the feasibility and usefulness of risk analysis for managing potential emergent misbehavior.

4. Research method

4.1. *Research motivation*

While many functions of autonomous systems have become accessible online, the increasing sys-

tem complexity creates new vulnerabilities and risks, which are greater than the sum of the risks of the individual components Axelrod (2013). An obvious challenge is how to identify risks that are yet unknown and to ensure resilience against them. Continuing the definition of risk in standard ISO 31000, the new ISO 31050 Jovanovic (2019) will further contribute to managing the changing risk landscape due to the emerging risks. While ISO 31000 deals with handling known risks, previously unknown risks also need proper assessment to be well managed by the companies and organizations. There are different types of risk analysis, which can be used to identify different problems and solutions. Our study focuses on the conceptual assessment of existing security and safety risks and on causal co-analysis methods to support and enhance emergent risk analysis. Thus, the goal of this research can be formulated as the following main research question: *can existing risk analysis approaches identify emergent risks?*

4.2. *Research design*

There are various integrated safety and security approaches used to analyze risks Kavallieratos et al. (2020). For our study, we selected methods that included causation in the risk analysis process. There are two important factors to consider: the source of the risk and its consequences. Their relationships can be understood by integrating causation into the risk analysis process, either to identify which causes/sources of risk contributed causally to a consequence or to predict the potential consequences of a cause. Figure 1 shows our focus on risk analysis methods that incorporate safety and security in their cause-effect analysis.

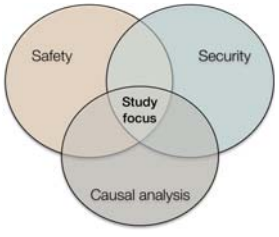

Fig. 1. The study focus on integrated safety and security in risk causal analysis.

Below we list a few well-known safety and security co-analysis methods for the risk analysis, which consider causation in their processes:

- System theoretic process analysis (STPA)-SafeSec Friedberg et al. (2017).
- Failure mode, vulnerabilities and effects analysis (FMVEA) Schmittner et al. (2014).
- Bayesian belief network (BBN) Fenton and Neil (2018).
- Unified framework for risk and vulnerability analysis Aven (2007).

5. Results

5.1. *The emergent risk concept*

5.1.1. *An emergent risk taxonomy*

Emergent risks or "black swans" as Taleb (2007) called them, have three main attributes: (i) they have extraordinary impact; (ii) they are unexpected; and (iii) people can explain them after the fact. We define emergent risks as unintentional with safety impact and intentional with security and potential safety impact. Safety is associated with accidental failures that could result in undesirable consequences for the system's environment. On the other hand, security is related to malicious misuse, particularly cyberattacks. Figure 2 shows a high-level emergent risk taxonomy.

Fig. 2. A high-level emergent risk taxonomy.

On the safety side, increased dependence on technology may lead to common risks arising from failures and human errors or under normal conditions. Systems are vulnerable to failure at scale, and a number of unanticipated emergent effects could arise. Han and DeLaurentis (2013) analyzed the complex propagation of failures through interdependencies between systems by integrating propagating failure rates with inherent individual failures. In addition, complex interactions between components within a system could cause accidents that may emerge from normal operations. For example, a mid-air collision between two flights over the Amazon happened without any root cause of catastrophic equipment failures or human errors Lacagnina (2009). Instead, the investigation showed that the accident occurred within the normal variability of the system's performance range, where a sudden unexpected combination or resonance of performance variations changed the system's functioning De Carvalho (2011).

On the security side, the intentional risks of being exposed by cyber-criminals could threaten society either with or without safety impact. In the emergent domain, attack operations can be

