SpringerBriefs in Applied Sciences and Technology
Safety Management

Jean-Christophe Le Coze · Stian Antonsen *Editors*

Safety in the Digital Age

Sociotechnical Perspectives on Algorithms and Machine Learning

FONCSI
Foundation for an
Industrial Safety Culture

Springer

SpringerBriefs in Applied Sciences and Technology

Safety Management

The *SpringerBriefs in Safety Management* present cutting-edge research results on the management of technological risks and decision-making in high-stakes settings.

Decision-making in high-hazard environments is often affected by uncertainty and ambiguity; it is characterized by trade-offs between multiple, competing objectives. Managers and regulators need conceptual tools to help them develop risk management strategies, establish appropriate compromises and justify their decisions in such ambiguous settings. This series weaves together insights from multiple scientific disciplines that shed light on these problems, including organization studies, psychology, sociology, economics, law and engineering. It explores novel topics related to safety management, anticipating operational challenges in high-hazard industries and the societal concerns associated with these activities.

These publications are by and for academics and practitioners (industry, regulators) in safety management and risk research. Relevant industry sectors include nuclear, offshore oil and gas, chemicals processing, aviation, railways, construction and healthcare. Some emphasis is placed on explaining concepts to a non-specialized audience, and the shorter format ensures a concentrated approach to the topics treated.

The *SpringerBriefs in Safety Management* series is coordinated by the Foundation for an Industrial Safety Culture (FonCSI), a public-interest research foundation based in Toulouse, France. The FonCSI funds research on industrial safety and the management of technological risks, identifies and highlights new ideas and innovative practices, and disseminates research results to all interested parties.

For more information: https://www.foncsi.org/

FONCSI

Foundation for an
Industrial Safety Culture

Jean-Christophe Le Coze · Stian Antonsen
Editors

Safety in the Digital Age

Sociotechnical Perspectives on Algorithms and Machine Learning

Springer

Editors
Jean-Christophe Le Coze
French National Institute for Industrial
Environment and Risks (Ineris)
Verneuil-en-Halatte, France

Stian Antonsen
Department of Industrial Economics
and Technology Management
NTNU Social Research
Trondheim, Norway

ISSN 2191-530X ISSN 2191-5318 (electronic)
SpringerBriefs in Applied Sciences and Technology
ISSN 2520-8004 ISSN 2520-8012 (electronic)
SpringerBriefs in Safety Management
ISBN 978-3-031-32632-5 ISBN 978-3-031-32633-2 (eBook)
https://doi.org/10.1007/978-3-031-32633-2

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Chapter 1
Safety in a Digital Age: Old and New Problems—Algorithms, Machine Learning, Big Data and Artificial Intelligence

Jean-Christophe Le Coze and Stian Antonsen

Abstract Digital technologies including machine learning, artificial intelligence and big data are leading to dramatic changes, in both the workplace and our private lives. These trends raise concerns, ranging from the pragmatic to the philosophical, regarding the nature of work, the professional identity of workers, our privacy, the distribution of power within organizations and societies. They also represent both opportunities and challenges for the work of producing safety in high-hazard systems. We highlight a number of pressing issues related to these evolutions and analyze the extent to which existing lenses from sociotechnical theory can help understand them.

Keywords Machine learning · AI · Algorithmic management · Regulation · Safety

This introduction is based on a slightly modified version of the call for the NeTWork workshop on "*safety in the digital age*". The call framed the invitation of researchers to debate then to write a chapter for a book. The intention of this workshop then of this publication was to start a collective discussion, based on empirical and conceptual reflection, about safety in this new stage of societies' trajectories, commonly described as "*the digital age*". These chapters are followed by a conclusion which develops a sociotechnical proposition of how to start thinking "*safety in the digital age*"…

Stian Antonsen's contributions in the production of the book were funded with support from the Research Council of Norway (grant no. 303489).

J.-C. Le Coze
Ineris, France
e-mail: jean-christophe.lecoze@ineris.fr

S. Antonsen (✉)
NTNU Social Research, Trondheim, Norway
e-mail: stian.antonsen@samforsk.no

J.-C. Le Coze and S. Antonsen (eds.), *Safety in the Digital Age*,
SpringerBriefs in Safety Management,
https://doi.org/10.1007/978-3-031-32633-2_1

Algorithms, machine learning, big data and artificial intelligence (AI) are key words of a current transformation of societies. Following a first wave of internet development coupled with the spread of personal computers in the 1990s, the 2010s brought a second level of connectedness through smart phones and tablets, generating a massive amount of data from private and public activities. It is this new environment built over thirty years made of big data produced by the daily activities of people working, traveling, reading, buying, communicating and amplified and captured by a growing market of the Internet of Things—IoT, which provides an opportunity for the proliferation of algorithms, machine learning and a new generation of AI [11, 12]. Without falling into the trap of technological determinism, this transformation through digitalisation clearly affects every sphere of social life including culture, economy, science, politics, art, education, health, family, business, identity and social relations.

One can easily find examples in these different spheres through which our daily private and public lives are affected. Social media (e.g., Facebook, Twitter, LinkedIn, ResearchGate), search engines (e.g., Google, Bing, Qwant) and websites in so many various areas including online shopping (e.g., Amazon, Fnac), music (e.g., Spotify, Deezer), news (e.g., New York Times, Le Monde, Financial Times), videos, series, cinema and programmes (e.g., Netflix, YouTube, DailyMotion) or activism (e.g., SumOfUs, Avaaz) are only a few examples. Because these ubiquitous online services reconfigure our ways of listening to music, of consuming, of reading, of communicating, of learning, of creating…we simply experience new ways of being in the world.

A digital society slowly emerges, somewhere between (1) reality, (2) proclaimed bright futures and (3) fears of dystopian trends in the next years or decades to come. In the call for this NeTWork workshop on "Safety in the digital age", we wished to remain grounded in reality. It has become indeed very clear for sociologists that we now empirically live in a mediated constructed reality (e.g., Hepp and Couldry [3], and Cardon [4]), while some wonder if these changes should be characterised as an evolution or a revolution (e.g., Rieffel [14]), others now warn of a re-engineering of humanity because of the extent of the material, cognitive and social modifications of our environment (e.g., Frischmann and Sellinger [7]).

In this respect, the rise of internet giants (GAFAM/N for Google, Apple, Facebook, Amazon, Netflix, Microsoft) triggers several concerns ranging from business monopoly through fiscal to data privacy and exploitation issues, which reveal increasing concern by civil societies and states. The thesis of a "*surveillance capitalism*" by Zuboff comes to mind [17], a thesis based on the careful study of the ideologies professed by the engineers behind the digital world. An example is selected by [17, p. 432], quoting Pentland, an MIT Professor.

> Pentland says that "continuous streams of data about human behaviour" mean that everything from traffic, to energy use, to disease, to street crime will be accurately forecast, enabling a « world without war or financial crashes, in which infectious disease is quickly detected and stopped, in which energy, water and other resources are no longer wasted, and in which governments are part the solution rather than part of the problem (…) Great leaps in health care, transportation energy, and safety are all possible.

Narrowing this panoramic view to work, organisations, business and regulations, the implications are potentially quite profound. They seem obvious in some cases but remain also still partly uncertain in other areas. For instance, how much of work as we know it will be changed in the future? Estimates range from 9 to 47% of current jobs that could disappear within the next few decades because of AI. Whatever the extent of this replacement or mutation, one can imagine that combining human jobs with AI, or simply relieving people from current tasks, will change the nature of work as well as the configuration and management of organisations. In addition, with this digital expansion comes growing cyber-security challenges.

In the platform, digital and gig economy (e.g., Amazon, Uber, Deliveroo, Airbnb), *algorithmic management* has for instance been coined to characterise employees' working conditions [9]. And some of these companies' practices have already been met by workforce resistance in several countries, a workforce fighting for what they consider to be their rights as employees. In many cases, in the US, the UK and France, the legal system has ruled favourably concerning workers' claims that they were in a traditional employer–employee relationship, and not in a context of companies contracting with self-employed workers.

Businesses are threatened in their market positions by innovative ways of interacting with their customers through social media and use of data, by new ways of organising work processes or by new start-up competitors redefining the nature of their activities. Consider, for example, the prospect of autonomous cars, which could completely redefine the ecosystem of companies. Car makers could well become secondary players of an industry revolving around data exploitation and management controlled by digital companies, which become the new dominant players. The insurance business could well fall into the hands of these new data masters too, in the same way as hotels chains had to cope with new digital players. Business leaders must therefore adapt to this digitalisation of markets, to potential disruptions based on big data, machine learning and AI. They must strategise to keep up with a challenging and rapidly changing environment [5].

The same applies to regulation. Because of the now pervasive use of algorithms, machine learning, big data and AI across society, notions of *algorithmic governmentality* [15] or *algorithmic regulation* [16] have been developed to identify and conceptualise some of the challenges faced by regulators. Cases of *algorithmic biases*, *algorithmic law breaking*, *algorithmic propaganda*, *algorithmic manipulation* but also *algorithmic unknowns* have been experienced in the recent past, including the Cambridge Analytica/Facebook scandal during the last US election and the "Diesel-Gate" triggered by Volkswagen's software fraud [2]. This creates new challenges for the control of algorithms' proliferation, and some have already suggested, in the US, a National Algorithm Safety Board (e.g., Macaulay [10]).

This last point connects digitalisation with safety. How can high-risk and safety–critical systems be affected by these developments, in terms of their activities, their organisation, management and regulation? What can be the safety-related impacts of the proliferation of big data, algorithmic influence and cyber-security challenges in healthcare (e.g., hospitals, drugs), transport (e.g., aviation, railway, road), energy production/distribution (e.g., nuclear power plants, refineries, dams, pipelines, grids) or production of goods (e.g., chemicals, food) and services (e.g., finance, electronic communication)? Understanding how these systems operate in this new digital context has become a core issue. It is the role of research to offer lenses through which one can grasp how such systems evolve, and the implications for safety.

There are many affected areas in which research traditions in the safety field can contribute to question, to anticipate and to prevent potential incidents but also to support, to foster and to improve safety performance within a digital context [1, 8]. For instance, tasks so far performed by humans are potentially redesigned with higher levels of AI-based automation, whether in the case of autonomous vehicles or human–machine teaming [12]. What about human error, human–machine interface design, reliability and learning in these new contexts (Smith and Hoffmann 2017)? What are the consequences of pushing the boundaries of allocated decision making towards machines? What are the implications for the distribution of power and decision-making authority of using new sources of information, new tools for information processing and new ways of "preprogramming" actions and decisions through algorithms?

The same applies to the organisational or regulatory angles of analysis of safety critical systems, such as those developed by the high-reliability organisation [13] and risk regulation regimes [6] research traditions. What happens when protective safety equipment, vehicles, individuals' behaviours and the automation of work schedules are interlinked through data and algorithmic management delegating to machines a new chunk of what used to be human decision making? What are the implications for risk assessment, learning from experience or compliance to rules and regulations, including inspection by authorities?

But quite importantly, what of this is realistic and unrealistic? What can be anticipated without empirical studies but only projections into the future? Which of these problems are new ones and which of them are old? The NeTWork workshop in September 2021 was an opportunity to map some of the pressing issues associated with digitalisation based on algorithms, machine learning, big data and artificial intelligence for the safe performance of high-risk systems and safety–critical organisations. The chapters in this book cover many of the hot issues one needs to have in mind when operating, managing and regulating safety in a digital age. They offer a unique treatment of this topic, one of the first to bring multiple disciplinary viewpoints to bear. Each chapter is now summarised to allow the reader to get a big picture of the multiple angles of analysis explored.

In "*The digitalisation of risks assessment: fulfilling the promises of predictions?*", David Demortain introduces risk assessment in risk regulation regimes. He reminds the reader of the importance of this activity at the intersection of private companies, states, civil society, science and expertise in a variety of sectors such as the food, nuclear or pharmaceutical ones. Assessing risks consists in building mathematical models which translate phenomena into equations in order to anticipate their effects. The relationship between data, experiments, computers and models is key to an understanding of risk assessment. David describes three such models when it comes to predicting the impact of chemicals on the living (quantitative structure–activity relationships—QSAR; physiologically based pharmacokinetics—PPBK; biologically based dose response—BBDR). These models rely on different epistemological, methodological, experimental and mathematical options to support their predictive capabilities. Already extensively computerised models, the addition of machine learning, big data and artificial intelligence proves to be a new exciting prospect for promoters of increasingly sophisticated models, an example of which is the Tox21 program. David discusses digitalisation in this context by considering critically, in turn, what appears, according to him and at this stage, realistic, and what is not. He ponders the excessive ambitions surrounding datafication, computational innovation and the systemic ambition of models.

In "*Key dimensions of algorithmic management, machine learning and big data in differing large sociotechnical systems, with implications of system wide safety management*", Emery Roe and Scott Fortmann-Roe translate the problem of safety in the digital age at the empirical level of software design strategies in distributed-type companies (such as Google, Netflix, Facebook or Amazon). These strategies are ones based on design trade-offs of software along four dimensions: (1) comprehensibility versus features; (2) human operated versus automated; (3) stability versus improvement and (4) redundancy versus efficiency. The fast pace of digital innovation pushes such companies towards the right end of this design spectrum, towards more features, automation, improvement and efficiency. From a safety point of view, the traditional approach favours the opposite end of this spectrum, preferring comprehensibility, human operated, stable and redundant systems. Emery and Scott challenge these taken for granted design assumptions, considering the problem of obsolescence (outdated software systems), when a system falls behind, becomes too rigid in its evolving environment and exposed, additionally, to high levels of cyber-security threats.

Olivier Guillaume illustrates with a case study the privacy aspect of the digital age in his chapter "*digitalisation, safety and privacy*". He first situates the advocated value of the digital by its promoters in the context of work in organisations. Indeed, the digital age recasts the old problem of autonomy, professionalism, standardisation, bureaucracy and its relation to safety. In principle, by providing more efficient ways, through smart phones, personal digital assistants (PDAs), connected glasses and wearable sensors to plan, track, monitor and control employees' activities, a greater level of reliability and safety could be achieved for managers. However, tracking employees' activities is regulated by the European directive on data protection (GDPR) and is met anyway by employees' reluctance regarding intrusive management. Olivier shows how privacy, intimacy and private life in employees' daily activities play an important role in the construction of professional and collective identity as well as expertise. In his case study, digital solutions which impinge on privacy are negotiated, and employees obtain from their employer, through their representatives, a decision to abandon options that they consider to be intrusive. Olivier warns that in work contexts without a tradition of negotiations, or exposed to high levels of power asymmetry, the balance between digital control and employees' privacy might lean towards the former at the expense of the latter.

Cécile Caron continues the discussion of privacy aspects of the digital age, in her chapter *Design and dissemination of blockchain technologies: the challenge of privacy*. She takes as her starting point the antagonism between the ideals of blockchain as an information infrastructure that is decentralised and without the need of a trusted third party, and the GDPR regulation's requirements to have a (centralised) data controller responsible for the processing of personal data. Being based on two different forms of trust, the relationship between the two presents important privacy dilemmas that will need some form of reconciliation in concrete application of blockchain technologies. Caron studies this by means of a sociological case study of a mobility service using a blockchain, IoT solutions and mobile and web applications to track the charging of electrical vehicles. By analysing qualitative data from service designers and service users, Caron identifies different themes or "tests" that illustrate the confrontations, negotiations and alliances involved in the tension between privacy and blockchains. Among these themes are the crucial question of balancing decentralisation and (re)centralisation in the governance of privacy, the requirements for data minimisation, consent and transparency in the processing of personal data. The three dilemmas identified in the case study illustrate that concrete practices of privacy protection are by and large a skill or a form of expertise that is distributed among a wide range of actors in innovation ecosystems. The ability to find satisfactory compromises across these actors requires a high level of collaboration and experimentation.

In *Considering severity of safety–critical outcomes in risk analysis: An extension of fault tree analysis*, David Kaber and colleagues draw our attention to the input data of risk analysis. Despite increases in available data in some domains, other domains are still characterised by an absence of empirical observations. Hence, there are situations, particularly in novel work systems, where the data is sparse relative to the number of decision variables that must be considered in risk analysis and safety practice (the "curse of dimensionality"). They ask whether new and advanced tools can be established to create precise projections of safety–critical system outcomes in such situations and describe and discuss a method for accomplishing such projections. The authors also discuss the extension of existing systems safety analysis methods into more digitalised industries, and the crucial role of the quality and quantity of input data in such methods.

Nicola Paltrinieri in some ways picks up where Kaber and colleagues leave off, in his chapter *Are we going towards "no-brainer" safety management.* Where Kaber and colleagues focus on the methods used to provide analyses, Paltrinieri emphasises the role of humans in interpreting the *results* of such methods. He shows how increases in available data and enhanced computational power can in fact be utilised for more continuous monitoring of industrial process conditions but, as the title of the chapter suggests, the safety management of Industry 4.0 is still dependent on human judgement. By providing examples of AI-based prediction in three domains (release of hazardous materials in land-based industry, accidental drive-offs in offshore drilling operations and alarm chattering in a chemical plant), he shows that the predictions in all three cases still need to be interpreted and that we are nowhere near the condition of autonomous safety management. In addition, and like Kaber and colleagues, Paltrinieri points to the critical role of input data for making predictions and the equally critical role of human judgement in preparing some types of data for analysis.

Turning to the health sector, Mark Sujan presents two examples of the use of AI in his chapter *Looking at the safety of AI from a systems perspective.* In the two examples (autonomous infusion pumps for intravenous medication administration and AI support in the recognition of out-of-hospital cardiac arrest), Sujan explains the specific functions of the two systems and relates these functions to their social and professional contexts. He shows that many of the challenges are highly familiar to safety researchers, such as the "ironies of automation" and the potential for "automation surprise". Still, modern AI systems also pose new challenges, in that these systems are not necessarily put in place to replace physical work but rather to augment human actions. This involves AI systems being given different roles compared to traditional automation and a different form of interaction between humans and technology. Sujan argues that these relationships between humans and technology, and the associated social, cultural and ethical aspects, will have greater importance for future AI applications than was the case with traditional automation. This, Sujan argues, calls for a transition from a technology-centric focus that contrasts people and AI, to a more systems-based approach where AI and humans are seen as integrated in a wider health system.

In *Normal cyber-crisis*, Sarah Backman provides a high-level, yet empirically grounded, discussion of the phenomenon of large-scale cyber-crises that can affect the functioning of critical infrastructures. Based on interviews with senior experts on cyber-security and critical infrastructures in Sweden, the UK and the USA, she argues that the consequence dynamics of such crises can be explained by Charles Perrow's Normal Accident framework. She shows how the transboundary nature of large-scale cyber-crises needs to be understood through several layers: (1) the technical layer, especially emphasising the role of legacy software and hardware, (2) the cognitive layer, referring to the difficulties of perceiving and recognising dangers when tight couplings and interactive complexity is a transboundary phenomenon, (3) the organisational layer, how centralisation can make accidents more consequential, while redundancy serves to create looser couplings and (4) the macro-layer, illustrating how supply chains can be exploited by cyber-threat agents.

Picking up some of the threads from Backman's chapter, Nævestad and colleagues examine how critical infrastructure organisations can reduce their digital vulnerability. The starting point of their chapter, *Information security behaviour in an organisation providing critical infrastructure: A pre-post study of efforts to improve information security culture*, is that people can be both a cause of information security incidents and a key element in protecting a system against such incidents. They examine the effects of interventions aimed at improving information security culture, with an aim to ultimately influence behaviour related to information security. By means of a multivariate regression analysis of survey data consisting of employees' perceptions of key dimensions of information security, and controlling for education, seniority, prior knowledge and which department the respondents belonged to, they find information security culture to be the most important variable influencing information security behaviour.

Yann Ferguson discusses how the introduction of artificial intelligence in the workplace can influence the empowerment and productivity of workers, including the preservation of job quality, inclusiveness, health and safety. His chapter is titled *AI at work, working with AI—First lessons from real use cases*. Based on 150 use cases of a specific application of AI five ideal-type "worker stories" are crystallised, all describing potential outcomes of the use of AI in the workplace. AI can involve employees being both *replaced*, *dominated, augmented*, *divided* and *rehumanised*. All these ideal types are viable outcomes from the introduction of AI in the workplace. However, which of these consequences, or which combination of them, was not only a matter of the technology itself but was strongly shaped by characteristics of the work and workers involved. Although not in a deterministic way, the application of AI was associated with a reconfiguration of work and the form of engagement between workers, work and the technology involved.

References

1. Almklov, Antonsen, Størkersen, Roe, Safer societies. Safe. Sci. **110**(Part C) (2018)
2. L. Andrews, Algorithms, governance and regulation: beyond 'the necessary hashtags', 2017, retrieved in October 2019 at https://www.kcl.ac.uk/law/research/centres/telos/assets/DP85-Algorithmic-Regulation-Sep-2017.pdf
3. D. Cardon, *Culture numérique* (Presses de Science Po, Paris, 2019)
4. N. Couldry, A. Hepp, *The mediated construction of reality* (Polity Press, Cambridge, UK, 2017)
5. C. Deshayes, *La transformation numérique et les patrons* (Paris, Presses des Mines, Les dirigeants à la manoeuvre, 2019)
6. P. Drahos (ed.), *Regulatory theory: foundations and applications* (ANU Press, Acton, 2017)
7. B. Frischmann, E. Sellinger, *Re-engineering humanity* (Cambridge University Press, Cambridge, 2018)
8. J.-C. Le Coze (ed), in *Safety science research.* Evolution, challenges and new directions (CRC Press, Taylor & Francis group, Boca Raton, FL, 2020)
9. M. Lee, K. Kusbit, D. Metsky, E. Dabbish, Working with machines: the impact of algorithmic, data-driven management on human workers (2015). Retrieved in October 2019 at https://www.cs.cmu.edu/~mklee/materials/Publication/2015-CHI_algorithmic_management.pdf
10. T. Macaulay, Pioneering computer scientist calls for National Algorithm Safety Board. Techworld, 31 May (2017). Retrieved in October 2019 at http://www.techworld.com/data/pioneering-computer-scientist-calls-for-national-algorithms-safety-board-3659664
11. NSTC, Big data: a report on algorithmic systems, opportunity, and civil rights. Executive Office of the president, 2016a. Retrieved in October 2019 at https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf
12. NSTC, Preparing for the future of AI. Executive Office of the president, 2016b. Retrieved in October 2019 at https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf
13. R. Ramanujam, K.H. Roberts (eds.), *Organizing for reliability: a guide for research and practice* (Stanford University Press, Stanford, CA, 2018)
14. R. Rieffel, *Révolution numérique, révolution culturelle?* (Gallimard, Paris, 2014)

15. A. Rouvroy, T. Berns, Gouvernementalité algorithmique et perspectives d'émancipation. Le disparate comme condition d'individuation par la relation ? Réseaux **1**(177), 163–196 (2013)
16. K. Yeung, Algorithmic regulation. a critical interrogation. Regulation Governance **12**, 505–523 (2017)
17. S. Zuboff, The age of surveillance capitalism: the fight for a human future at the new frontier of power. Public Affairs (2019)

Chapter 2
The Digitalisation of Risk Assessment: Fulfilling the Promises of Prediction?

David Demortain

Abstract Risk assessment is a scientific exercise that aims at anticipating hazards. Prediction has always been a rallying call for the scientists that gave birth to this interdisciplinary movement in the 1970s. Several decades later, the broad movement of digitalisation and the promises of artificial intelligence seem to be pushing the limits of risk assessment and herald an era of faster and more precise predictions. This chapter briefly reviews the history of chemical risk assessment methods developed by regulatory bodies and associated research groups, and the complex ways it has digitalised. It unpacks digitalisation, to probe how its various aspects—datafication, computational innovation and modelling theories—align to meaningfully transform it, and determine whether the ever-revamped technological promise of prediction is within a closer reach than it was before.

Keywords Computational risk assessment · Regulation · Digitalisation · Modelling

2.1 Introduction

Risk assessment is a scientific exercise that aims at anticipating hazards. By convention, it entails a specification of this hazard, the collection of data about known occurrences and various calculations allowing to extrapolate the frequency, severity, probability of future occurrences for particular persons or organisations from a baseline of data.

Risk assessment has turned out to be constitutive of a type of regulation, known as risk regulation [1, 2], [3]. It took form progressively in the 1970s and 1980s thanks to the contributions of a series of more established disciplines such as actuarial sciences, geography and natural disaster research, physics, operational research, all

D. Demortain (✉)
Laboratoire Interdisciplinaire Sciences Innovations Sociétés (LISIS), INRAE, CNRS and Univ. Eiffel, Marne-La-Vallée, France
e-mail: david.demortain@inrae.fr

J.-C. Le Coze and S. Antonsen (eds.), *Safety in the Digital Age*,
SpringerBriefs in Safety Management,
https://doi.org/10.1007/978-3-031-32633-2_2

of which converged towards the notion that the probability of hazardous events could be computed, and they may be prevented thanks to these calculations. Prediction has been one of the rallying calls for the scientists that united to give birth to the interdisciplinary movement of risk assessment [1, 4, 5].

Several decades later, the broad movement of digitalisation and the promises of artificial intelligence seem to be pushing the limits of risk assessment and herald an era of faster and more precise predictions. The conversion of more diverse and larger sets of information into storable, classifiable and analysable digital form, and the design and adoption of IT technologies allowing organisations to perform such tasks at a quick pace and minimal cost revive the ambitions of risk assessors to anticipate risks with precision and reliability. And indeed, most risk assessment practitioners have joined the call to accelerate and deepen the movement of digitalisation, embracing like many other sciences the age of big data [6]. The imaginary of continuous, non-human-mediated production of data to train and feed predictive machines [7], and quickly discover new cause-and-effect relationships in complex systems, has penetrated risk assessment and risk analysis [8].

Digitalisation, however, is a complex of several intertwined transformations. It entails a phenomenon of datafication (the generation of data about an increasing diversity of organisational activity), of computational innovation (the introduction of new computing technologies and infrastructures) and of theoretical modelling of the world (with the rise, notably, of a systems-vision). This chapter briefly reviews the history of chemical risk assessment methods developed by regulatory bodies and associated research groups, and the complex ways in which this would-be science has digitalised over the years. It does so to identify what is currently happening in this area and to better determine whether the ever-revamped technological promise of prediction is within closer reach than it was before. Clearly, digitalisation runs through the history of risk assessment. But the computing technologies available, the data generated and the theoretical visions thanks to which we can make sense of these data and turn them into meaningful predictions, are perhaps more aligned nowadays than they used to be, producing this sense of a fast and deep transformation of technologies for knowing what is safe.

2.2 Assessing and Computing Risks

From aircraft to chemicals plants, through food ingredients and chemicals, most of the technologies that are recognised as potentially hazardous are submitted to some form of risk assessment. Risk assessment is the informational element of risk regulation regimes [2]. It involves dedicated techniques and routine processes, through which the conditions of appearance of hazards may be determined (their frequency, severity, publics and places most affected…), and corresponding regulatory controls legitimately decided.

It can be applied ex ante to the development of the technology in question, informing the decisions to put it on the market or in use more generally. It may also take place alongside the use of the technology in question (it is then called monitoring, surveillance or vigilance). However, the epistemic and regulatory ideal behind risk assessment is that of the prediction of hazards before they occur (Shapiro and Glicksman 2003).

Risk assessment is a process that has always made massive use of science and particularly of modelling. Chemicals cannot or must not be tested directly in the human body, in the environment or in the industrial conditions in which they will be used, but are on the contrary tested in experimental, simplified conditions. The toxicity of the chemical is tested on animals (and toxicologists and chemists speak, tellingly, of the various "animal models" (i.e., species) that can be used to perform these experiments) or in vitro. In this sense, risk assessment is a model-based science, if we understand models as a simplified, scale-reduced analogue, or representation, of a system.

Modelling is a process that consists in formulating a series of equations to capture the functioning of a system, and informing the parameters of the equations with various measurements and data, in such a way that various states of the system may be simulated. Modelling allows extrapolating from the data points available (of which there may be a relative paucity), to other situations, scales and time periods. It is closely dependent on current knowledge of a system and on the capacity to imagine risks and accidents occurring within that system [9]. Risk assessment follows a systems perspective. It simplifies systems and the behaviour of agents in this system. In practice, it follows from the social construction of a "risk object" [10]: a technological element excerpted from this system, to which a number of potential hazardous effects may be attributed.

2.3 Layers of Transformation: A Historical Perspective on Digitalisation

The field of risk assessment broadly evolves towards an ideal of continuous modelling of large sets of data, to analyse and simulate processes at various scales of a system; a sort of integrated form of simulation, where one aims to describe and predict a greater number of aspects of a system, at a fine-grained level [11]. This broad ambition, however, concatenates several transformations that have affected risk assessment since it emerged four decades ago: the material capacity to generate data in greater quantities and great variety, thanks to the diffusion of sensors across the environment and living organisms, or datafication; the design and increasing use of new mathematical models to capture the complexity of systems and the occurrence of hazards within them; the rise of a complex-systems vision of things. Looking back at what has been developed in the field helps appreciate the path of technological development and epistemic change through which current applications have taken shape.

2.3.1 Mathematical Models: Technologies of Computing

The first computational tool used in this area was one that aimed to characterise the properties of molecules through systematic analysis of the relationships between their structure and their biological effects—so-called structure–activity relationships (SARs) [12]. A quantitative SAR is a statistical analysis of the biological activity of a group of two or more chemicals that have some structural similarity, as captured through a chosen descriptor of the chemical. The modelling of causal relations between chemical properties and biological impacts is rooted in fundamental chemistry. It rests on the conduct of multiple, strictly standardised experiments on molecules with the same kind of structure (cogeneric molecules). Once a sufficiently powerful set of data has been produced, a statistical analysis can be run, to capture the correlations between structural properties and the biological effects. The resulting correlations can then be used to formulate a mathematical equation—a model—to predict the effects of a molecule without physically testing it. The challenges that QSAR research is facing typically concern the generation of sufficiently large sets of comparable data across a whole class of chemicals (a highly intensive endeavour), and the availability of both training sets and alternative data sets to validate the models. Without such data, modellers end up producing an over-fitted or under-fitted model [13]. Connecting model development to larger sets of data made available by pharmaceutical companies is one of the key hopes here.

A second technique aimed at modelling dose–response relationships in biological organisms. The technique is known as physiologically based pharmacokinetics (PBPK). PBPK models consist in simplified descriptions of the physiological system exposed to a chemical substance. By modelling the organism and the biological mechanisms involved in the metabolism of the substance, one can compute the dose at which the substance will produce hazardous effects in the organism. Models represent relevant organs or tissues as compartments, linked by various flows (notably blood flows) in mathematical terms. The parameters are calibrated with data emerging from animal experiments or clinical observations. PBPK modelling really started in the 1970s, once sufficient data and computer tools became available to establish the doses at which anticancer medicines could be delivered to various organs. The application of PBPK to industrial chemicals started at the beginning of the 1980s, to define so-called reference doses for chemicals: the levels of concentration at which they can safely be considered to not cause harm. This could be done because of the accumulation of data about volatile chemicals (then under threat of regulatory restrictions): data about people's inhalation of chemicals, data about biological metabolisation of these chemicals and data about the quantity of chemicals eliminated by the human body and exhaled. These data originated, notably, from the use of costly inhalation chambers. Once databases were elaborated, models started to be elaborated and calibrated in more reliable ways, for more chemicals, allowing to envisage the possibility to model together the chemical and the human body. In this field, the main challenge has always been the capacity to calibrate the model with realistic and varied biological data, to counterbalance the drive to make predictions based on more quickly produced, but less representative average values.

A third technique consists in developing what is called biologically based mechanistic models, to analyse the functioning of the human body and biological pathways inside those, as well as their interactions with substances. The resulting "biologically based dose response" (BBDR) models pursue the same kind of aim as PBPK—doing better than animal tests in terms of prediction of risk thresholds. Indeed, some of its champions are the same as for PBPK [14], and BBDR was also developed to counter or moderate regulatory drives on critical chemicals such as dioxin [1]. Instead of capturing biology through equations, as PBPK does, it banks on rapidly evolving knowledge of the cellular pathways through which chemical substances trigger potential toxicological issues. These theoretical models of biological organisms are supposed to guide the interpretation of empirical toxicological data. Much like PBPK, the reliability of this sort of modelling is limited by the data that are being used, and their capacity to represent "inter- and intraindividual heterogeneity" [15].

2.3.2 Datafication

All of the above techniques, as briefly mentioned, have been limited by the slow and costly generation of data through in vivo or in vitro tests, as well as by the quality of the hypotheses that guide their interpretation. In terms of toxicity data, the game-changer has come from the genomics (and the corresponding *toxico*genomics) revolution, namely from tools that can generate massive sets of data points about genetic events from a single experiment, and at high speed. "Omic" techniques, such as microarrays, make it possible to represent all the events in a biological system associated with the presence of a chemical substance. Robots allow multiple assays to be run on dozens or hundreds of substances day after day, generating massive sets of data, to be modelled by biologists. This toxicogenomic effort emerged a little after 2000s, after the three others introduced above.

Under the impetus of the chief of the US National Toxicology Program, Chris Portier (a biostatistician who had, among other things, worked in the area of PBPK and BBDR), a draft strategy was elaborated in 2003 "*to move toxicology to a predominantly predictive science focused upon a broad inclusion of target-specific, mechanism-based, biological observations*". The Environmental Protection Agency embarked on a similar effort a few years later. These institutions soon developed together a vast effort known as Tox21, to conduct hundreds of assays on thousands of substances thanks to high-throughput technologies. The central character in this program is a robot from the Swiss company Stäubli that autonomously manipulates plates containing dozens of mini-petri-dishes, to conduct multiple assays on multiple chemicals at several dosages. The result is an immense set of data, in which biological patterns can hope to be detected. This is done, notably, through open data challenges: the Tox21 institutions have called for teams of computational biologists around the world to search through their data to generate such models. This is where machine learning enters the picture: models are being constructed from the ground up, through supervised exploration of the mass of data to identify (or learn) patterns [16].

2.3.3 Computational Risk Assessment: The Integrated Vision

At about the same time as the Tox21 effort took off, a panel of top toxicologists and specialists of the field of toxicity testing, led by Melvin Andersen, rationalised computational toxicology.

The addition of high-throughput toxicogenomic to previous developments allowed to envision a future in which data would be available for many possibly toxicity pathways concerning multiple substances, to radically change how the toxicity of chemical substances would be tested: not as an isolated object with defined properties, but as elements of a biological system acting at low doses through diverse biological pathways. In other words, a knowledge system that would be representative of the reality of how biological systems function in the current chemicalised environment. The risk assessment of chemicals, thus, has evolved in the same manner as supporting disciplines such as biology, towards a more computational, systems-based style of analysis [17, 18].

The resulting "vision" was published by a branch of the US National Academies (the National Research Council) and heralded as the right guiding vision [19]. Interestingly, the vision seems to cap all previous efforts in the area of model-based, predictive toxicology: efforts in QSAR (to characterise properties of a substance), PBPK and BBDR (to formulate biomathematical models of the organism) and in high-throughput in vitro testing were now the building blocks of a knowledge system allowing to "*evaluate relevant perturbations in key toxicity pathways*" [19, p. 7], as opposed to simply measure the levels at which an object, taken in isolation, may prove harmful.

2.4 Discussion

The current development of artificial intelligence rests on a discourse about the all-powerful machine learning methods, and their unabridged capacity to learn from data, thanks to powerful computers. Risk modellers often resort to short-cutting claims such as the one that they can predict risks thanks to better maths and bioinformatics. In holding that discourse, modellers in the area of chemical risk assessment illustrate the fact that digitalisation colonises risk assessment of chemicals, just as it has colonised other areas of scientific practice.

Historians of science and technology have noted that the digital is a *lingua franca* in sciences; a form of generic technology that produces comparable epistemic effects across disciplines [20]. In the case that is documented here, one can see that computing technologies and theoretical, systems-based visions were both, in some ways, borrowed from neighbouring spaces. In the present case, one sees the application of deep learning late in the process, in the context of the Tox21 program. One also sees the importation of a robotic technology from industrial fields. To give one further example: PBPK modelling has developed and gained credibility thanks to the use of generic programming languages (e.g., Fortran), allowing more people to engage in this area, generate more models, creating an emulation/comparison of models, resulting in the improvement of the technique altogether.

This all too short historical overview has tried to specify, in contrast with this discourse, what are the area-specific conditions of a digital transformation. I have emphasised, first, that risk assessment has been, from the very start, a computational practice: a kind of science that asserted its scientificity through the development of gradually more complex modes of calculation of risks, moving towards the mathematisation and modelling of more and more aspects of the functioning of biological systems. There is certainly a degree of novelty in the current introduction of a variety of machine learning methods, but risk assessment has always used some means of computation, and the artificial intelligence methods that are being experimented now have a certain degree of continuity with previously used methods.

Second, it appears that the application of sophisticated means of computation and machine learning algorithms may not fulfil the promise of prediction, if it is not matched by equivalent investments in datafication. Computational modelling, indeed, does not mean doing without data, and without the various means available—including experimental ones—to generate, collect, curate and classify them. What one learns from the history above is that the generation of data is a necessary condition for moving towards more digital risk assessment. In fact, as can be gathered from the brief descriptions above, the various families of modelling techniques have been restricted by the same problem: the availability, diversity and representativeness of the data that are being modelled.