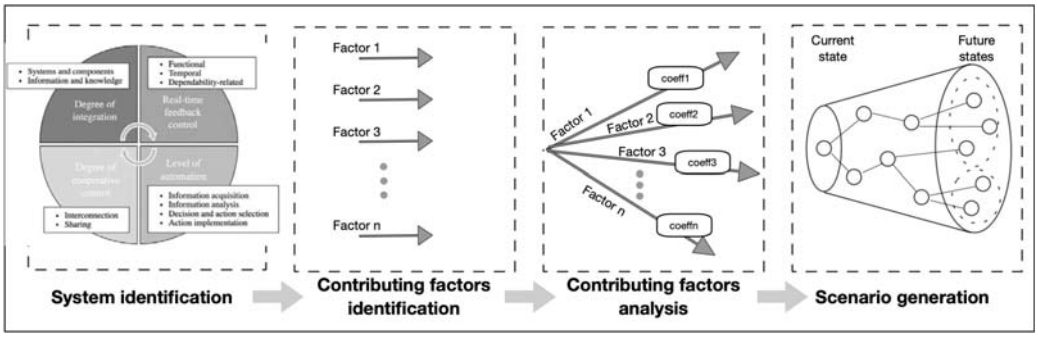

Fig. 3. The general process of analysis of emergent misbehavior in autonomous systems.

divided into known and novel threats. For example, in the known threats, a traditional botnet attack scenario can be considered a form of intentional emergent misbehavior. For example, distributed denial-of-service (DDoS) attacks only emerge when a large number of compromised bots are acting in harmony to achieve a common attack goal of service disruption. For novel threats, in more autonomous infrastructures, such attacks can have different effects, broader and more catastrophic than in the past Kaloudi and Li (2020). Advanced botnets can benefit from emergent collective behavior to cause more robust intent-based attacks based on the swarm mentality Manky (2018). Such swarm botnets using artificial intelligence (AI) technologies with learning capability can involve feedback loops across different levels, combining feedback from the environment.

5.1.2. *Developing an emergent risk profile*

To prepare for unknown risks, it is useful to develop an emergent risk profile, which may include one or more related risks. A common method of developing a risk profile is scenario analysis, which can reveal many important aspects of future situations based on current knowledge. Following Kosow and Gaßner (2008), we propose the following process for establishing an emergent risk horizon context for the risk management process, providing insights for better risk coverage.

- *System identification.* We examine the current state of the autonomous system, considering its characteristics using the conceptual framework of smartness dimensions and its groups of characteristics to evaluate smartness in cyber-physical systems Kalluri et al. (2020).
- *Contributing factors identification.* The International Risk Governance Council (IRGC) Graham et al. (2010) has suggested twelve factors pertinent to emerging risks. For our analysis, we include the seven most relevant factors: scientific unknowns, loss of safety margins, positive feedback, temporal complications, communication, information asymmetries, and malicious motives and acts. These factors are directly connected with autonomous systems. We exclude the other five factors (varying susceptibilities to risk, conflicts about interests, values and science, social dynamics, technological advances, and perverse incentives) because they are not related to the systemic nature of a cyber-physical autonomous system. Rather, they address geographical, political, societal, regulative, and economic aspects, respectively.
- *Contributing factors analysis.* For different autonomous systems, these factors may have different weights. Thus, analysis can be used to add coefficients to the factors for a better evaluation.
- *Scenario generation.* The output will be a series of *what-if* scenarios that can describe possible futures on how the autonomous system might develop. This will allow emergent risk identification through scenario analysis using the "funnel model", in which each identified factor contributes to a better understanding of the possible future states of the system.

Figure 3 presents an approach for understanding emergent risks in autonomous systems. The output of the process can be used as input to the general risk management process Purdy (2010) to better understand the context of the risk.

5.2. *A swarm-of-drones scenario*

We analyze the emergent misbehavior in a swarm of drones using the process in Figure 3. This analysis allows to look at misbehavior in autonomous systems in a systematic way and can be used as a guideline for improving risk assessment models.

5.2.1. *System identification*

The first step is to define the issue in the system for which we are building the potential scenarios. We then analyze the relevant threats of errors adding up, following the smartness framework Kalluri