A simple conclusion to draw from this is that artificial intelligence will represent an innovation and a new leap in modelling capacities, in so far as it is matched by the parallel deployment of larger infrastructures of data allowing to document the various elements of these complex systems, rather than an isolated risk object and its effects. Failing the full datafication of the systems that scientists want to model, prediction will stay focused on these particular objects, as they have always been. As can be gathered from the brief description above, various risk objects are construed by computational systems over time. QSAR looks at the properties of molecules and models classes of chemicals. PBPK looks at the dose of chemicals in the human body and models physiological systems. In Tox21, it is the biological pathway that is the object of knowledge. These are heterogeneous objects, and the systems that are in place to know these objects are distinct, and not necessarily compatible. They may be, quite simply, the incarnation of different ways of modelling or predicting [21].

In the case of Tox21, even though a holistic vision has emerged, eventually, there is no assurance that these knowledge systems can be further integrated, or that the current development of artificial intelligence will bring coherence to past developments. It is so because there is ontological politics involved: a search, which may be contentious, for a realistic definition of what the problem is. A risk can be defined in reductive ways, assigned to an object that is deemed easier to regulate and control (i.e., the molecule). Or a risk can be defined in a more diffused, systemic manner and lead to the exploration of chains of causation between objects forming a complex system. The more one evolves towards modelling complex systems, the more complex it becomes to intervene in and regulate these systems, since modelling will reveal complex chains of causation and an intertwinement of causes. In the present historical case, this is illustrated by the fact the ontology of the "dose", "threshold" and of the risky object—the chemical substance to which a risk can be attributed—loses ground. This raises the issue of how decision criteria are forged in the space of knowledge systems that are designed to turn out complex correlations, rather than to isolate linear causation chains between an agent and an effect.

Overall, then, the ideal of digitalisation and the epistemic ambition to predict what is happening in systems may be capped by the establishment of data systems. A gap remains between the imaginary of digitalised prediction and the actual breadth of data systems. The various levels at which digitalisation unfolds—datafication, computational turn, theoretical visions of what may be modelled—reinforce one another, but they are not necessarily accorded in practice. For instance, with digitalisation and the big data revolution comes the "end-of-theory" claim: the notion that data-driven sciences will be fully empirical, learning from the bottom-up, by the mere, intensive exploration of data, to recognise patterns in complex systems, without the assistance of a priori theory about how these models are constituted and work.

Again, the history outlined above shows that this is unlikely to happen, as there is no pure and atheoretical exploration of data: data are generated by infrastructures that enact a certain theoretical vision of the world, namely in this case, a vision in terms of chemical substances and the measurement of doses of chemicals in bodies and environments. Generating data according to different theories is a process that will take ample time. Epistemologically, it would be legitimate to think that this is simply not possible: data and metrics necessarily enact a certain theoretical vision of what needs to be counted, of all that is happening out there in the world.

References

1. D. Demortain, *The Science of Bureaucracy: Risk Decision-Making and the US Environmental Protection Agency* (The MIT Press, 2020)
2. C. Hood, H. Rothstein, R. Baldwin, *The Government of Risk: Understanding Risk Regulation Regimes* (Oxford University Press, 2001)
3. S. Shapiro, R. Glicksman, *Risk Regulation at Risk: Restoring a Pragmatic Approach* (Stanford University Press, 2003)
4. S. Jasanoff, *The Fifth Branch: Science Advisers as Policymakers* (Harvard University Press, 1990)
5. A. Rip, The mutual dependence of risk research and political context. Sci. Technol. Stud. **4**(3), 3–15 (1986)
6. R. Kitchin, Big data, new epistemologies and paradigm shifts. Big Data Soc. **1**(1), 1–12 (2014). https://doi.org/10.1177/2053951714528481

7. B. Benbouzid, D. Cardon, Machines à prédire. Reseaux, no **211**(5), 9–33 (2018)
8. T. Aven, R. Flage, Foundational challenges for advancing the field and discipline of risk analysis. Risk Anal. **40**(S1), 2128–2136 (2020)
9. J. Downer, 737-Cabriolet: the limits of knowledge and the sociology of inevitable failure. Am. J. Sociol. **117**(3), 725–762 (2011). https://doi.org/10.1086/662383
10. S. Hilgartner, The social construction of risk objects, in *Organizations, Uncertainties and Risk*, ed. by J. Short, L. Clarke (Westview Press, 1992), pp 39–53
11. S. Ruphy, Simulations numériques de phénomènes complexes: Un nouveau style de raisonnement scientifique?, in F. Varenne (Éd.), Modéliser & simuler – Tome 2. Éditions Matériologiques.
12. H. Boullier, D. Demortain, M. Zeeman, Inventing prediction for regulation: the development of (quantitative) structure-activity relationships for the assessment of chemicals at the US environmental protection agency. Sci. Technol. Stud. **32**(4), 137–157 (2019)
13. B. Laurent, F. Thoreau, Situated expert judgement: QSAR models and transparency in the European regulation of chemicals. Sci. Tech. Stud. **32**(4), 158–174 (2019). https://doi.org/10.23987/sts.65249
14. M.E. Andersen, K. Krishnan, R.B. Conolly, R.O. McClellan, Biologically based modeling in toxicology research. Archives of Toxicology. Archiv Fur Toxikologie. Supplement **15**, 217–227 (1992)
15. K.S. Crump, C. Chen, W.A. Chiu, T.A. Louis, C.J. Portier, R.P. Subramaniam, P.D. White, What role for biologically based dose-response models in estimating low-dose risk? Environ. Health Perspect. **118**(5), 585–588 (2010)
16. R. Huang, M. Xia, Editorial: tox21 challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental toxicants and drugs. Front. Environ. Sci. **5**(3), 2–3 (2017). https://doi.org/10.3389/fenvs.2017.00003
17. H. Kitano, Systems biology: a brief overview. Science **295**(5560), 1662–1664 (2002)
18. E.T. Liu, Systems biology, integrative biology. Predictive Biology. Cell **121**(4), 505–506 (2005)
19. NRC, *Toxicity Testing in the 21st Century: A Vision and a Strategy*. (National Academies Press, 2007)
20. J. Lenhard, G. Küppers, T. Shinn, *Simulation: Pragmatic Constructions of Reality*, vol. 25. (Springer Science & Business Media, 2007)
21. S. Aykut, D. Demortain, B. Benbouzid, The politics of anticipatory expertise: plurality and contestation of futures knowledge in governance. Sci. Technol. Stud. **32**(4), 2–12 (2019)

Chapter 3
Key Dimensions of Algorithmic Management, Machine Learning and Big Data in Differing Large Sociotechnical Systems, with Implications for Systemwide Safety Management

Emery Roe and Scott Fortmann-Roe

Abstract The time is ripe for more case-by-case analyses of “big data”, “machine learning” and “algorithmic management”. A significant portion of current discussion on these topics occurs under the rubric of Automation (or, artificial intelligence) and in terms of broad political, social and economic factors said to be at work. We instead focus on identifying sociotechnical concerns arising out of software development in the topic areas. In so doing, we identify trade-offs and at least one longer-term system safety concern not typically included alongside notable political, social and economic considerations. This is the system safety concern of obsolescence. We end with a speculation on how skills in making these trade-offs might be noteworthy when system safety has been breached in emergencies.

Keywords Algorithmic management · Sociotechnical systems · Safety management · Automation

3.1 Roadmap and Introduction

After this section’s background preliminaries, we briefly examine the consequences of treating Automation/AI as an overarching rubric under which to frame discussions of algorithmic management, machine learning and big data. We then move to the bulk of the chapter to identifying and discussing the three key topics and associated

E. Roe (✉)
University of California, Berkeley, CA, USA
e-mail: emery.roe@berkeley.edu

S. Fortmann-Roe
Text Blaze, Berkeley, CA, USA
e-mail: scott@insightmaker.com
URL: https://insightmaker.com/

J.-C. Le Coze and S. Antonsen (eds.), *Safety in the Digital Age*,
SpringerBriefs in Safety Management,
https://doi.org/10.1007/978-3-031-32633-2_3

trade-offs within a sociotechnical context of hardware and software developers in highly distributed systems such as Google, Netflix, Facebook and Amazon's technical infrastructure. We conclude with discussing how our case- and topic-specific perspective helps reframe discussions of algorithmic management, machine learning and big data, with a special emphasis on system safety management implications.

Algorithmic management, machine learning and big data are fairly well-defined concepts. In contrast, the popularised term "AI" is in respects more a hype-driven, marketing term than a meaningful concept when discussing real-world digital issues focused on in this chapter. We will not discuss AI further, nor for brevity's sake are we going to discuss other key concepts the reader might expect to be included alongside algorithmic management, machine learning and big data.

In particular, this chapter does not discuss "expert systems" (expert-rule systems). This omission is important to note because expert systems can be thought of as the opposite in some respects of big data/machine learning. In the medical context, an expert system may be developed by having medical professionals define rules that can identify a tumor. A big data/machine learning approach to the same problem might start with the medical professionals marking numerous images of potential tumours as being either malignant or benign. The machine learning algorithm would then apply statistical techniques to create its own classification rules to identify which unseen tumours were malignant and benign. In case it needs saying, expert systems are also found in other automated fields, e.g., in different autonomous marine systems [10].

These differences matter because sociotechnical systems differ. In autonomous marine systems (among others), there are components (including processes and connections) that must never fail and events that must never happen (e.g., irreversible damage to the rig being repaired by the remotely operated vehicle) in order for the autonomous system to be reliable. Redundant components may not be readily available at rough seas. In highly distributed systems, by way of contrast, each component should be able to fail in order for the system to be reliable. Here redundancy or fallbacks are essential.

3.2 Limitations of "Automation" as a Covering Concept

To get to what needs to be said about algorithmic management, machine learning and big data, we must usher one elephant out of this chapter (albeit very much still found elsewhere in this volume), that of capital-A "Automation". This very large topic is the subject of broad political, social and economic debates (for much more detailed discussions of the interrelated debates over "automation" writ large, see: Benanav [2, 3]; McCarraher [5, 6]:

- *Economic*: It is said that Automation poses widespread joblessness for people now employed or seeking to be in the future.

- *Social*: It is said that Automation poses huge, new challenges to society, not least of which is answering the existential question, "What is a human being and the good life?"
- *Political*: It is said that Automation poses new challenges to the Right and Left political divide, e.g., some Right free-market visionaries are just as much in favor of capital-A Automation as some elements on the Left, e.g., "Fully Automated Luxury Communism" (for more on the possible political, social and economic benefits, see Prabhakar [7]

This chapter has nothing to add to or clarify for these controversies. We however do not see why these concerns must be an obstacle to thinking more clearly about the three topic areas.

The *sociotechnical context*, this chapter seeks to demonstrate, is just as important. Large-scale sociotechnical systems, not least of which are society's critical infrastructures (water, energy, telecommunications….), are not technical or physical systems only; they must be managed and operated reliably beyond inevitably baked-in limitations of design and technology [8, 9]. The sociotechnical context becomes especially important when the real-time operational focus centres on the three subject areas of algorithmic management, machine learning and big data in what are very different large sociotechnical systems that are, nevertheless, typically conflated together as "highly automated".

If we are correct—the wider economic, political and social contexts cannot on their own resolve key concerns of the sociotechnical context—then the time is ripe for addressing the subjects of concern from perspectives typically not seen in the political, social and economic discussions. The section that follows is offered in that spirit.

3.3 Developers' Perspective on a New Software Application

We know software application developers make trade-offs across different evaluative dimensions. The virtue of the dimensions is that each category can be usefully defined from the developer's perspective and that each fits into a recognisable trade-off faced by software developers in evaluating different options (henceforth, "developers" being a single engineer, team or company).

This section focuses on a set of *interrelated* system trade-offs commonly understood by software developers, including their definitions and some examples. Many factors will be familiar to readers, albeit perhaps not as organised below. No claim is made that the set is an exhaustive list. These well-understood dimensions are abstracted for illustrative purposes in Fig. 3.1.

1. **Comprehensibility/Features Dimension**

    - **Comprehensibility (Left Side)**: Ability of developer to understand the system, all bounded by human cognitive limits. Highly distributed systems

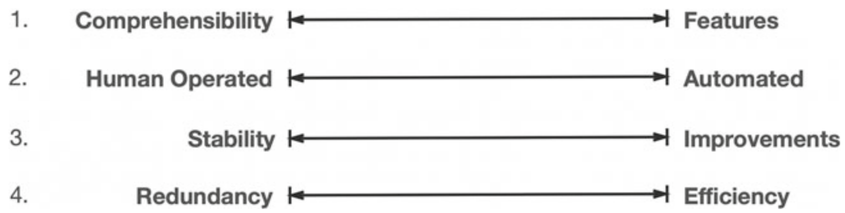


Fig. 3.1 Four key interrelated trade-offs for software developers

are often beyond the ability of one team, let alone individual, to fully know and understand as a system.

- **Features (Right Side)**: Capabilities of the system. Additional features provide value to users but increase the system's sociotechnical complexity,[1] thereby reducing comprehensibility.

2. **Human Operated/Automated Dimension**

- **Human Operated (Left Side)**: Changes to the system configuration are carried out by human operators. For example, in capacity planning, servers may be manually ordered and provisioned to address forecasted demand.
- **Automated (Right Side)**: The system may dynamically change many aspects of its operation without human intervention. For instance, it may automatically provide or decommission servers without the intervention of human operators.

3. **Stability/Improvements Dimension**

- **Stability (Left Side)**: System operates at full functionality without failure. Beyond strict technical availability, stability may also include the accessibility of the systems to operators trained on an earlier version of the system without requiring retraining.
- **Improvements (Right Side):** Changes to the system are made to provide new and enhanced features, or other enhancements such as decreased latency (response time).

4. **Redundancy/Efficiency Dimension**

- **Redundancy (Left Side):** The possibility of the system to experience the failure of one or more system components (including processes and connections) and still have the capacity to support its load. An example is a system provisioned with a secondary database ready to take over in case the primary one fails.

[1] Complexity in terms of these digital systems is indexed in terms of the elements, their functions and the interrelationships between elements and functions in the systems (for this classic definition of sociotechnical complexity, see LaPorte [4]. These features are also captured by (Michael) Conway's law, "organizations which design systems … are constrained to produce designs which are copies of the communication structures of these organizations".

- **Efficiency (Right Side):** The ability of the system to provide service with a minimum of cost or resource usage. It is paying for what you are using only.

These four dimensions are relied upon by software builders, where the trade-offs can be explicitly codified as part of the software development and application process. Consider Google's Site Reliability Engineer (SRE) "error budget", where applications are given a budget of allowed downtime or errors within a quarter time period. If exceeded—the application is down for longer than budgeted—additional feature work on the product is halted until the application is brought back within budget.[2] This is an explicit example on the **Stability/Improvements** dimension.

For each of the four dimensions, current technology and organisation processes occupy one or more segments along the dimension. These respective segments expand/intensify as new technology and processes are developed.

By way of illustration, consider the **Human Operated/Automated** dimension. Technology and new services have provided additional opportunities to automate the management of increasingly complex sociotechnical systems:

- In the 2000s, the advent of Cloud providers such as Amazon Web Services (AWS) and Google Cloud Platform (GCP, initially with App Engine) provided significant opportunities to provision hardware via application programming interfaces (API's) or technical interfaces making it relatively simple to spin up/down hardware instances based on automated heuristics.
- More recently, big data and machine learning have provided additional opportunities to manage systems using opaque ML algorithms. DeepMind has, for example, deployed a model that uses machine learning to manage the cooling of Google's data centres leading to a 40% reduction in energy use.[3]
- Processes, such as Netflix's Chaos Monkey, enable the organisation to validate the behavior of their highly complex systems under different failure modes. By way of example, network connectivity may be deliberately broken between two nodes to confirm the system adapts around the failure, enabling them to operate increasingly complex and heavily automated architectures.

The expansion of a dimension's segments is dominated by an asymmetrical expansion of activities and investments on the right side. The importance of Cloud providers, big data and machine learning in driving the expansion has been mentioned. Other factors include sociotechnical shifts such as agile methodologies, the rise of open source, the development of new statistical and machine learning approaches, and the creation of more recent hardware such as GPU's and smartphones.

[2] "Error budgets are the tool SRE uses to balance service reliability with the pace of innovation. Changes are a major source of instability, representing roughly 70% of our outages, and development work for features competes with development work for stability. The error budget forms a control mechanism for diverting attention to stability as needed." (https://sre.google/workbook/error-budget-policy/).

[3] https://deepmind.com/blog/article/deepmind-ai-reduces-google-data-centre-cooling-bill-40

3.4 What's the Upshot for System Safety? Obsolescence as a Long-Term Sociotechnical Concern

System safety is typically taken to be on the left side of the developer's trade-offs, located in and constituted by stability, redundancy, comprehensibility and recourse to human (manual) operations. Since the left side is also expanding (due in part to advances not reported here outside the three topic areas), we can assume the left-side expansion contributes to advances in safety as well.

If the left side is associated with "system safety", then the right side can be taken as the maximum potential to generate "value" to the developer/company. Clearly, increases in both right-side features and right-side efficiencies can increase the ability of the system to provide value for its operators or users, other things considered.

Now, look at that left side more closely, this time from the perspective of the designer's long term versus short term.

Software applications are littered with examples of stable and capable systems that were rendered obsolete by systems that better met users' needs in newer, effective ways. If the current electrical grid rarely goes down, that is one form of safety. But do we want a system that is stable until it catastrophically fails or becomes no longer fit for new purposes? Or would we prefer systems to fail by frequent small defects that, while we can fix and solve in real time, nonetheless produce a steady stream of negative headlines?

More generally and from a software designer's perspective, we must acknowledge that even the most reliable system becomes, at least in part and after a point, outdated for its users by virtue of not taking advantage of subsequent improvements, some of which may well have been tested and secured initially on the right side of the trade-offs.

In this way, obsolescence is very much a longer-term system safety issue and should be given as much attention, we believe, as the social, political and economic concerns mentioned at the outset. Cyber-security, for example, is clearly a very pressing right-side issue at the time of writing but would still be pressing over the longer term because even stable defences become obsolete (and for reasons different than current short-term ones).[4]

[4] Cyber-security, however, deserves its own treatment and pushes us beyond the remit of this chapter (see how *societal safety* and *societal security* overlap and differ in Almklov et al. [1].

3.5 A Concluding Speculation on When System Safety is Breached

Since critical infrastructures are increasingly digitised around the three areas, it is a fair question to ask: Can or do these software developer skills in making the four trade-offs assist in immediate response and longer-term recovery after the digital-dependent infrastructure has failed in disaster? This is unanswerable in the absence of specific cases and events and, even then, answering would require close observation over a long period. Even so, the question has major implications for theories of large-system safety.

The crux is the notion of trade-offs. According to high reliability theory, system safety during normal operations is non-fungible after a point, that is, it cannot be traded-off against other attributes like cost. Nuclear reactors must not blow up, urban water supplies must not be contaminated with cryptosporidium (or worse), electric grids must not island, jumbo jets must not drop from the sky, irreplaceable dams must not breach or overtop, and autonomous underwater vessels must not damage the very oil rigs they are repairing. That disasters can or do happen reinforces the dread and commitment of the public and system operators to this precluded-event standard.

What happens, though, when even these systems, let alone other digitised ones, fail outright as in, say, a massive earthquake or geomagnetic storm and blackout? Such emergencies are the furthest critical infrastructures get from high reliability management during their normal operations. In disasters, safety still matters but trade-offs surface all over the place, and skills in thinking on the fly, riding uncertainty and improvising are at their premium.

If so, we must speculate further. Do skills developed through making the specific software trade-offs add value to immediate response and recovery efforts of highly digitised infrastructures? Or from the direction: Is the capacity to achieve reliable normal operations in digital platforms—not by precluding or avoiding certain events but by adapting to electronic component failure most anywhere and most all of the time—a key skill set of software professionals and their wraparound support during emergency management for critical infrastructures? Answers are a pressing matter, as when an experienced emergency manager in the US Pacific Northwest itemised for one of us (Roe) just how many different software critical to the emergency management infrastructure depend on one platform provider major in the region (and globally for that matter).

References

1. P.G. Almklov, S. Antonsen, K.V. Størkersen, E. Roe, Safer societies. Safety Scie. **110**(Part C), 1–6 (2018)
2. A. Benanav, Automation and the future of work--Part one. New Left Rev. **119**(Sept/Oct), 5–38 (2019a)

3. A. Benanav, Automation and the future of work--Part two. New Left Rev. **120**(Nov/Dec), 117–146 (2019b)
4. T.R. La Porte, *Organized Social Complexity: Challenge to Politics and Policy* (Princeton University Press, 1975)
5. E. McCarraher, Automated vistas (I). Raritan **39**(1), 18–42 (2019)
6. E. McCarraher, Automated vistas (II). Raritan **39**(2), 102–126 (2019)
7. A. Prabhakar, In the realm of the Barely feasible—complex challenges acting the nation require a new approach to ramping up innovation: solutions R&D. Issues in Science and Technology XXXVII **1**(Fall), 34–40 (2020)
8. E. Roe, P.R. Schulman, *High Reliability Management: Operating on the Edge* (Stanford University Press, Stanford, CA, 2008)
9. E. Roe, P.R. Schulman, *Reliability and Risk: The Challenge of Managing Interconnected Infrastructures* (Stanford University Press, Stanford, CA, 2016)
10. I.B. Utne, I. Schjølberg, E. Roe, High reliability management and control operator risks in autonomous marine systems and operations. Ocean Eng. **171**(1), 399–416 (2019)

Chapter 4
Digitalisation, Safety and Privacy

Olivier Guillaume

Abstract In order to increase the safety of industrial facilities and people, firms and their managers traditionally pay attention to the visibility of activities and the intentions of workers. Firms and managers can use connected objects worn by workers to collect this data. Analysing the introduction of smart glasses and smart shoes in an industrial site, this contribution explains how workers can use these tools without sacrificing their autonomy and privacy. In this site, performance as well as the safety of the activities is based on a combination of high individual autonomy and solidarity between colleagues strengthened by a private life at work. The sociotechnical context of this industrial site and the desire of professionals to control their privacy at work have a strong impact on the trajectory of these technologies. They insist on the use of technologies with their colleagues which strengthen the bonds of cooperation and solidarity and are resistant to technologies that could geolocate them or trace their movements. Moreover, the association of user spokespersons is an essential condition for the success of the design and dissemination of digital technologies. Finally, the control of privacy and private life at work by the workers contributes to the reinforcement of the performance and the safety of production.

Keywords Privacy in the workplace · Digitalisation · Safety management

4.1 Introduction

The safety of installations and people is traditionally the result of increased attention paid by firms and their managers to the visibility of activities, their actual sequences and the intentions of workers. The increased transparency ideally allows a reduction of uncertainties and potential errors in the execution of activities and improved anticipation, programming and control of the realisation of processes. Transparency

O. Guillaume (✉)
EDF R&D & Laboratoire Printemps, Université Versailles Saint-Quentin en Yvelines, Paris, France
e-mail: olivier.guillaume@edf.fr

J.-C. Le Coze and S. Antonsen (eds.), *Safety in the Digital Age*,
SpringerBriefs in Safety Management,
https://doi.org/10.1007/978-3-031-32633-2_4

also allows faster and cheaper remedial actions while optimising business execution schedules by reducing downtime, incompatibilities and sources of error.

The challenge for firms, or rather the utopia of some of them, to ensure the safety of their facilities and their staff while optimising their productivity has always been to detect, capture and analyse more and more data on actual activities, movements and locations of workers. Even more, to capture and analyse more and more data on the sequencing of activities to ensure the safety of workers or to anticipate any risk of industrial incidents, as well as to rationalise the movement of individuals and goods, maintenance and execution programmes to increase productivity.

With this in mind, individualised digital tools embedded on workers—smart phones and PDAs, smart glasses and soon wearable sensors—have flourished over the past decade because they simultaneously allow companies to capture live data on the actual activities carried out, sequences, execution times or errors in order to optimise the programming of activities and reduce faults while optimising shifts or the timing of tasks (Kogan, [12] on road transport). The improvement of organisational performance by on-board digital tools is also based on the sending in real time of technical and precise information to professionals, easier access to precise documentation or dialogue with remote experts advising workers. The geolocation function on these tools can also help to locate and rescue them in the event of a perilous situation. At the same time, employees can claim to be equipped with these digital tools in order to effectively carry out their professional missions due to several trips between establishments or spaces within the same factory. Embedded and individualised digital tools allow them to receive remotely and continuously information in order to carry out their activities reliably or to send it back to the organisation and their managers in order to inform them about the carrying out of activities or their location.

Managers and employees are then faced with the contradiction of having to use technologies that can provide information and expert assistance to carry out efficient and reliable activities, while strengthening hierarchical control over time, travel or compliance with procedures and encouraging rationalisation of work. However, behind this classic theme of the confrontation between professional autonomy and organisational rationalisation, wearable digital tools explore new areas of control by entering the private lives of workers. Indeed, sensors integrated into watches, vests or smart glasses can, for example, assess the locations, movements, timing of movements or the state of stress of the worker by measuring their heart rate or pupil dilation. The wearable digital objects then enter the bodily intimacy of workers and their private lives, and limit and control the space-times necessary for rest, reflection, creativity and therefore performance and safety, as we will show below.

Considered as fundamental and an individual freedom, privacy is nevertheless protected in the professional sphere[1] by several laws and regulations which arbitrate a priori the tensions between the organisations and the professionals using these tools.

[1] The protection of the personal life of the employee at work has been affirmed on numerous occasions by the European Court of Human Rights or the Social Chamber of the French Court of Cassation, specifying that “the employee has the right to, even at the time and place of work, respect for the privacy of his private life” [10]. More specifically, personal data at work has long been

However, this programmatic vision is undermined by reality: people in organisations are not always familiar with these laws, have difficulties interpreting their meaning in specific situations or do not always have an interest in mobilising them.

These points raise a number of questions which must be empirically analysed. Is it possible that these technologies are increasing workers' efficiency without sacrificing their autonomy to a rationalised and disempowering organisation? How do workers using these technologies protect and manage their privacy at work and their personal data?

4.2 Individualised Digital Tools and Privacy at Work

Individualised and wearable technologies renew the classic issue of professional autonomy constrained by organisational rationalisation, which is based on visibility of activities, movements, locations or the real intentions of employees. These technologies potentially immerse themselves in the intimacy and privacy of employees and then unveil them before ultimately turning against the objectives of enhanced safety and performance.

To understand this, it is necessary to clarify some terms. We use the term intimacy in this chapter to refer to the inner space of the individual. Secret, it withdraws from the gaze and control of others. Private life is an area where exchanges between intimacy and the public take place and in which confidential information passes to a small circle of individuals who respect secrecy and discretion [8]. Privacy is the potentiality of controlling intimacy and private life. It is understood as an immutable right of each individual "to be left alone" [24]. It can be conceived as a selective control of access to oneself [2] in order to minimise their vulnerability during interaction with others [14]. It represents the degree to which a person has access to others and their information. It is protected when others have limited access to a person's thoughts, bodies and possessions [19].

In the workplace, privacy-related analyses mainly focus on the use of digital tools and the collection of their data. They particularly question the potential for monitoring and restricting the autonomy and the private life of employees [16] by devices that make social phenomena more visible, individuals more calculable, exploitable and governable [7]. For example, the geolocation functionality of mobile phones can discipline and regulate the practices of technicians by forcing them to report each

protected in France by the Data Protection Act, which imposed five main principles to be respected. These principles are now extended in the General Data Protection Regulation (GDPR) implemented in 2018. This text protects European citizens from violations of their privacy. Their personal data are considered as "any information relating to an identifiable natural person, directly or indirectly". This text specifies that the processing of their data (the collection, recording, conservation, extraction, consultation, dissemination) must be lawful, fair and transparent with regard to the subject, for specific purposes, explicit and legitimate. This text gives citizens new rights: collection of specific free consent, informed and unambiguous if no legal basis for data processing, right to information, right of access, rectification and erasure, right to portability.

intervention or by measuring the time spent on each site or during journeys, before data are transmitted to managers in order to compare individual performances [13]. Construction workers are reluctant to wear smart vests that locate them in the event of an incident, for fear that they will generate data to monitor their downtime [6]. Physiological data from sensors can potentially be used to infer driving behaviours and performance [10].[2]

Other authors stress the potential for articulation between professional and private life offered by digital tools. On one hand, they allow employees to regulate their private affairs or carry out their relational and social activities during their working day [21]. On the other hand, they manage crises and tensions at work or break loneliness at work by enabling calls to friends and spouses [4].

However, it would be unfortunate to limit analyses of privacy at work to the impact of technological objects, digital traces collected and potential for control. Palm [17] thus defines privacy at work as local and informational. Local privacy includes the possibilities for the employee to retire to isolated areas of work and to use some of them in order to protect themselves from intrusions and observations. Indeed, the public gaze could stifle the expression of intimate feelings and the possibility of "governing" oneself by acting without the approval of others, developing one's own standards or rearranging one's thoughts in order to prepare our "public performances". Informational privacy is the individual's ability not to publicly reveal personal data and to retain some of them in order not to interfere with relationships with colleagues, employers, consumers or clients. Information privacy includes limits on data explaining when, where and how employees carry out their activities, which give the employer a detailed picture of their productivity. Ultimately, as Palm [17, 18] specifies, controlling their privacy allows workers to restrict others' access to themselves, build their professional role or prepare for their public performance. Thanks to this, an individual can understand themselves and develop autonomous acts, form their own standards and act in accordance with them while entering into a relationship with others considered as an equal. They can also form and develop their own goal, express their identity and their value, ensure peace and personal reflection, in particular to develop safe or more effective acts.

4.2.1 An Empirical Analysis

To illustrate these ideas and answer the proposed questions, we will refer to an empirical study carried out on an industrial production site. Its technical design, which is unique in France, includes specific equipment on which maintenance technicians, sometimes with little experience on the site, must intervene, which sometimes leads

[2] Even if other contributions show that digital tools can pool the knowledge and practices of actors who rarely spoke and met each other, facilitate the emergence of new work collectives [3] and do not necessarily strengthen management control and sanctions due to poor analysis of digital traces by controllers and managers, who must understand and handle the data collected, have the time and the inclination, which is not always the case [11, 12].

them to experience difficulties either in locating the equipment on which to intervene or in carrying out certain operations. Technicians can sometimes work alone, accompanied by work documentation that is not always up to date or even available. The industrial performance as well as the safety of the activities is based on a combination of high individual autonomy in the activities to organise them, to seek information, to train by companionship… and solidarity between colleagues to transfer quickly the precise information on materials and activities. In order to reinforce this solidarity, the technicians develop a private life at work which includes rituals of coffee breaks and outdoor outings (aperitifs, sports activities, etc.) allowing the expression of feelings of comfort, support and mutual aid, which facilitate cooperation during work activities. The technicians know each other intimately and trust each other. This facilitates mutual assistance and the rapid transmission of technical information to carry out activities with performance and safety.

However, the organisation is proposing reforms that shake up local and information privacy. Managers have created a single open plan workspace for all technicians near their own offices, eliminating the possibility for technicians to fully control the relevant information on activities, i.e., the tricks of the trade to carry out the tasks. Indeed, managers can easily go to the open space and attend technicians' preparatory activities and meetings. Space reform is also a power issue for managers. Technicians have reacted to protect their informational privacy on personal work data by taking refuge in "interstitial spaces" [9]—like corridors and entrance halls—where they can exchange this knowledge or take refuge in secluded "intimate spaces"—such as offices and meeting rooms—where they can think about their activities and the way to do them. During the same period, the organisation and the managers are testing digital technologies which will allow workers to enter into dialogue with colleagues to help them carry out activities, or to locate them in order to help them in the event of an incident.

The sociotechnical context of this industrial site and the desire of professionals to control their privacy at work have a strong impact on the trajectory of these technologies. Maintenance technicians are completely in favor of being equipped with digital technologies that allow them to enter into dialogue and cooperation with colleagues from their site but are not favorable to increased collaboration with external experts. The latter do not know the technical specificities of the site, and the absence of informal relationships and the associated interpersonal trust with these unknown experts would not allow the technicians to expose their doubts on their activities or their professional shortcomings. They insist on the use of technologies with their colleagues to strengthen the bonds of cooperation and solidarity necessary for the performance and security of the organisation. Moreover, they clearly express that these technologies should not be continuously active, but just a temporary help.

Finally, the maintenance technicians of this factory are resistant to other technologies that could locate them or even trace their movements. They wish to protect their autonomy of movement on site, not for the pleasure of strolling around the site, but to reinforce the performance and the safety of the installations by going to check an equipment, to make a tour of the facility, help a colleague… In this situation, professional autonomy and movement allow the control of the installation and the acquisition of knowledge to strengthen security.

Scholars emphasise that the major pitfalls for the acceptance and use of wearable digital technologies are the lack of association and consideration of user expectations from the design stage, as well as the potential for reinforcement controls and rationalisation of work. These pitfalls are overcome in organisations where strong ethical rules combined with cooperative relations between actors make it possible to involve “spokespersons” from the design phase and to “translate” users’ expectations in order to integrate them into technical systems [1]. In this sense, the integration of local teams from the design of the project and throughout the use of technologies is essential. Their integration makes it possible to facilitate the collection and implementation of their expectations and their uses during the design and use phases of technologies. Moreover, an appropriate discourse addressed to users in the development phase of wearable digital technologies is a key component in the acceptability of these tools. This consists of emphasising the legitimate purpose of technology and its role as a decision aid and not as a means of controlling activity.

These lessons can be found in the empirical case mentioned. Helped by ergonomists and sociologists, the representatives of the technical professions discuss the issues they face in situation, before imagining specific uses and purposes for new applications implemented on a new smartphone-like device. The technology is all the more accepted as it does not duplicate other uses in the installation. It is not perceived at the time of our study as a potential instrument for rationalising and controlling work that would limit privacy and intimacy at work. The technicians only imagine using it occasionally to receive information in order to become more competent and autonomous in their activities, without the devices generating data concerning their work performance. It is different with the second technology, smart shoes, which are less well perceived because they duplicate the “deadman” safety technology that they already use, and for which technicians do not imagine specific uses that would help resolve their professional challenges. Worn all the time, they also raise fears of control technology that would undermine their privacy at work. Overcoming user reluctance also results from the ability of organisations and their teams to respond to users’ questions about unwanted uses of technologies. This capacity is established thanks to the existence of places favouring contradictory debates between the different categories of actors to ultimately organise “joint regulations” [22].

4.3 Conclusion

Wearable digital tools can provide information and expert assistance to carry out efficient and reliable activities, while strengthening managers’ control over time, travel or compliance with procedures and encouraging a rationalisation of work. Behind this classic theme of the confrontation between professional autonomy and organisational rationalisation, these digital tools explore new territories of control by entering the private lives of workers. Privacy at work is not reducible to the digital data harvested by technologies, but concerns broader control of oneself, of one's information or of one's private activities, which can be weakened by the reform of the spatial organisation of work, managerial presence or the recording of precise data by digital tools, limiting the possibilities for workers to relax, to decompress, to experiment with new activities, to organise the sequence of activities themselves or to develop new knowledge, contributing to the performance and safety of installations.

Issues of privacy at work cannot be reduced to digital data from technology, but include workers' ability to control use of their personal information or data on their work activities. In the context of the empirical case we described, the spatial organisation of work and managerial presence participate in the control of privacy, which forces the worker to fit together different symbolic spaces. Those devoted to private life allow the development of socialisation rituals. Interstitial spaces allow the exchange of specific professional knowledge. The intimate spaces allow the professional to reflect on the evolution of their activities. The control of privacy and private life at work by the worker contributes to the reinforcement of the performance and the safety of production giving the worker the possibility to experiment new activities, to organise them or develop new knowledge.

In addition, we wanted to show that if the association of user spokespersons is an essential condition for the success of the design and dissemination of digital technologies, the technical and organisational context also has an essential impact on the trajectory of these technologies. The constraints of the activities, the psychological tension of carrying out complex activities alone explain the importance of a private life at work which reinforces solidarity between technicians of the same team. In this context, technologies reinforcing the exchange between peers are well accepted, contrary to those which individualise and reinforce interactions with external experts.

The production of digital personal data on work, resulting from the use of technologies, nevertheless increases the difficulties of protecting the privacy of workers since they can be memorised, potentially aggregated or searchable. Even if these risks are limited by European and national regulations (GDPR, labor code), experience shows that the rules are not always an absolute guarantee because they may be poorly known, understood or interpreted. Workers and organisations therefore have an interest in adopting new approaches to guarantee the protection of workers' digital data. First, this kind of risk can be limited in companies with a strong social tradition of negotiation and where negotiation relations with worker representatives are institutionalised, strengthening managers to comply with the ethical obligations and regulatory protection of "privacy at work". Secondly, this kind of risk can be limited when employees are able to develop relationships of trust with the managers to be sure that some data will not be used contrary to their interests. However, not everyone is able to develop these kinds of relationships, and sometimes the organisational conditions are not favorable.