et al. (2020). Within the *"real-time feedback control"* dimension, the erroneous state estimation indicates that the system cannot identify false data and report them on time (i.e., identification). This means that the filtering algorithm fails to prevent the attack (i.e., prevention). The false data in the sensor signals shows the need for minimum failure rates in sensors (i.e., safety). Within the *"level of automation"* dimension, the false data indicates the system's inability to gather meaningful information (i.e., information collection). The erroneous state estimation shows that the false data injection in several sensors suggests that the swarm is not capable of managing its mission (i.e., self-regulation). While erroneous control commands indicate the inability of the drones to make logical inferences (i.e., reasoning) and send safe control inputs to actuators (i.e., decision-making). The control system was not successful in understanding its environment and identifying erroneous data (i.e., sensing and context-awareness). Within the *"degree of cooperative control"* dimension, the state estimation is falsely affected by the operation of other interconnected components (i.e., co-regulation). Within the *"degree of integration"* dimension, we identify the issue of effective use of resources (i.e., self-optimization) due to the reliance of the system on a single-sensor input. Figure 4 illustrates a scenario of an emergence incident in a swarm of drones due to false data injection Gu et al. (2021).

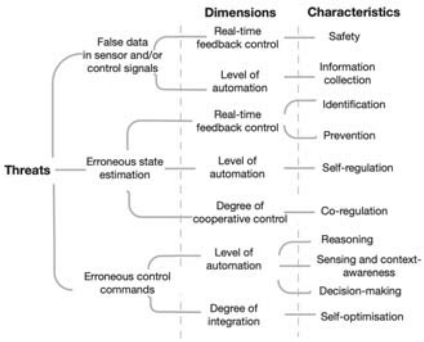

Fig. 4. False data injection threats linked to the framework of smartness dimensions and characteristics.

5.2.2. *Contributing factors identification and analysis*

The next step involves identifying the factors to be analyzed for each issue to predict possible future scenarios. Due to space limitations, we have selected three of the identified characteristics (i.e., safety, identification, and decision-making) to be mapped with their factors.

(1) In the context of false data, factors like "loss of safety margins" and "positive feedback" may have a greater impact on safety in a swarm of drones due to multiple interactions between non-linear drone systems.
(2) In the context of erroneous state estimation, factors like "information asymmetries" and "malicious motives" may contribute more to identification due to vulnerable dynamics behind the intentional undetected changes.
(3) In the context of erroneous control commands, factors like "temporal complications" and "communication" may have a greater effect due to temporal issues for changes in system structure over time.

5.2.3. *Scenario generation*

For analyzing emergent misbehavior, it is essential to predict the future states of the system based on the possible changes in existing variables. Based on the previous analysis, we provide some future risk scenarios extrapolated from the identified contributing factors. Some examples of scenarios assessing the effects of changing input variables are the following:

- S1: How would the errors be amplified from their intended malicious values if the system is non-linear?
- S2: How would the estimated state change if a sophisticated threat agent acted stealthily against the system?
- S3: How would the manipulated state variables be affected if bad data detection was implemented?

5.3. *Comparative analysis*

To address emergent risks in our context (i.e., a swarm-of-drones), we qualitatively evaluated the contributing factors using the four risk analysis methods. A summary of the comparative analysis of the pros and cons of each approach for predicting emergent properties is shown in Table 1.

5.3.1. *F1 - Scientific unknowns*

Dealing with emergent misbehavior requires preparing for unknown, unanticipated risks. The term "*scientific unknowns*" includes the degree of **uncertainty** arising from emergence due to insufficient available information in risk analysis. Therefore, the likelihoods of tractable unknowns should be included in assessments.

The unified framework for risk analysis can perform uncertainty analysis of causes and consequence analysis addressing the uncertainties Aven (2007). BBN provides good quantitative capabilities based on Bayesian inference, which can identify conditions under uncertainty Fenton and Neil (2018). However, there are methods (for example,

STPA-SafeSec and FMEVA) that cannot handle uncertainty or inaccuracy due to their qualitative data nature.

5.3.2. *F2 - Loss of safety margins*

Due to increased interdependency and connectivity, systems are becoming more tightly coupled, because the interactions among system components are non-linear. A change in one component can quickly have a strong impact on the related components Perrow (2011). Therefore, the **tight coupling** in a system leads to a loss of safety margins, increasing the likelihood of emergent risks.

To deal with complexity, STPA-SafeSec analyzes not only system losses caused by single components failures and/or threats, but also analyzes the non-linear interactions among the components. However, FMEVA deals with component-based incidents only, without considering their interactions Erik Nilsen et al. (2018). The unified framework for risk analysis uses event and attack trees likely to describe likely failures and threats in linear systems without considering complex interactions. Similar, BBN cannot identify likelihoods of dependency relations between system components Fenton and Neil (2018).