Bibliography

1. M. Akrich, M. Callon, B. Latour, A quoi tient le succès des innovations. Gérer et Comprendre. Annales des mines, 11–12 (1988)
2. I. Altman, *The Environment and Social Behavior: privacy, Personal Space, Territory, and Crowding* (Brooks/Cole Publishing, Monterey, CA, 1975)
3. A. Boboc, Numérique et travail: quelles influences? Sociologies Pratiques **1**(34), 3–12 (2017)
4. S. Broadbent, *L'intimité au travail*, Fyp Editions (2011)
5. A. Casilli, Contre l'hypothèse de la « fin de la vie privée ». La négociation de la privacy dans les médias sociaux. Revue française des sciences de l'information et de la communication, 3/ 2013. http://rfsic.revues.org/630
6. B. Choi, S. Hwang, S. Lee, What drives construction workers' acceptance of wearable technologies in the workplace?: Indoor localization and wearable health devices for occupational safety and health. Autom Constr **84**, 31–41 (2017)
7. B. Doolin, Power and resistance in the implementation of a medical management information system. Information Systems J. **4**, 343–362 (2004)
8. J.P. Durif-Varembont, L'intimité entre secrets et dévoilement. Cahiers de psychologie clinique **1**(32), 57 à 73 (2009)
9. P. Fustier, L'interstitiel et la fabrique de l'équipe, Nouvelle revue de psychosociologie 2(14), 85 à 96 (2012)
10. R. Greenfield, E. Busink, C.P. Wong, E. Riboli-Sasco, G. Greenfield, A. Majeed, P.A. Wark, Truck drivers' perceptions on wearable devices and health promotion: a qualitative study. BMC Public Health **16**(1), 677 (2016)
11. G. Jemine, L'outil face au manager: le contrôle du travail à l'ère du numérique, un terrain controversé? Les Cahiers du numérique **15**(4), 137–162 (2019)
12. A.F. Kogan, TIC, tac, tic, tac…Du temps traqué au travail contrôlé: le cas du transport routier de marchandises, Tic & société **10**(1) (2016)
13. A. Leclercq Vandelannoitte, H. Isaac, M. Kalika, Mobile information systems and organizational control: beyond the panopticon metaphor? Eur J Inf Syst **23**(5), 543–557 (2014)
14. S.T. Margulis, Conceptions of privacy—current status and next steps. J. Soc. Issues **33**(3), 5–21 (1977)
15. H. Nissenbaum, *Privacy in Context. Technology, Policy, and the Integrity of Social Life* (Stanford University Press, 2009)
16. W.J. Orlikowski, Integrated information environment or matrix of control? The contradictory implications of information technology. Account., Manag. Inform. Tech. **1**, 9–42 (1991)
17. E. Palm, Privacy expectations at work—what is reasonable and why? Ethical Theory and Moral Pract **12**(2), 201–215 (2009)
18. E. Palm, Securing privacy at work : the importance of contextualized consent. Ethics Inform. Tech. **11**, 233–241 (2009). https://doi.org/10.1007/s10676-009-9208-8
19. J. Persson Anders, S.O. Hansson, Privacy at work—ethical criteria. J. Bus. Ethics **42**, 59–70 (2003)
20. B. Rey, *La vie privée à l'ère du numérique*. Paris, Lavoisier, Coll. Traitement de l'information (2012)
21. B. Rey La vie privée au travail. Les Cahiers du numérique **9**(2),105–136 (2013)
22. J.D. Reynaud, Les règles du jeu, Paris, Armand Colin (1988)
23. D.J. Solove, *The Digital Person: Technology and Privacy in the Information Age* (New York University Press, New York, 2004)
24. S. D. Warren, D.L. Brandeis, *The Right to Privacy*, vol 5 (Harvard Law Review, 1890).

Chapter 5
Design and Dissemination of Blockchain Technologies: The Challenge of Privacy

Cécile Caron

Abstract Presented as trust technologies, blockchains, by allowing immediate secure peer-to-peer exchanges without a trusted third party, have strong disruptive potential, but raise privacy issues. We illustrate some challenges that this antagonism raises and the sociotechnical compromises made to overcome them, by analysing the design of a mobility service by a consortium of some fifteen operators, and its experimentation with the employees of these operators. The service seeks to respond to the new needs linked to the electrification of company fleets, by tracking the recharging of (personal) electric vehicles at work or (professional) vehicles at home with a view to reimbursing employees' professional expenses by relying on a blockchain. Privacy management is a skill, based on emerging expertise, distributed across a range of professions and users, which requires compromises between different conceptions of technology and data to be guaranteed. For blockchain designers, these compromises have limited the disruptive potential of blockchain technology by recentralising data management and losing the open nature of blockchain. However, in the eyes of other designers and users, they have allowed unexpected uses and benefits to emerge, such as reinforcing the choice of blockchain technology as a "*privacy solution*".

Keywords Blockchain · Privacy

Blockchain is a technology for storing and sharing information, based on the recording of data in the form of blocks linked to each other in the chronological order of their validation, making it possible to certify with certainty the date of the transaction. These blocks are processed in a decentralised manner and are protected by cryptographic methods. Each piece of data deposited in the blockchain is verified by intermediaries (the "miners") according to a precise protocol. The infrastructure is thus distributed within a network ("distributed ledger technology"), which makes it possible to do without a trusted third party when a transaction is carried out. In the

C. Caron (✉)
EDF Lab, Paris, France
e-mail: cecile.caron@edf.fr

J.-C. Le Coze and S. Antonsen (eds.), *Safety in the Digital Age*,
SpringerBriefs in Safety Management,
https://doi.org/10.1007/978-3-031-32633-2_5

context of the energy transition, FinTech blockchain technologies are seen as likely to be disruptive innovations for the energy sector. By allowing secure, immediate and almost free of charge peer-to-peer exchanges without the intermediary of a trusted third party, blockchains have strong disruptive potential [29] for tracking and transferring assets or for executing *smart contracts* (autonomous programs that automatically execute the terms and conditions of a contract, without human intervention).

However, they are controversial, especially in terms of privacy [17]. Indeed, if blockchains offer sufficient guarantees that no external attack can access personal information [20] and allow "*respect for privacy through the proactive use of cryptography"* [27], they raise questions of compliance with the European General Data Protection Regulation (GDPR) around the adequate processing of personal data [11], but also in terms of responsibility and explicability. The GDPR requirement to specify a data controller when processing personal data is incompatible with the decentralised operation of the blockchain. The right granted to users to delete and modify their personal data conflicts with the immutability of the registry, while the requirement of explicability is difficult to apply to the implementation of complex cryptographic algorithms.

Blockchains and the GDPR constitute two relatively antagonistic proposals for trust models. The principles of blockchain were conceived in the context of a crisis of confidence in institutions, particularly banks [2]. Gathered in the Bitcoin white paper [25], they combine a series of technical and social properties (trust and distributed consensus, infallibility and auditability of the register) which are similar to a proposal for trust in an "expert system" characteristic of modernity [16]: trust is no longer placed in a person, but in a system. For Giddens, our modern, anonymous, highly complex and functionally differentiated society has led to a radical transformation of the status of trust: social order is no longer based solely on familiarity and personal trust but also on trust in abstract systems. In contrast (see Table 5.1), the GDPR reflects a conception of trust as the "empire of the third party" [21], with primary social relations coming under the aegis of the instituted third party (as authority in a third-party position and as internalisation of the condition of a legal subject).

These issues of regulatory compliance are accentuated by the emergence of a growing concern among users about the protection of their privacy [24]. "In the vein of "surveillance studies" [5], a body of work argues that the increase in technological capabilities [has] broken down the boundaries that protect us from an Orwellian world" [8]. The technologisation of surveillance is said to be a constant threat to individual freedoms and privacy. Other approaches link the end of privacy to the very extension of the norms of authenticity and the public sphere that are derived from it, which reduces the possibilities of preserving one's freedom behind social roles that deliver the self to the "tyrannies of intimacy" [31]. While departing from this hypothesis of the "end of privacy", recent work associated with the emergence of surveillance capitalism [30] or with the analysis of social and digital practices attests to profound changes both in societies' perception of privacy and in the articulations between its different components [4, 9, 23]. The work on privacy reflects the plurality of dimensions that it incorporates. It is a right, notably to tranquillity [28], a commodity that although contested can be commensurable and exchangeable [6];

Table 5.1 Summary of the principles underlying trust in blockchains and in systems concerned by the GDPR

Principles of blockchains			Principles of GDPR	
Principles	Expression		Principles	Expression
Decentralization	Public (without third parties), private (with a central entity) or semi-private (consortium) blockchains	versus	Accountability of processing	Appointment of a Data Protection Officer and maintenance of a register of processing operations and purpose limitation
Transparency	Unforgeable history of all transactions and anonymity	versus	Protection on of personal data	Collection of consent if no legal basis, minimisation of data and their retention, exercise of rights (to information, to erasure, to correction)
Security	Cryptographic algorithms and secure transmission protocols	versus	Explicability	Transparency of algorithms
Trust in "expert systems" [16]		versus	Trust in "the empire of the third-party" [21]	

a state that allows personal spaces to be preserved from the intrusion of others by reserving access to limited groups of people [32], a capacity to manage social capital in a negotiated form [10] and a capacity for control [4]. All of these dimensions are affected by major technological, regulatory and societal developments that interact and generate vulnerability for individuals, but also for organisations, in terms of privacy management [24, 4, 33).

In a context where privacy issues, from a regulatory and societal point of view, impact the design and use of emerging technologies such as blockchain, this chapter analyses the way in which the actors involved in the design and experimentation of a service deal with these tensions between blockchain and privacy.

To do this, we will study a use case, the design and experimentation of a mobility service based on a blockchain. The service seeks to respond to new needs linked to the electrification of corporate fleets, by tracking the recharging of (personal) electric vehicles at work or (professional) vehicles at home with a view to reimbursing employees' expenses. Indeed, the electrification of business fleets implies a change in the business model. Whereas the management of a combustion car fleet is based on a "just-in-time" model (employees have a petrol card or are reimbursed for mileage allowances), the management of an electric fleet is based on an "anticipatory" model which requires that the cars be sufficiently charged at the time of departure to make the journey. Recharging takes place either at the workplace or at the employee's home, and the cost of recharging is passed on to the energy bill at the workplace or at home. It is difficult to distinguish the cost of recharging on bills that aggregate a range of uses and therefore for employers to reimburse or charge for the cost of recharging.

This service combines several technologies to meet this new need for tracing and certifying electric car recharges:

- a blockchain that allows validated charges to be written and information to be stored in a secure and reliable manner;
- communicating objects (IoT) installed in the vehicles;
- a mobile application for employees allowing them to declare the start and end of the vehicle's charge and authorise the cross-referencing of this information with electricity consumption data from the Linky meter (which certifies to the employer the existence of the home charging);
- a web application that allows company managers to monitor recharging.

We will mobilise the results of a survey carried out between November 2020 and January 2021 in the Nantes region of France, concerning the design of this mobility service by a consortium of some 15 operators from three main activity sectors: transport, energy and new technologies, and its experimentation with the employees of these operators.

The survey was carried out in two phases:

- interviews with a dozen designers of the service belonging to the various companies in the consortium (from the world of new technologies, mainly blockchain specialists; from the world of energy, electricity suppliers and distributors; from the world of mobility, transport companies);
- interviews with a dozen or so experimenters of the service (employees of the consortium companies testing the service).

This hybrid collective coalescing around new technologies will be confronted with the issue of privacy. How have the designers dealt with the blockchain's compliance with regulations protecting privacy? Have privacy issues undermined the ways in which trust in the service is built? More generally, has the trajectory of diffusion of blockchain technology been affected by the options chosen?

We will show that the issue of privacy protection gives rise to a series of "tests" in the sense of the sociotechnical approach to innovation [1] which will punctuate the "trajectory" [26] of the service's design and experimentation. These tests give rise to confrontations between actors (on the way they envisage privacy and the use of technologies),but also, to negotiations and new alliances that enable the tensions between privacy and blockchains to be resolved. Here we can observe the social dynamics and normative mediations that run through the trajectories of innovations [3], but these contribute to shaping sociotechnical compromises that are able to articulate the regulatory and acceptability requirements of privacy protection with the particularities of the technology. These compromises are made at the cost of reducing the initial promises associated with the blockchain technology studied here, but allow the emergence of solutions that are the subject of consensus within the consortium of actors. In a first part we will study the way in which the actors, here the developers, manage to define a governance mechanism that meets the GDPR obligations concerning responsibility. Then, in a second part, we will study the solutions found around the management of personal data. Finally, in the last part, we will discuss the challenges that security and explicability represent for this group of actors.

5.1 A First Privacy Test: Defining Governance

Among the general obligations of the GDPR, as soon as the presence of personal data is identified, is the identification of a data controller. This obligation is not self-evident, especially when it comes to blockchain technology. Indeed, blockchain mobilises a series of actors around the data. For example, the "miners" (who validate transactions and create blocks by applying the rules of the blockchain), especially on public blockchains (where anyone can carry out a transaction, participate in the block validation process or obtain a copy of the blockchain) could, in a completely decentralised system, be qualified as a data controller. Nevertheless, the recommendations of the French data protection agency CNIL (2018) on how to define the data controller on the blockchain indicate that "participants, who have a right to write on the chain and decide to submit data to be validated by miners can be considered as data controllers". Despite this indication, the question of how to define a controller has challenged the consortium.

5.1.1 The Appointment of a Controller, a "Test" for the Consortium

This first test concerns above all the world of service design; this world of designers is shared between designers from the digital, electrical and mobility sectors who have joined forces to design this service for tracking the recharging of electric vehicles based on a blockchain. These designers include a range of blockchain specialists from start-ups and large companies who have joined forces in a consortium.

The consortium members initially had a technical reading of the problem of processing responsibility, imagining that start-ups specialising in blockchain technology would take on this role in data governance. The "privacy" deliverable entrusted to one of the consortium's start-ups specialising in blockchain was thought to be a way of delegating the management of the issue to someone specialising in the technology.

> What organisation do we want in terms of GDPR? Who is the controller?" And there, no one raised their hand, whereas we thought it was going to be the start-ups or the software developers. (Designer, electricity sector)

Nevertheless, the lawyers of the large companies participating in the consortium will, in accordance with the recommendations of the CNIL, redirect the responsibility for the processing to the companies that designed the service, rather than to the start-ups, which occupy a position of "subcontractor" within the consortium. The definition of governance puts the consortium to the test, on the one hand because it forces a hierarchy of roles among the members of the consortium who could previously think of themselves as equal partners, and on the other hand because it requires one of the members of the consortium to take responsibility for the processing of the data and run the risk of a penalty (which could potentially be as high as 20 million euros or 4% of the annual worldwide turnover).

> And sometimes this is not necessarily obvious. When there are projects where the stakeholders are somewhat intertwined, to determine who is really responsible for processing, who is a subcontractor, to see if there are potentially cases of joint responsibility, i.e., the parties determine together the purposes and means of processing. And this, all this governance of the GDPR, is not necessarily very simple to apply to a technology such as blockchain either. (Lawyer, electricity sector)

5.1.2 A Form of Recentralisation Contrary to the Imagination of Blockchain Designers

The definition of a data controller thus introduces a form of recentralisation of the consortium's operations by attributing responsibility for processing to one of its members. Responsibility is no longer shared equally among all the members of the consortium, which clashes with the sociotechnical conception [13] associated with the technology by the blockchain designers [7]. Its inventors "*trace or dream of a network and a community operating without intermediaries, claiming a desire for anonymity and total security of transactions*" [15]. The blockchain designers we interviewed testify to this shared ethic with libertarian roots. They see blockchain as a technology that can enable unmediated exchange within a horizontal society and thus forms of democratic administration independent of unrepresentative or failing centralised institutions.

This attachment to decentralised forms of organisation leads them to prefer public blockchains to consortium or private blockchains, which restrict the use of the technology to a small, closed community and hinder its wide dissemination. They regret the choice of creating a consortium blockchain, preferred to a public blockchain by designers from the electricity and mobility sectors, unfamiliar with blockchain and worried about the negative images associated with the technology (particularly with regard to bitcoin in terms of money laundering and energy sobriety) and anxious to keep control of the service being designed.

> For me, when I came into this subject, I said to myself, and I think I'm not the only one who said it to myself, we tried to put blockchain where it wasn't necessarily needed. For me, the pure blockchain use case would be the one that could not be replaced by a centralised system. (Blockchain designer, IT department, Energy World)

The choice of relying on blockchain was not made by the service designers solely on the basis of the technology's properties, but because this technology, which is perceived as having value, attracts public funding (in this case, a call for projects financed by future innovation programmes), which supports innovation on a territorial scale. Blockchain designers believe that the use case does not necessarily lend itself to the use of a blockchain; while other designers are unfamiliar with the properties and promises of the technology.

5.1.3 The Compromise of Choosing the Consortium Blockchain

The designation of a data controller within the consortium, in compliance with the requirements of the GDPR, reinforces in the eyes of blockchain designers the compromise that the choice to create a consortium blockchain represented. It contributes to foregoing the disruptive promise of a perfectly decentralised technology. Nevertheless, the experimentation will displace these representations of the technology to validate its contributions in the eyes of the service designers. On the one hand, blockchain is less costly than managing a centralised platform, mobilising teleoperators who supervise the management of information, which lends credibility to the economic model of the service (which is of little value, since it concerns small transactions, the cost of an electric recharge being low). On the other hand, consortium blockchain appears to be a way of securing data storage and guaranteeing trust within a consortium of various partners.

While the blockchain designers keep the public and decentralised blockchain as their horizon, the other service designers rally around the technology on the basis of its restricted nature, limited to the consortium, and on the classic governance modalities that are associated with data management. The blockchain, backed by the requirements, appears to all the designers of the service as a technique allowing interoperability and guaranteeing compliance.

5.2 Second Privacy Test: Management of Personal Data

The governance and responsibility for processing aims to ensure that personal data is properly handled. Around this service, a large amount of data can be qualified as personal data.

> Personal data is anything that can be linked, directly or indirectly, to a natural person. In the context of the service, this can be, for example, a number plate, an IP (Internet Protocol) address, a telephone number, an e-mail address, a surname, a first name, an identification number, I don’t know, a contract number, for example, for someone who has an electricity contract, that sort of thing. So, this goes very far, i.e., in practice, there is an enormous amount of information that can be qualified as personal data. For example, the load curve of someone, of a customer, of an individual, is personal data, i.e., it is an imprint of his electricity consumption. It is linked to a natural person. (Lawyer, energy sector)

The GDPR aims to guarantee the right to information, deletion, correction and portability of data to those whose data are collected and processed. As we have already mentioned, these rights are difficult to apply on a blockchain because of the immutable nature of the register and the impossibility of deleting what is written on the blockchain.

The designers have resolved this intrinsic contradiction in two ways. On the one hand, by setting up an off-chain storage system, i.e., a data management system independent of the blockchain, and on the other hand by seeking to "minimise" the data that will be registered on the blockchain so that it can no longer be qualified as personal data.

5.2.1 Setting up an Off-Chain System to Store the Data

The implementation of an "off-chain" management system emerged as a compromise solution that was the subject of a form of consensus among the designers of the service. They agree on the practice of not recording personal data on the blockchain, which allows GDPR compliance. This obligation leads to the use of servers, in addition to the blockchain, to manage off-chain personal data.

Designers from the energy and mobility sectors saw this as an opportunity to adhere to a strict legal framework and to curb challenges associated with energy data [12] or geolocation data that informs on user behaviour.

> We must not forget that we are under the spotlight and that although it is an experiment, we are never safe. We know that Linky is a really sensitive subject for the media. And today, the CNIL is not very favourable. It finds that everything that is blockchain is not necessarily protective of personal data. So we have been very vigilant in trying to be as protective as possible. (Designer, energy sector).

But this solution is also valued by blockchain designers because it allows the open nature of the blockchain to be preserved in part and its potential transfer, at a later stage (when the service is industrialised), to a public blockchain. Blockchain designers are distinguished by a very specific conception of the processing of personal data in line with the libertarian ethics that guide their representation of "privacy". They want to allow people to keep control of their data. They anchor this vision in a conception of private property as a property of the self [14] which, when extended to data, and in particular personal data, proclaims the right of each person to dispose of it for themselves. In line with this reading, blockchain technology should make it possible, via the establishment of exchanges between peers, to avoid the constitution of economic monopolies and the capture of the value of data by digital companies. For them, surveillance capitalism [30] is the antimodel that blockchain technology should make it possible to thwart.

Nevertheless, this solution, which articulates blockchain with a classical off-chain data management system, is not optimal in their eyes. They also recommend algorithmic solutions to preserve confidentiality, such as Zero Knowledge Proof methods, referring to their belief in the neutrality of the technology. The algorithmic authority and automation of processes contained in the technology are perceived as guarantees of objectivity. The representations of blockchain actors are therefore part of a form of technical solutionism; they intend to solve the problems of trust in a market through technical solutions. Thus, blockchain designers demonstrate a professional culture that aggregates values, representations and practices specific to the worlds of design [18]. These also shape their reading of privacy.

5.2.2 Data Minimisation

The second direction chosen to manage personal data was to strongly minimise the data recorded on the blockchain. In particular, the charging curves (which, when cross-referenced with the user's charging declaration, certify the existence of a charge) are stored in an off-chain system to comply with the GDPR. Only the duration of the recharge has been recorded in the blockchain, as the duration is not considered as personal data, given that a person cannot be identified from this information alone.

This desire to minimise the data retained for legal reasons allowed the conditions of acceptability by the final users to be considered. Thus, the employees involved in the experiment did not wish to show their employers their recharging hours or to be geolocated (two options that were retained at the start of the experiment). Indeed, this data could inform their employer about their presence at home or their travels; but they see no problem in transmitting the charging times via the service. The blockchain set-up provides "privacy by design" in accordance with the users' reading of it; however, it does not meet the ambition of the blockchain designers to keep the data as close to its owner as possible. The experimentation has made it possible to articulate compliance and acceptability by considering the notion of privacy that users have.

5.3 Third Privacy Test: A User Pathway Tested for Explicability and Security

5.3.1 Three Requests for Consent

The demands of compliance with privacy legislation gave the lawyers a major role in the design of the experiment. The latter argued for a strict, even extensive application of the GDPR by requiring multiple consent requests: via the signature of an experimentation agreement, via the customer management applications authorising access to Linky meter data and via the mobile application during each recharge declaration by the employee.

> "In the end, all that was put in was very classic GDPR. Basically, nothing was created or invented." There, for example, on the application, we told them that they had to tick "I accept"; but inevitably we're also going to make them sign a little paper in which they actually also agree to communicate their load curves." Take a belt and braces approach! (Lawyer, mobility Sector)

For their part, the experimenters believe that the consents collected as part of the experiment to authorise access to and processing of their data do not constitute a guarantee for the user, but rather a guarantee for the institutions that collect and process the data. Transparency, security and data minimisation constitute the triptych of trust with regard to privacy issues as expressed by the experimenters.

> In fact, what goes through my head when I read this kind of thing is: "what information am I disclosing, and to whom?" (User of pilot system, male, mobility sector).

This multiplication of consents, as well as the intertwining of technologies, has contributed to shaping a complex customer journey that is unrealistic for a service that is to be developed industrially.

5.3.2 An Opaque Security Key System

Faced with this complex user journey and their representation of blockchain as a technology that is difficult to explain and controversial, the designers have chosen to make the technology invisible to experimenters.

However, this choice is, in fact, relatively questionable. Users have expressed a series of fears and misunderstandings about the blockchain's key system (each participant has a public key and a corresponding private key: the public key is similar to an identifier, an address; the private key allows the user to sign a transaction. This provides security in the exchange but also privacy by anonymising the identity of the participants in the exchange).

> The only information I have is my profile, public key and all that. I don't have much. I didn't understand what it was for. (User of pilot system, female, mobility sector)

The requirements of the GDPR were thus apprehended through the collection of numerous consents, rather than through the requirement of explicability.

5.4 Conclusion

Confronted with three dilemmas that the designers had to decide upon according to the objectives of the project and the constraints attached to it, the protection of privacy on a blockchain requires the implementation of sociotechnical compromises between designers and experimenters with different representations. The first is decentralisation versus responsibility, which arises around the designation of a data controller.

The second is that of anonymity and identification, which arises around the off-chain storage of personal data. The third is transparency versus confidentiality.

As we have seen from this experiment, managing privacy protection is a skill, based on emerging expertise, distributed across a range of professions and users, which requires collaboration to be implemented.

Privacy is a distributed issue in innovation ecosystems, generating sociotechnical compromises. The service designers from the mobility and energy worlds do not defend a purist vision of technology as absolutely guaranteeing transparency and decentralisation based on an open protocol. Rather, they defend a vision of the technology as a tool for interoperability (making data available in a secure and technically simple way to multiple stakeholders), corresponding to the use case of the experiment, which mobilises data from a variety of sources (meters, production facilities, vehicles, data centres, etc.) operated by multiple players (individuals, SMEs, major accounts, public services, local authorities, etc.). Consortium blockchain has emerged as a technical compromise between these two visions. In the eyes of blockchain designers, these compromises have limited the disruptive potential of blockchain technology, by recentralising data management and losing the open nature of blockchain. However, they have allowed other designers to see unexpected uses and benefits of the technology, such as appearing as a privacy "solution".

> We don't necessarily do without intermediaries, but at least the intermediaries between them have a protocol to trust each other. (Designer, mobility sector)

Ethics Statement Informed consent was obtained from the people interviewed for this study, and their identity has been anonymised. The study protocol was approved by a manager in EDF's research division. Ethics board approval is not required for this type of study in France.

References

1. M. Akrich, M. Callon, B. Latour, *Sociologie de la traduction.* (Presses de l'Ecole des Mines, Paris, 2006)
2. Y. Algan, P. Cahuc, *La société de défiance: comment le modèle social français s'auto-détruit* (PSE Ecole d'Economie de Paris, 2007)
3. N. Alter, *L'innovation ordinaire* (PUF, Paris, 2000)
4. D. Antony, C. Campos-Castillo, C. Horne, Toward a sociology of privacy. Ann. Rev. Sociol. (2017)
5. K.S. Ball, K.D. Haggerty, D. Lyon, *The Routledge Handbook of Surveillance Studies* (Routledge, New York, 2012), pp.1–11
6. L. Barraud De Lagerie, E. Kessous, La mise en marché des données personnelles. In: Steiner P, Trespeuch M (dir.) *Marchés contestés. Quand le marché rencontre la morale* (Presses Universitaires du Mirail, Paris, 2014)

7. J. Bohr, M. Bashir, Who uses bitcoin? An exploration of the bitcoin community. In: 2014 Twelfth Annual International Conference on Privacy, Security and Trust, pp. 94–101. IEEE (2014)
8. F. Castagnino, Critique des *surveillances studies*. Éléments pour une sociologie de la surveillance. Déviance et Société **42**, 9–40 (2018)
9. A. Casilli, Contre l'hypothèse de la fin de la vie privée. La négociation de la *privacy* dans les médias sociaux. Revue française des sciences de l'information et de la communication **3** (2013)
10. A. Casilli, P. Tubaro, Y. Sarabi, (en) *Against the Hypothesis of the End of Privacy : An Agent-Based Modelling Approach to Social Media* (Cham, Springer, 2014), 57 p
11. F. Chafiol, A. Barber-Massin, La blockchain à l'heure de l'entrée en application du règlement général sur la protection des données. Dalloz IP/IT **2017**, 637 (2017)
12. A. Danieli, La "mise en société" du compteur communicant. Innovations, controverses et usages dans les mondes sociaux du compteur d'électricité Linky en France. Sociologie. Université Paris Est Marne-la-vallée (2018)
13. P. Flichy, *L'innovation technique. Récents développements en sciences sociales. Vers une nouvelle théorie de l'innovation* (Paris, La Découverte, 2003)
14. J. Gharbi, C. Sambuc, Propriété de soi et justice sociale chez les libertariens. Cahiers d'économie Politique **62**, 187–222 (2012)
15. C. Gasull, Des racines libertariennes à la bienveillance du monde économique: aperçu des idéologies dans le développement des blockchains. In: Toledano J. (dir.), *Les enjeux des blockchains* (France Stratégie, juin 2018)
16. A. Giddens, *Les conséquences de la modernité* (L'Harmattan, Paris, 1994)
17. O. Lasmoles, La difficile appréhension des *blockchains* par le droit. Revue internationale de droit économique **t. xxxii**(4), 453–469 (2018)
18. S. Levy, L'éthique des *hackers*. Globe (2013)
19. A. Manas, Y. Bosc-Haddad, La (ou les) *blockchain*(s), une réponse technologique à la crise de confiance. Annales des Mines - Réalités industrielles **3**, 102–105 (2017)
20. S. Moatti, Technologies de la confiance. L'Économie politique **75**(3), 5–7 (2017)
21. F. Ost, *Le droit ou l'empire du tiers* (Dalloz, Paris, 2021)
22. R. Sennet, *Les tyrannies de l'intimité* (Seuil, Paris, 1974)
23. D.J. Solove, I've got nothing to hide and other misunderstandings of privacy. San Diego Law Review **44**, 745–772 (2007)
24. B. Rey, *La vie privée à l'ère du numérique*. Lavoisier, coll. Traitement de l'information (2012) 297 p
25. N. Satoshi, Bitcoin: a peer-to-peer electronic cash system. www.bitcoin.org
26. A. Strauss, Hospital and his negotiated order, traduction (1992). Baszanger, I., *La trame de la négociation. Sociologie qualitative et interactionnisme*, Paris L'Harmattan.
27. J. Toledano (sous dir.), *Les enjeux des Blockchains* (France Stratégie, juin 2018)
28. S.D. Warren, D.L. Brandeis, The right to privacy. Harv. Law Rev. **4**(5), 193–220 (1890)
29. C. Zolynski, *Blockchain* et *smart contracts*: premiers regards sur une technologie disruptive, RD banc. fin. 2017. Dossier 4 (2017)
30. S. Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (PublicAffairs, January 2019)
31. R. Sennett, The fall of public man. (New York, Norton, 1974)
32. A.F. Westin, Social and Political Dimensions of Privacy. J. Soc. Iss. **59**, 431–453 (2003). https://doi.org/10.1111/1540-4560.00072
33. A.F. Westin, Privacy and freedom. (New York, Atheneum, 1967)

Chapter 6
Considering Severity of Safety-Critical System Outcomes in Risk Analysis: An Extension of Fault Tree Analysis

David B. Kaber, Yunmei Liu, and Mei Lau

Abstract With the advent of digitalisation and big data sources, new advanced tools are needed to precisely project safety-critical system outcomes. Existing systems safety analysis methods, such as fault tree analysis (FTA), lack systematic and structured approaches to specifically account for system event consequences. Consequently, we proposed an algorithmic extension of FTA for the purposes of: (a) analysis of the severity of consequences of both top and intermediate events as part of a fault tree (FT) and (b) risk assessment at both the event and cut set level. The ultimate objective of the algorithm is to provide a fine-grained analysis of FT event and cut set risks as a basis for precise and cost-effective safety control measure prescription by practitioners.

Keywords Fault tree · Risk assessment · Event consequences

6.1 Motivation

"Digitalisation" in contemporary industrial processes and workplaces has been defined as access to "big data", artificial intelligence (AI) tools/machine learning (ML), and algorithmic methods for the purposes of operation and management of safety-critical systems [5]. In this definition, algorithmic method is a reference to "algorithmic management", or the use of new technological tools for remote workforce organisation and tasking. Such methods have been identified as a developing practice also having the potential to create new "hazards" in complex human-automated systems [5]. As such, algorithmic methods, and the broader technology of AI, imply a need for advances in systems safety practice and, specifically, methods

D. B. Kaber (✉) · Y. Liu
University of Florida, Gainesville, FL 32611, USA
e-mail: dkaber@ise.ufl.edu

M. Lau
John Hopkins Applied Physics Lab, Laurel, MD 20723, USA

J.-C. Le Coze and S. Antonsen (eds.), *Safety in the Digital Age*,
SpringerBriefs in Safety Management,
https://doi.org/10.1007/978-3-031-32633-2_6

for (near) real-time and highly accurate assessments of risks associated with digital process hazards.

In some industrial domains, such as continuous flow manufacturing and petrochemical processing, there has emerged an abundance of digital technologies, including low-cost data sensors and new real-time signal processing devices. The flow of data in these domains has driven the need for development of new AI tools, such as deep-learning neural networks, for real-time system state classification and prescription of control actions, based on large numbers of process measures. One concern that has emerged based on this big data and availability of new computational analytical tools is how management and systems safety practices may be affected. For example, how do we ensure that data streams do not reflect spurious phenomenon and the output of AI algorithms is accurate and valid as a basis for safe operational decision-making.

Having posed this concern, it should be observed that the emergence of big data in some existing industrial applications may not extend to new highly automated processes involving human–autonomy teaming scenarios, as in algorithmic management approaches. That is, there may be an absence of empirical observation of novel work systems and, consequently, sparse data relative to the number of decision variables that must be considered in risk assessment and safety practice. In data analytics, this problem is referred as the "curse of dimensionality", where available process data records are sparse compared with the number of attributes being measured. The curse of dimensionality poses fundamental issues of sensitivity and reliability of any data analytics or statistical analysis. Thus, in some truly unique algorithmic management applications, there may also be a need to identify alternative AI or advanced analytical methods by which to generate field-relevant data/estimates as a basis for application of modeling and decision analysis tools.

It is anticipated that new computational capabilities and AI algorithms may also provide a platform for more computationally sophisticated (real-time) safety analysis methods. For example, more demanding modeling approaches, including high dimensional structures, can be processed by new AI supercomputers in a matter of milliseconds. However, such data-driven computational safety analysis methods may lack the scientific basis and systematic nature of traditional systems safety analysis (SSA) methods. For example, results of ML methods tend to be unique to a specific system, rather than generalisable across application domains. On the other hand, existing traditional SSA methods may not be able to accommodate big data sources. For example, traditional fault tree analysis (FTA) methods do not account for potential event consequences, which can be estimated based on big (process) data.

In general, there is a need for enhanced risk assessment methods to ensure that new management approaches and safety controls, as part of digital industrial processes, are compatible and jointly effective. The specific research question we seek to address is whether new advanced tools can be created for precise projections of safety-critical system outcomes, given high-dimensional decision spaces? Such spaces are characterised by sources of harm in a work system leading to numerous safety-related events (or mechanisms) to negative outcomes. Ultimately, we also need to investigate how such technologies can be exploited to better support work performance and safety.

6.2 Background

A fundamental challenge for safety science has been to determine how to minimise hazard exposure without compromising work productivity. Unfortunately, traditional SSA methods for mitigating exposure have trailed behind industry applications, not even considering the various forms of digitalisation identified by Le Coze and Antonsen [5]. In the early 1980s, the Air Force developed a collection of formal SSA methods as part of MIL-STD-882 [8]. These methods primarily focused on the frequency of hazard exposure vs. severity of outcomes or degree of loss. It took another 14 years for NIOSH (National Institute for Occupational Safety and Health) to develop a guide to further disseminate these methods [2]. Unfortunately, there have been limited advances in formal SSA methods since that time.

The persistence of workplace deaths and injuries underscores the need for enhancement of existing SSA methods to better support industry in precision loss assessment and control when applying digital technologies and process management. This research need is further motivated by economics. The Liberty Mutual (LM) 2019 Worker's Compensation Claims (WCC) Index revealed $46.93 B per year in losses for US industry from the top-10 most disabling worker injuries. The insurance carrier estimated industry costs at over $1 B per week [6]. This level of loss is unacceptable with respect to industry financial stability and US global competitiveness. Comprehensive and valid safety critical engineering methods represent one potential solution to reducing total incident rates (TIRs) in industry.

6.3 Objective

We present a new algorithm for advancing an existing SSA technique, specifically fault tree analysis (FTA), by integrating consideration of the severity of consequences of safety-critical system events for design and operational risk assessment. Such considerations can now be made based on big data sources in digital process scenarios. As the reader may be aware, traditional FTA only accounts for the likelihood of occurrence of system fault states but not the probability of various degrees of system or operational loss, given the occurrence of a fault [2]. Although FTA has realised substantial application in government/military and industrial applications since its conception, the method has always represented an incomplete risk assessment approach. Furthermore, it requires substantial analyst time and effort, even when addressing a single fault event. Another limitation of FTA is that the method considers only the outcome of an identified undesirable and credible ("top") fault event for which an analyst must have prior knowledge. There is no analysis of potential negative outcomes of predecessor (intermediate) events to the top event that are causal in nature. Consequently, any FTA provides an incomplete picture of the total level of system risk posed by a top event, intermediate events and basic sources of harm in a work environment.

The way that these FTA shortcomings have been addressed is by an analyst applying other safety analysis techniques, including but not limited to FMEA (Failure Modes and Effects Analysis (FMEA), Event-tree Analysis (ETA), Probabilistic Risk Assessment (PRA), Cause-consequence Analysis (CCA), etc., in conjunction with FTA [3]. Most of these additional methods make use of FTA outputs to facilitate evaluation of the severity of specific system faults. The common FTA outputs include identification of a minimum set of basic initiators and intermediate events guaranteeing occurrence of the top event [i.e., minimal cut sets (MCSs)] or the exact probability of a top event. Related to this, we only found one study in the literature that addressed severity of hazard exposure in application of FTA. Lindhe et al. [7] presented a dynamic fault tree (FT) with hazard risk determined by means of Monte Carlo simulation [7]. However, any intermediate events in the FT were not identified as having unique outcomes relative to the top event.