5.3.3. *F3 - Positive feedback*

Feedback loops are essential to the proper functioning of systems. When positive feedback dominates within a system by amplifying a perturbation, this tends to be a destabilizing emergent factor. Thus, it is important for analysts to identify the **feedback dynamics** in system structure and assess their function under different conditions.

Through the interaction of positive and negative feedback loops, STPA-SafeSec can explain the system dynamics related to changes in behavior over time. However, some methods have limited capacity for dynamic modeling of complex systems (for example, BBN and FMVEA). The unified framework for risk analysis also does not consider dynamic analysis because it uses event and fault trees that depend on their static nature Kriaa et al. (2015). It should be noted that the dynamic version of fault tree analysis could contribute significantly.

5.3.4. *F4 - Temporal complications*

Anticipating how an emergent risk will evolve, the **time course** can play a crucial role in the safe and secure system operation. The risk profile needs to address temporal issues.

The STPA-SafeSec provides a way to model dynamic processes and deal with structural dynamics related to system structure changes over time. But, the other methods (FMVEA, BBN, and the unified framework for risk analysis) are limited in the dynamic modeling of complex systems due to their static nature Kriaa et al. (2015). For instance, the BBN model is obtained manually based on the analyst's understanding of the system, and needs to be reconstructed to model different risk events or system changes.

5.3.5. *F5 - Communication*

Different types of communication, both external and internal, are essential factors for assessing emergent misbehavior. Besides the information exchanged internally in the system, **environmental influences** can demonstrate how emergent risks are amplified through external communication issues. To achieve effective communication about emergent risks, the process of bottom-up learning can be more useful than the top-down approach. As Bonabeau (2002) says, traditional top-down approaches fail to explain the behavior of emergent phenomena. Since bottom-up is starting from the local interactions of individuals to conclude about the group behavior.

The unified framework for risk analysis uses both forward and backward search methods Aven (2007). A backward search, which is a top-down approach, starts with the undesirable state and identifies its causes. A forward search, a bottom-up approach, starts from the component-based initiating events and uses them to develop scenarios. BBN is characterized by bottom-up or diagnostic inference, aiming to determine the impact of failures in low-level equipment or software systems Fenton and Neil (2018). FMEVA is an approach suitable for system design, focusing on components, and can be used for early analysis Erik Nilsen et al. (2018). However, top-down approaches like STPA-SafeSec, while able to examine control actions under different possible conditions, might lack detail because they focus on a conceptual mission Kaneko et al. (2018).

5.3.6. *F6 - Information asymmetries*

Autonomous decision-making enhances information asymmetries, where useful information may not be equally distributed among the system's parts. Systems with **autonomous capabilities** that think by themselves but act collectively can amplify an emergent risk's likelihood or severity. Identifying and evaluating these asymmetries caused by failures or intentional concealment should be considered in a risk assessment process.

However, STPA-SafeSec does not consider elements with autonomous capabilities within a system that might hide information (e.g., neural-network-based control software). Similarly, FMVEA is unable to analyze systems with learning capability due to their black-box nature Erik Nilsen et al. (2018). BBN and the unified framework for risk analysis can generate an uncertainty assessment where the probabilities could be conditioned based on the influence of information partially available to attackers Aven (2007).

Table 1. A comparison of risk analysis approaches with the contributing factors to assess the emergence in an autonomous system.

Approach	F1	F2	F3	F4	F5	F6	F7
STPA-SafeSec	-	++	++	++	+	-	+
FMVEA	-	-	-	-	+	-	+
BBN	++	-	-	-	+	+	+
Unified framework for risk and vulnerability analysis	++	-	-	-	++	+	+

Note: In Table 1, ++ (+, -) indicates that the particular method addressed thoroughly (addressed partially, did not address) the corresponding emergence factor.

5.3.7. *F7 - Malicious motives and acts*

The growing use of ICTs and AI is leading to increased **emerging vulnerabilities** to malicious acts in autonomous systems Kaloudi and Li (2020). The automation of the attack process allows conducting attacks at a wider scope, faster speed, and larger scale. Thus, risk processes should consider both known and novel threats.