In summary, with the advent of digitalisation and big data sources, it may now be possible to predict system event consequences and probabilities for more accurate risk assessment methods. However, there remains an absence of systematic and structured approaches to account for this information as part of FTA. This methodological situation limits the accuracy of SSA as a basis for risk assessment and supporting system design, which should be particularly critical for highly automated and high-risk work environments.

6.4 Review of Traditional FTA

In the traditional FTA approach, we initially construct a FT diagram (see Fig. 6.1) and identify relationships among all process events deduced to contribute to the top event (fault state). Logic (AND/OR) gates are used to create connections between the events (see Fig. 6.2).

We subsequently assign probability values to the occurrence of basic sources of harm in the environment (or initiators). Typically, an analyst makes use of a historical database or safety records for probability data. In the case of novel human–autonomy teaming applications, such data may not be available and, consequently, expert judgements and advanced analytical methods may be necessary to generate probability estimates for analysis (we say more on this below). In traditional FTA, we

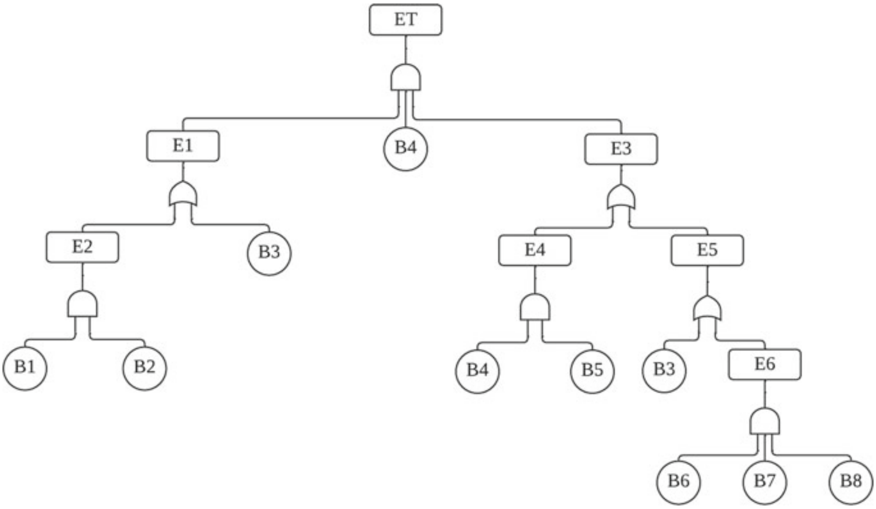


Fig. 6.1 Example FT. *Note* Tree has five layers with eight basic events/initiators (B_i), six intermediate events (E_i), and one top event (ET). An example AND gate integrates inputs from $E1 \cdot B4 \cdot E3$; an example OR gate integrates inputs from $E2 + B3$

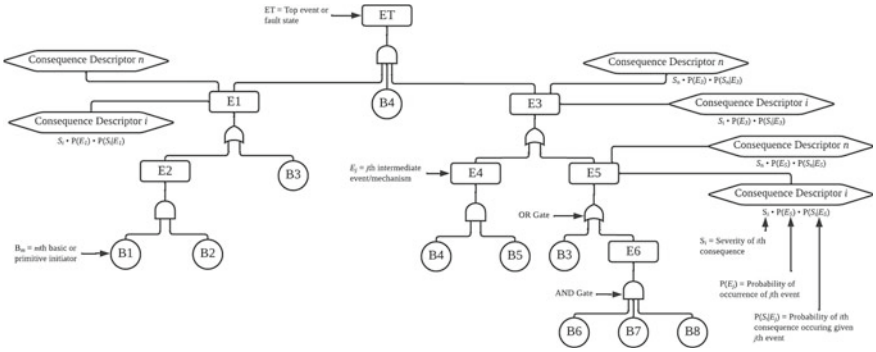

ET = Top event or fault state
ET
Consequence Descriptor *n*
Consequence Descriptor *i*
$S_i \cdot P(E_1) \cdot P(S_i|E_1)$
E1
B4
E3
Consequence Descriptor *n*
$S_n \cdot P(E_3) \cdot P(S_n|E_3)$
Consequence Descriptor *i*
$S_i \cdot P(E_3) \cdot P(S_i|E_3)$
E2
B3
E_j = *j*th intermediate event/mechanism
E4
E5
Consequence Descriptor *n*
$S_n \cdot P(E_5) \cdot P(S_n|E_5)$
Consequence Descriptor *i*
$S_i \cdot P(E_5) \cdot P(S_i|E_5)$
OR Gate
B_m = *m*th basic or primitive initiator
B1
B2
B4
B5
B3
E6
S_i = Severity of *i*th consequence
$P(E_j)$ = Probability of occurrence of *j*th event
AND Gate
B6
B7
B8
$P(S_i|E_j)$ = Probability of *i*th consequence occuring given *j*th event

Fig. 6.2 Each intermediate event in a FT may represent a top event in another FT. For each intermediate event, we specify a set of negative outcomes and levels of associated severity as bases for the CSPIM analysis

calculate the probability of the top event and intermediate events, based on initiator probabilities ($P(B_i)$), and use of Boolean algebra.

The value of the event probabilities is a sufficient basis for calculating the importance of initiating events and cuts sets (minimum combinations of events causing the top event). It is necessary to make use of a method to obtain cut sets (MOCUS) or Boolean equivalent tree construction to identify the MCSs for a FT (but event probabilities are not necessary). For the example tree in Fig. 6.1, we determined three MCSs, including: $\{B_3, B_4\}$, $\{B_1, B_2, B_4, B_5\}$ and $\{B_1, B_2, B_4, B_6, B_7, B_8\}$.

These are guaranteed pathways to the top event or system fault state and represent decision alternatives/priorities for a safety engineer. In traditional FTA, MCSs represent areas of greatest system vulnerability and the most likely targets for safety countermeasures.

For each of the MCSs, we determine a probability of occurrence as well as an importance value and ranking of the cut sets. Traditional FTA limits MCS importance to the impact on the top event probability of occurrence. The probability of a cut set can be calculated as the product of the probabilities of embedded initiators ($P_k = \prod_{B=1}^{m} P_{\mathrm{B}}$), where P_k is the k th cut set probability and $\prod_{B=1}^{m} P_{\mathrm{B}}$ represents the product of the probabilities of m initiators in k th MCS. The importance is the ratio of the MCS probability to top event likelihood: $I_k = \frac{P_k}{P_{\mathrm{T}}}$, where P_{T} is the estimated probability of the top event (based on data or expert judgment) and I_k is the k th cut importance value.

On the basis of this analysis, cut sets can be ranked in terms of importance and the ranking can be used as a basis for safety resource allocation. The MCS with the greatest importance ratio is the most likely set of initiators to cause the top event. An analyst might recommend changing the system structure (or FT event relationships) to increase the number of initiators in the top-rank MCS and reduce the probability of its occurrence. An analyst might also recommend changing system components, materials or procedures to reduce the likelihood of initiators as part of the top-rank MCS.

These traditional FTA outcomes are all fine and good but, as we noted in the literature review, the method does not capture the contributions of initiators to the severity of consequences of the intermediate and top events, nor does it support event risk assessment; i.e., $R_i = P(E_i) \cdot S$, where R_i is the hazard risk for event E_i and S is the severity of outcome of exposure. However, this kind of information is very important in most safety analyses as those initiators primarily contributing to the occurrence of an event may not be one in the same with those factors contributing to the degree of loss, given the occurrence of the event.

6.5 A Consequence Severity-Probability Importance Measure Algorithm for FTA

On the basis of the literature and review of traditional FTA, we propose an algorithmic extension of FTA with the purposes of: (a) analysis of the severity of consequences of both top and intermediate events as part of a FT and (b) risk assessment at both the event and cut set level. The ultimate objective of the algorithm is to provide a fine-grained analysis of FT event and cut set risk as a basis for precise and cost-effective safety control measure prescription by practitioners. Here, it is important to note that we elect to extend the existing FTA method due to its scientific basis and systematic approach to generating results for specific system fault states that may generalise to other domains. These features are desirable relative to a ML/neural network approaches that can only reveal combinations of initiators guaranteeing a system fault but do not identify various event pathways to the fault.

The CSPIM algorithm begins just like traditional FTA, including constructing a FT diagram and assigning probability values to initiators. The FT identifies relationships among the top event and hardware failures, environmental factors, system command faults, etc. For initiator probabilities, we make use of any available system performance data, observations on legacy systems, benchmarking of competing technologies, or expert estimates. In the latter case, estimates are used to address the curse of dimensionality that can occur in FTA, specifically the number of pathways to a top event may exceed the data available for estimating the likelihood of events in a pathway. Consequently, there is a need for estimates and advanced analytical approaches to facilitate accurate decision making on safety controls. Lavasani et al. [4] showed how a fuzzy estimation approach can be applied to judgements of event possibilities by a group of experts. Linguistic terms (of probability) are translated to predetermined intervals of possibility values using fuzzy sets. Possibility estimates are aggregated across experts with mean values being defuzzified to crisp possibility scores. These scores are mathematically transformed to initiator probability values. Here, it is important to note that experts are typically senior systems operators, who are initially interviewed individually for event likelihood estimates. This step is followed by expert group review and discussion on likelihoods in which some individual estimates may be adjusted. Lastly, all estimates are transformed to probability scores, and aggregate values are calculated. This process is most applicable to unique work systems with limited data for safety decision analyses.

At this stage, CSPIM departs from traditional FTA. We identify consequences associated with each event (intermediate and top) and qualitatively categorise consequences according to levels of severity (S). (See Fig. 6.2 for a conceptual representation of this step, as an extension of traditional FTA.) For this purpose, we use the Hazard Severity Category (HSC) scheme provided in MIL-STD-882E (1984; $S_1 =$ “catastrophic”; $S_2 =$ “critical”, $S_3 =$ “marginal”; $S_4 =$ “negligible”). An analyst may also use other severity ranking scales appropriate to the specific industry or develop a custom scale based on prior company safety experience/accidents. The consequences of ET and E_i may occur at many different levels of severity (S_c). Therefore, it is necessary to assign a set of conditional probabilities for levels of severity of outcome for each intermediate event ($P\left(S_1|E_j\right)$; $P\left(S_2|E_j\right)$; $P\left(S_3|E_j\right)$; $P\left(S_4|E_j\right)$) as well as the top event, producing a matrix of severity probabilities for the FT. Ordinarily, these data would also be based on prior system experience but expert judgements can be used as well for novel applications.

As with traditional FTA, the CSPIM algorithm requires that we use initiator probabilities and Boolean algebra to calculate the top event and intermediate event probabilities. These values are then used as bases for calculating and assessing composite event risk values. We calculate the product of the probability of the fault event, the probability of a specific severity of consequence (assuming fault occurrence) and the severity level, written as: $R_{j\mathrm{c}} = P\left(E_j\right) \cdot P\left(S_\mathrm{c}|E_j\right) \cdot S_\mathrm{c}$, where c is the level of consequence severity. As all the variables in this equation should actually appear in a risk matrix for all FT events and all consequence severity levels, the equation should be rewritten using matrix notation as: $\boldsymbol{R_E} = \boldsymbol{P}(\boldsymbol{E}) \cdot \boldsymbol{P}(\boldsymbol{S}|\boldsymbol{E}) \cdot \boldsymbol{S}$.

The composite risk value of events are compared with an established (or custom) Hazard Risk Index (HRI), such as that from MIL-STD-882E. This index facilitates classification of risk outcomes in terms of hazard exposure and severity of outcomes and identifies HRI levels. Using the risk assessment matrix, an analyst codes the composite risk values as representing "low", "medium", "serious" and "high" exposures. The MIL-STD index also provides for general safety control actions based on HRI levels.

At this stage, the analyst is looking for any events that represent serious or high-level risks. According to the MIL-STD, high risks are "unacceptable" and serious risks are "undesirable", which means response countermeasures should be taken as soon as possible to mitigate the probability or severity of a specific outcome.

As with the traditional FTA method, we use MOCUS or Boolean equivalent tree construction, as part of the CSPIM algorithm, to identify the MCSs for the FT. It is also necessary for an analyst to identify any and all intermediate events triggered by initiators as part of different MCSs. In Fig. 6.3, we have marked in "red" the $\{B_3, B_4\}$ MCS as well as the intermediate events triggered by this cut set. This step should also be applied to the other two identified MCSs. (It should be noted that this diagram does not represent a complete visualisation of the new CSPIM outcomes, as modelled in Fig. 6.2. A full CSPIM visualisation includes presentation of all negative consequences for each intermediate, and the top event, along with the severity of outcome indicators and probabilities for each consequence.)

Subsequently, we make a major departure from traditional FTA by determining a composite probability for each MCS, which includes the likelihoods for triggered

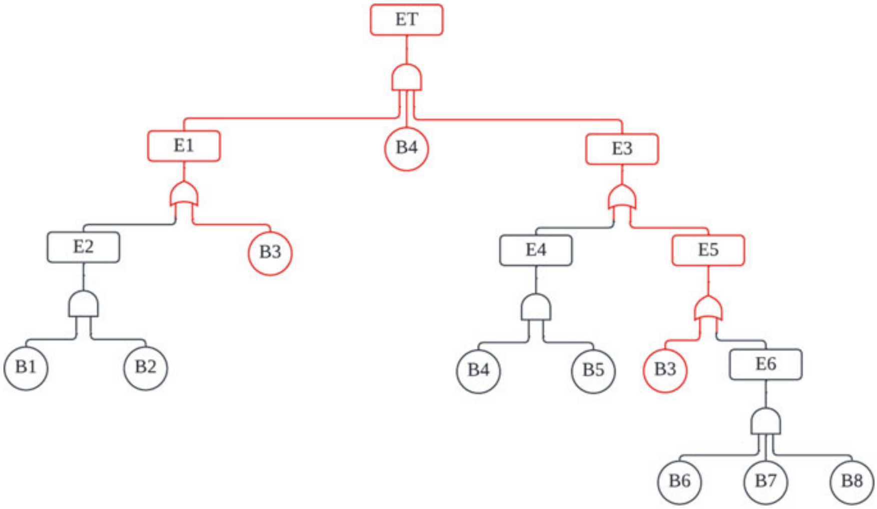


Fig. 6.3 Minimal cut set $\{B_3, B_4\}$ and triggered events (in "red") according to CSPIM method

intermediate and top events. (Here it should be noted that intermediate event probabilities account for the MCS initiator probabilities.) Related to this, the intermediate event probabilities will vary among cut sets depending on the different initiators included in the cut set and activated pathways to the intermediate event. Therefore, it is not appropriate to directly transfer the $P\left(E_j\right)$ value from one cut set to another. We define $P\left(E_{kj}\right)$ to represent the probability of E_j in the kth cut set. The composite probability value accounting for all events associated with the cut set that can lead to negative consequences is then defined as $\boldsymbol{P}(\boldsymbol{E}_k)$ for all j.

We then move on to calculation of the composite risk associated with the MCS. This step requires risk values for all intermediate and top events for the cut set at each level of severity of outcome, given as $R_{kj\mathrm{c}} = P\left(E_{kj}\right) \cdot P\left(S_{\mathrm{c}}|E_{kj}\right) \cdot S_{\mathrm{c}}$. For each specific level of severity (c), the composite risk of event E_j in the kth cut set is noted as $\boldsymbol{R}_{kj}$. We can then obtain the total composite risk for the kth cut set at a given severity level as: $\boldsymbol{R_k} = \sum_{j=1}^{m_k} \boldsymbol{R_{kj}}$, where m_k is the number of events (intermediate and top) associated with kth cut set. These risk values are then used to determine and rank importance ratios for all MCSs.

The importance of each MCS will vary for different outcome severity levels; consequently, all of these values need to be calculated separately. We define the importance of the kth MCS at the cth severity level as the ratio of the composite risk value for the cut set to the total composite risk value for the given system with negative outcomes caused by all triggered events occurring at the cth severity level: $I_{k\mathrm{c}} = \frac{R_{k\mathrm{c}}}{R_{\mathrm{c}}}$, where R_{c} is the composite risk level for the system across all cut sets for the severity level c and $I_{k\mathrm{c}}$ is the importance of kth cut set at the cth severity level. On the basis of this analysis, cut sets can be ranked in terms of importance (as in traditional FTA) and the ranking can be used for safety resource allocation.

Following an almost identical methodology, an analyst can calculate and evaluate the importance of each initiator in an MCS. This additional analysis is used to identify primary risk factors contributing to negative system outcomes and to further refine risk mitigation strategies and target control measures.

As with the event risk analysis, the DoD HRI can be applied to the calculated cut set and initiator risk values. The values are classified as low, medium, serious and high risks, revealing initiators that should be a focal point for control measures to reduce the probability of occurrence or severity of associated event outcomes.

6.6 Conclusions

The contributions of this work include identification of the need for enhanced SSA methods for comprehensive risk analysis in digital industry processes, and a step-by-step algorithm extending traditional FTA. This new method can exercise big data and/or process estimates to account for both the likelihood of hazard exposure and severity of event outcomes and preserves the scientific and systematic nature of the original FTA method. The CSPIM algorithm involves determining risks for all basic, intermediate and top events in a FT based on additional data sources from new AI methods and/or expert judgements and advanced analytical techniques. Implicitly, the algorithm addresses all consequences of intermediate and top events and leads to an overall system risk assessment. The CSPIM algorithm generates new FTA outputs that are sensitive to the severity of event outcomes and allows for prioritisation of fault events for safety controls on the basis of potential risk. These capabilities are currently absent from traditional FTA and substantially enhance the method for new digital industry applications.

Such enhanced SSA methods facilitate higher resolution/precision risk assessment and control formulation maximising industry effectiveness in use of safety resources. At the same time, the CSPIM approach can minimise misallocation of design and administrative measures for safety-critical systems. Consequently, our response to the specific research question identified in the Motivation Section is that it is possible to create new advanced tools for comprehensive safety-critical systems analysis that can accommodate/exploit additional "big" data on event consequences and likelihoods by starting from existing formal SSA techniques.

Although there are advantages to the CSPIM algorithm, the approach is mathematically complex and computationally intensive. In this work, we have only considered application to a small FT; however, in actual industrial applications, such as offshore oil rig design, a FT may literally have thousands of intermediate events included in pathways to an undesired top event of platform failure. In this case, the risk calculations as part of the CSPIM algorithm could be extremely large and complex. Related to this, the recent advances in computational technologies and AI methods may support application of such analysis of high-dimensional system structures, even in the presence of sparse historical data for event probability assignment. As described, expert subjective judgements can be used to generate valuable estimates to support application of such methods when faced with the curse of dimensionality. As noted earlier, in real-world FTA applications, the number of decision variables, or pathways to a top event, can exceed the data available for estimating the frequency of occurrence of specific initiators or intermediate events comprising pathways as well as the data for estimates of event consequences and likelihoods of consequences. New advanced analytical and AI-based methods can be used to generate additional field-relevant data to limit the impact of the curse of dimensionality of safety risk analysis and mitigation strategy decisions.