Methods like BBN and the unified framework for risk analysis, which use probabilistic models, have the potential to ignore conditions that have a low probability Silva Castilho (2019). This assumption of independence in estimations might leave out stealthier threats. The STPA-SafeSec does not adequately capture a potential attacker's view, focusing only on system vulnerability, without analyzing threats leading to the activation of unsafe control actions Kaneko et al. (2018). Meanwhile, FMVEA considers only threats based on the STRIDE model and how an individual component could potentially be misused Erik Nilsen et al. (2018).

6. Discussion

As systems are getting more interconnected with more autonomous capabilities, undesirable emergent behavior make failures and attacks more broad and serious. Despite a significant amount of research on ways to analyze and predict emergent phenomena, there has been little research on risk analysis of emergent misbehavior, especially from the security point of view Husted and Myers (2014). Risk analysis can be a valuable tool to support risk-based decisions for controlling or eliminating emergent misbehavior. In particular, emergence risk analysis can help (i) identify the potential causes leading to consequences and/or predict possible consequences from a cause, and (ii) quantify the degree of risk associated with the undesirable emergence. Consequently, the system's level of trustworthiness can be increased by considering emergent risks in the risk management process.

To help readers select the most appropriate approach out of several risk causal analysis approaches, we selected the factors critical for assessing emergence and evaluated their fulfillment based on the literature findings. It is often difficult to use conventional methods to determine causation between the source(s) of the emergent risk and its consequences. These methods need to be improved by better understanding the process, by considering the systemic factors and the effects of multiple initiating events on non-linear component interactions, and by quantifying the impact of the disruptions caused by different types of adversaries. We believe that the approaches identified here can be used in a complementary manner to better analyze emergent misbehavior.

7. Conclusion and future work

As a result of the integration of new technologies, a system's emergence characteristics can create new types of risks called *emergent risks*, which affect both the system and society. Therefore, systems might face situations for which they are unprepared, significantly compromising their security, safety, and resilience. This paper analyzes the emergent risk concept and the pros and cons of four risk causal analysis approaches for better predicting future risks in autonomous systems. In future work, we intend to apply the most promising risk analysis methods to a specific case in order to compare their performance.

References

Allan, N., N. Cantle, P. Godfrey, and Y. Yin (2013). A review of the use of complex systems applied to risk appetite and emerging risks in erm practice: Recommendations for practical tools to help risk professionals tackle the problems of risk appetite and emerging risk. *British Actuarial Journal*, 163–269.

Aven, T. (2007). A unified framework for risk and vulnerability analysis covering both safety and security. *Reliability engineering & System safety 92*(6), 745–754.

Axelrod, C. W. (2013). Managing the risks of cyber-physical systems. In *2013 IEEE Long Island Systems, Applications and Technology Conference (LISAT)*, pp. 1–6. IEEE.

Bonabeau, E. (2002). Predicting the unpredictable. *Harvard Business Review 80*(3), 109–116.

Council, N. R. (2001). *Embedded, everywhere: A research agenda for networked systems of embedded computers*. National Academies Press.

Dallard, P., T. Fitzpatrick, A. Flint, A. Low, R. R. Smith, M. Willford, and M. Roche (2001). London millennium bridge: pedestrian-induced lateral vibration. *Journal of Bridge Engineering 6*(6), 412–417.

De Carvalho, P. V. R. (2011). The use of functional resonance analysis method (fram) in a mid-air collision to understand some characteristics of the air traffic management system resilience. *Reliability Engineering & System Safety 96*(11), 1482–1498.

Erik Nilsen, T., J. Li, S. O. Johnsen, and J. A. Glomsrud (2018). Empirical studies of methods for safety and security co-analysis of autonomous boat. *Safety and Reliability-Safe Societies in a Changing World*.

Fenton, N. and M. Neil (2018). *Risk assessment and decision analysis with Bayesian networks*. Crc Press.

Ferreira, S., M. Faezipour, and H. Corley (2013). Defining and addressing the risk of undesirable emergent properties. In *2013 IEEE International Systems Conference (SysCon)*, pp. 836–840. IEEE.

Friedberg, I., K. McLaughlin, P. Smith, D. Laverty, and S. Sezer (2017). Stpa-safesec: Safety and security analysis for cyber-physical systems. *Journal of information security and applications 34*, 183–196.

Fromm, J. (2005). Types and forms of emergence. *arXiv preprint nlin/0506028*.

Graham,
J. D., H. Fineberg, D. Helbing, T. Homer-Dixon, W. Kröger, M. Maila, J. McNeely, S. Michalowski, E. Millstone, M. Wilson, et al. (2010). *The Emergence of Risks: Contributing Factors*. Number REP_WORK. International Risk Governance Council (IRGC).

Gu, Y., X. Yu, K. Guo, J. Qiao, and L. Guo (2021). Detection, estimation, and compensation of false data injection attack for uavs. *Information Sciences 546*, 723–741.

Han, S. Y. and D. DeLaurentis (2013). Development interdependency modeling for system-of-systems (sos) using bayesian networks: Sos management strategy planning. *Procedia Computer Science 16*, 698–707.

Huang, H.-M. (2004). Autonomy levels for unmanned systems (alfus) framework volume i: Terminology version 2.0.

Husted, N. and S. Myers (2014). Emergent properties & security: The complexity ofsecurity as a science. In *Proceedings of the 2014 New Security Paradigms Workshop*, pp. 1–14.

Jovanovic, A. S. (2019). Managing emerging risks for enhanced resilience: Aligning approaches internationally. Proceedings of the 29th European Safety and Reliability Conference. Research Publishing, Singapore.

Kalluri, B., C. Chronopoulos, and I. Kozine (2020). The concept of smartness in cyber–physical systems and connection to urban environment. *Annual Reviews in Control*.

Kaloudi, N. and J. Li (2020). The ai-based cyber threat landscape: A survey. *ACM Computing Surveys (CSUR) 53*(1), 1–34.

Kaneko, T., Y. Takahashi, T. Okubo, and R. Sasaki (2018). Threat analysis using stride with stamp/stpa. In *The international workshop on evidence-based security and privacy in the wild*.

Kavallieratos, G., S. Katsikas, and V. Gkioulos (2020). Cybersecurity and safety co-engineering of cyberphysical systems—a comprehensive survey. *Future Internet 12*(4), 65.

Kosow, H. and R. Gaßner (2008). *Methods of future and scenario analysis: overview, assessment, and selection criteria*, Volume 39. DEU.

Kriaa, S., L. Pietre-Cambacedes, M. Bouissou, and Y. Halgand (2015). A survey of approaches combining safety and security for industrial control systems. *Reliability engineering & system safety 139*, 156–178.

Kushner, D. (2013). The real story of stuxnet. *ieee Spectrum 50*(3), 48–53.

Lacagnina, M. (2009). Midair over the amazon. *AeroSafety World*, 11–15.

Manky, D. (2018). Order vs. mad science analyzing black hat swarm intelligence. RSA Conference.

MITRE (2018). Treating systems of systems as systems.

Mogul, J. C. (2006). Emergent (mis) behavior vs. complex software systems. *ACM SIGOPS Operating Systems Review 40*(4), 293–304.

Perrow, C. (2011). *Normal accidents: Living with high risk technologies-Updated edition*. Princeton university press.

Purdy, G. (2010). Iso 31000: 2009—setting a new standard for risk management. *Risk Analysis: An International Journal 30*(6), 881–886.

Schmittner, C., Z. Ma, and P. Smith (2014). Fmvea for safety and security analysis of intelligent and cooperative vehicles. In *International Conference on Computer Safety, Reliability, and Security*, pp. 282–288. Springer.

Silva Castilho, D. (2019). *Active STPA: integration of hazard analysis into a Safety Management System Framework*. Ph. D. thesis, Massachusetts Institute of Technology.

Taleb, N. N. (2007). Black swans and the domains of statistics. *The American Statistician 61*(3), 198–200.

Williams, A. P. and P. D. Scharre (2015). *Autonomous systems: Issues for defence policymakers*. HQ SACT.