In general, more SSA tools like the CSPIM algorithm need to be created to support safety engineers in dealing with the broad challenges of digitalisation and algorithmic management in industry, which Le Coze and Antonsen [5] have identified. This represents a challenge for the safety science community, which may be addressed through new computational systems and AI tools that are currently being developed by computer scientists for big data analysis. For example, new AI-based sensing methods may be effective for generating field-relevant data for analysis with advanced risk assessment methods. In addition, new AI-based simulation methods make possible the development of "digital-twins" for highly realistic models of actual systems and processes that support testing of safety controls in advance of actual implementation. These are new digitalisation trends that will have a major impact on the future application of systems safety analysis methods.

References

1. BLS, in *National Census of Fatal Occupational Injuries in 2019* (News Release 12/16/20; BLS, Washington, D.C., 2020)
2. P. Clemens, R. Simmons, *Systems Safety and Risk Management: A Guide for Engineering Educators* (NIOSH Instruction Manual; US Department of Health and Human Services, Cincinnati, OH, 1998)
3. F. Khan, S. Rathnayaka, S. Ahmed, Methods and models in process safety and risk management: past, present and future. Process Saf. Environ. Prot. **98**, 116–147 (2015)
4. S.M. Lavasani, N. Ramzali, F. Sabzalipour, E. Akyuz, Utilisation of fuzzy fault tree analysis (FFTA) for quantified risk analysis of leakage in abandoned oil and natural-gas wells. Ocean Eng. **108**, 729–737 (2015)
5. J.-C. Le Coze, S. Antonsen, Algorithms, machine learning, big data and artificial intelligence, in *Safety in a Digital Age: Old and New Problems*, eds. by J.-C. Le Coze, S. Antonsen (Springer, 2023)
6. Liberty Mutual, in *Workplace Safety Index Reveals Workplace Injuries Cost US Companies Over $1 billion per week* (Press release: Avail. Online 8/19/21; Liberty Mutual, Hopkinton, MA, 2019)
7. A. Lindhe, L. Rosén, T. Norberg, O. Bergstedt, Fault tree analysis for integrated and probabilistic risk analysis of drinking water systems. Water Res. **43**(6), 1641–1653 (2009)
8. MIL-STD-882B, in *System Safety Program Requirements* (Technical report) (Department of Defense, Washington, DC, 1984)

Chapter 7
Are We Going Towards "No-Brainer" Safety Management?

Nicola Paltrinieri

Abstract Industry is stepping into its 4.0 phase by implementing and increasingly relying on cyber-technological systems. Wider networks of sensors may allow for continuous monitoring of industrial process conditions. Enhanced computational power provides the capability of processing the collected "big data". Early warnings can then be picked and lead to suggestion for proactive safety strategies or directly initiate the action of autonomous actuators ensuring the required level of system safety. But have we reached these safety 4.0 promises yet, or will we ever reach them? A traditional view on safety defines it as the absence of accidents and incidents. A forward-looking perspective on safety affirms that it involves ensuring that "as many things as possible go right". However, in both the views there is an element of uncertainty associated to the prediction of future risks and, more subtly, to the capability of possessing all the necessary information for such prediction. This uncertainty does not simply disappear once we apply advanced artificial intelligence (AI) techniques to the infinite series of possible accident scenarios, but it can be found behind modelling choices and parameters setting. In a nutshell, any model claiming superior flexibility usually introduces extra assumptions ("there ain't no such thing as a free lunch"). This contribution will illustrate a series of examples where AI techniques are used to continuously update the evaluation of the safety level in an industrial system. This will allow us to affirm that we are not even close to a "no-brainer" condition in which the responsibility for human and system safety is entirely moved to the machine. However, this shows that such advanced techniques are progressively providing a reliable support for critical decision making and guiding industry towards more risk-informed and safety-responsible planning.

Keywords Uncertainty · Safety assessment · Machine learning · Decision-making

N. Paltrinieri (✉)
Department of Mechanical and Industrial Engineering, NTNU, Trondheim, Norway
e-mail: nicola.paltrinieri@ntnu.no

J.-C. Le Coze and S. Antonsen (eds.), *Safety in the Digital Age*,
SpringerBriefs in Safety Management,
https://doi.org/10.1007/978-3-031-32633-2_7

7.1 Introduction

At the beginning of the 90 s, Prof. Diekmann [7] stated the following. "New analysis tools are emerging, which have the potential to allow complex risk analyses to be performed simply. These new tools, which are underpinned by decision analysis and, lately, expert-systems technology, may lead to powerful, yet simple, approaches to the representation of risky problems". This optimistic prediction on the future of risk analysis was accompanied by the suggestion of a possible interdisciplinary direction: "Future approaches to risk analysis will certainly rely more on the advances being made in Artificial Intelligence (AI) and the cognitive sciences. New computer tools and knowledge-representation schemes will unquestionably lead to new techniques, insights and opportunities for risk analysis".

In the same decade (1997), the Russian chess grandmaster Garry Kimovich Kasparov (former World Chess Champion, ranked world No. 1 from 1984 until his retirement in 2005) lost a chess game with IBM's chess playing computer Deep Blue, which was an example of Good Old-Fashioned Artificial Intelligence (GOFAI) [16]. On that game, [17] later stated the following: "Deep Blue was intelligent the way your programmable alarm clock is intelligent. Not that losing to a 10-million-dollar alarm clock made me feel any better".

Industrial risk analysis and safety management have tried to make use of AI, but they have unevenly progressed since the described events. They neither respected Diekmann's prediction (methodological gaps are still present [24]), nor turned into "programmable-alarm-clock intelligence" thanks to the progressive refinement of machine learning models and the increase in available computing power [12].

This contribution aims to outline what AI can bring to risk analysis and safety management by illustrating a series of examples (with emphasis on benefits and limitations) where AI techniques are used to continuously update the evaluation of the safety level in an industrial system.

7.1.1 Artificial Intelligence and Machine Learning

AI is intelligence demonstrated by machines, and it is divided into subfields based on technical considerations, such as particular goals (e.g., "robotics" or "machine learning"), the use of particular tools ("logic" or artificial neural networks) or deep philosophical differences.

This contribution focuses on the subfield of machine learning (ML). ML refers to techniques aiming to program computers to learn from experience [32]. Some of its models (e.g., deep learning) aim to simulate the learning model of the human brain [12]. Such models are composed of multiple processing layers to learn representations of data with multiple levels of abstraction.

A computer may be trained to assess risk for safety-critical industries such as Oil and Gas through these learning techniques. This would allow processing a large amount of information in the form of indicators from normal operations and past unwanted events (from mishaps to major accidents), which would be used for training. Due to the subjectivity of the definition of risk [40], a risk level cannot be assigned to each event with certainty and expert supervision is needed. Once the model has learned risk categorisation, it uses its knowledge to evaluate real-time risk from the state of the monitored system.

7.1.2 Monitoring of Early Deviations and Past Events

Increasing attention has been dedicated to monitoring safety barrier performance through indicators, as a way to assess and control risk. Indicators may report a series of factors: physical conditions of a plant (equipment pressure and temperature), number of failures of an equipment, maintenance backlog, number of emergency preparedness exercises run, amount of overtime worked, etc. A number of indicator typologies are theorised and used in the literature [23]. Øien et al. [23] affirm that we can refer to risk indicators if: they provide numerical values (such as a number or a ratio), they are updated at regular intervals; they only cover some selected determinants of overall risk, in order to have a manageable set of them. That being said, the latter feature is quickly becoming outdated due to the extensive collection carried out in industry and the attempts to process large numbers of them [30].

Øien et al. [23], Paltrinieri et al. [25, 26] and Landucci et al. [19] have produced several reviews on risk and barrier indicators. They show that the definition and collection of risk indicators have become consolidated practices in "high-risk" sectors, such as the petroleum and chemical industries. For instance, the Norwegian Petroleum Safety Authority (PSA) requires indicators describing the technical performance of safety barriers within the Norwegian Oil and Gas industry since 1999 [31], while the European directive "Seveso III" [9] on the control of major-accident hazards involving dangerous substances suggests their use for sites handling hazardous substances [10]. Such a trend towards the definition and collection of higher numbers of indicators [30] demonstrates the mentioned challenge on big data processing for risk level assessment.

7.2 Examples of AI-Based Prediction

Three examples of AI-based prediction with safety-related purposes are described in the following. The cases depict not only the application of machine learning techniques, but also the criticality of input data and implicitly the human efforts in preparing the data.

7.2.1 Consequence Class Associated with a Hazardous Material Release

ML techniques were applied by Paltrinieri et al. [28, 29] to a database of past accidents with the purpose of simulating its application on the national databases managed by the Seveso competent authorities. The data set used is the Major Hazard Incident Database (MHIDAS) [1] launched by the UK Health and Safety Executive in 1986 and developed by AEA Technology until the mid-1990s. The events included are based on public domain information sources, and their characteristics are registered using keywords.

MHIDAS includes about 8972 hazardous events from 1916 to 1992, with the attributes listed in Table 7.1. Some attributes use a taxonomy to systematically categorise the event. While the actual quality of the data could not be fully verified across the recorded hazardous events, this database is characterised by a high quality of data model, i.e., high semantic quality allowing for clear boundaries and relevant properties of the problem domain and the requirements of the task. Given that it takes a high amount of creativity and vision to design a solution that is robust, usable and can stand the test of time [15], the high semantic quality of MHIDAS could only be reached by significant knowledge and experience of the field.

The attributes listed in the upper part of Table 7.1 were used as inputs to the ML models to predict the consequences—lower part of Table 7.1. The details of data preprocessing are explained elsewhere [34]. The study focused on the number of

Table 7.1 Attributes used to record hazardous events in MHIDAS [1]

Attribute	Description	Category from taxonomy
Date	Date of the event	
Location	Location of event	
Substance	Substances involved in the event	X
Event type	Typology of event	X
Origin	Area of the plant and type of equipment from which the event started	X
Section	Plant section in which the event occurred	X
Quantity	Amount (ton) of released substance	
General causes	General causes the led to the event	X
Specific causes	Specific causes the led to the event	X
Evacuated	Number of people evacuated	
Consequences		
Damage	Economic damage to the property or production loss	
Injured	Number of people injured by the event	
Killed	Number of people killed by the event	

Specific keywords are used to describe some of the attributes

Table 7.2 Severity categories considered by the study

Severity categories	
0	Event with no fatalities
1–10	Event with a number of fatalities between 1 and 10
10–100	Event with a number of fatalities between 10 and 100

people killed and aimed to predict the occurrence of a hazardous event within one of the severity categories listed in Table 7.2 based on the considered inputs. Only categorical data are used.

7.2.2 *Wellhead Damage Frequency in a Drilling Rig*

To avoid potential damage during drilling operations for a new offshore Oil and Gas well, a semisubmersible drilling unit should maintain its position above the wellhead. This is particularly critical if the platform is in shallow waters, where small changes of position lead to higher riser (pipe connecting the platform to the subsea drilling system) angles. Exceeding physical inclination limits may result in damages to the wellhead, Blowout Preventer (BOP—sealing the well) or Lower Marine Riser Package (LMRP—connecting riser and BOP) [5].

Platform position is maintained in an autonomous way (without mooring system) by a set of thrusters controlled by the Dynamic Positioning (DP) system. Input for the DP system is provided by the position reference system (Differential Global Positioning System—DGPS and Hydroacoustic Position Reference—HPR), environmental sensors, gyrocompass, radar and inclinometer [5]. A Dynamic Positioning Operator (DPO) located in the Marine Control Room (MCR) is responsible for constant monitoring of DP panels and screens and carrying out emergency procedures if needed [11]. Platform position may be lost due to several reasons.

In this case study, Paltrinieri et al. [29] assume that the platform thrusters exercise propulsion in a wrong direction, leading to a “drive-off” scenario. If the rig moves to an offset position, specific alarms turn on and suggest that the DPO stop the drive-off scenario by deactivating the thrusters and initiate the manual Emergency Disconnect Sequence (EDS) to disconnect the riser from the BOP. If the manual EDS fails, the automatic EDS activates at the ultimate position limit allowing for safe disconnection [5].

A number of works [21, 24, 26] address the details of occurrence and development of drive-off scenarios. Relevant indicators are defined to assess the performance of safety barriers and related systems. Examples of these indicators are the following.

- thruster control failures in the last three months;
- thruster monitoring sensors failures in the last three months;
- simulator hours carried out by the DPO in the last three months;
- inadequate DPO communication events in the last three months;

- delays in DPO shifts in the last three months;
- percentage of time in the last three months with more than one operator monitoring.

Collection of a wide variety of indicators may lead to challenges related to data integrity. Lack of accurate data may be due to several reasons, such as time and financial constraints experienced by database managers responsible for recording relevant indicators. As companies are expected to do more with less, developers must make decisions about the extent to which they are going to implement and evaluate quality considerations [15].

Simulations of drive-off indicator trends for a period of 30 years can be found in the literature [24]. They are inspired by the typical bathtub curve for technical elements [41] and relevant expert judgement for the remaining elements.

As shown by Bucelli et al. [4], indicator values may be aggregated based on relative weights and hierarchical barrier models, in order to enable dynamic update of barrier failure probabilities. This can be used to update, in turn, occurrence frequencies of potential outcomes. Outcome frequencies are an expression of the scenario probability and, in turn, of the risk. If we assume that the other factors are constant, this represents a simplified risk model. However, Matteini [21] points out a certain complexity within the hierarchical barrier model, which may be due to a tangled structure and an unclear approach to assign relative weights to single model elements. For this reason, a machine learning approach bypassing the construction of such hierarchies and aggregation rules is suggested by Paltrinieri et al. [29].

7.2.3 Alarm Chattering in an Ammonia Plant

Alarm data from a section of an ammonia production process [39] are analysed by Tamascelli et al. [38]. Due to the large quantity of hazardous substances stored and handled during normal activity, the plant has been classified as an “upper tier” Seveso III establishment. Extensive use of methane, hydrogen and ammonia (anhydrous and aqueous solution) occurs in the plant section. Furthermore, due to the intrinsic properties of the processes involved, severe operating conditions (i.e., high pressure and high temperature) are often associated with corrosive substances. Additional information about ammonia production and the considered site can be found at: [2, 42].

The alarm database consists of alarm data collected during an observation period of more than four months. In this case, both data and data model are of high quality as they are acquired from consolidated monitoring systems. Human effort would instead reside in the interpretation and the definition of appropriate priorities among the provided data.

Each row of the database represents an alarm event (26,473 observations in total), and each column (36 in total) represents a piece of information (i.e., an “attribute”) about the alarm. The most meaningful attributes are presented in Table 7.3.

Table 7.3 Alarm database attributes

Attribute	Meaning
Time stamp	Date and time (GMT) of the alarm event
Source	Source that triggered the alarm (measuring instrument, PLC function…)
Jxxx	The safety interlock logic associated with the alarm
Message	The message that is shown to the operator contains the following five attributes: (1) the source; (2) a concise description of the equipment involved; (3) the safety interlock logic (Jxxx); (4) the value and units of measures of the process variable; (5) the alarm identifier (e.g., HHH, HTRP, LLL, LTRP, ACK, etc.)
Active time	Date and time (GMT) of the first alarm occurrence
Data value	The value of the process variable
Eng. unit	The units of measure of the process variable

The Alarm Identifier (point 5. of the "Message" attribute) is a code that defines the alarm status. Examples of Alarm Identifiers are "HHH" (which means that the measured variable has exceeded the "high level" setpoint), "HTRP" (the measured variable has exceeded the "very high level" alarm setpoint and automatic block intervention procedures might be triggered), "IOP" (which indicates an instrumental failure or out-of-range measure), "LLL" and "LTRP" (same as "HHH" and "HTRP" but referring to a "low/very low level").

According to [18], an alarm event is uniquely identified by three attributes only: Time Stamp, Source, and Alarm Identifier (e.g., HHH, HTRP, LLL, LTRP, etc.). The combination of a "source" and an "alarm identifier" is called a "unique alarm".

More than 96% of the alarms registered in the database occurred within one month only, when a considerable number of floods and chattering alarms must have occurred. In fact, only ten alarm sources (out of 194 in total) were responsible for more than 80% of the alarms recorded.

Chattering alarms are alarms "that repeatedly transitions between active state and inactive state in a short period of time" [3]. Therefore, chattering alarms have the potential to produce a large count of alarms and reducing their number is a key step to improve the performance of the alarm system during alarm floods.

Kondaveeti et al. [18] proposed a method for quantifying alarm chatter based on run lengths distributions. Although effective, this technique produces static results (i.e., chattering is quantified based on historical alarm data, but no conclusion can be drawn about the alarm's future behaviour). This Chattering Index approach is modified by Tamascelli et al. [38] to predict chattering behaviour by means of standard ML models.

7.3 Method

ML classification models were used for the three examples in Sect. 7.2. Moreover, comparison among different ML models is also beneficial. Results from multiple linear regression (MLR) were compared to the relatively more sophisticated deep neural network (DNN) models.

Both MLR and DNN aim at modelling the relationship between two or more independent feature variables and a label dependent variable. While the former model fits a linear equation to observed data, the structure of the latter model is similar to the organisation of neurons in the brain, arguably the most powerful computational engine known today [20].

An algorithm uses part of the available data to train the ML model to predict the specific label variable based on the feature variables and test the result on the remaining data. Model performance needs to be evaluated before employing it for actual applications. The result might be far from perfect, and this may be due to poor data quality or indicate the need to tune the model to the actual application.

7.3.1 Metrics

The performance of the classification models used is assessed during the evaluation phase. As an example, consider a situation where accidents must be classified into two classes A or B. A positive prediction occurs when the model predicts the class A. Instead, a negative prediction occurs when the model predicts the class B. Whenever the model predicts the class of an object, there are four possible outcomes:

- TP = True Positive—i.e., predicted label = A, true label = A;
- TN = True Negative—i.e., predicted label = B, true label = B;
- FP = False Positive—i.e., predicted label = A, true label = B;
- FN = False Negative—i.e., predicted label = B, true label = A.

The sum of True Positives and True Negatives represents the number of correct predictions, while the sum of False Positives and False Negatives indicates the number of wrong predictions. True Positives, True Negatives, False Positives, and False Negatives are used to obtain three performance indicators:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{7.1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7.2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7.3}$$

Accuracy represents the fraction of objects that have been correctly classified. Precision indicates the success rate of a positive prediction. Recall denotes the fraction of actual positives that have been correctly identified.

It is worth mentioning that metrics and indicators depend on the probability threshold used by the classification models. For example, if the decision threshold is lowered, the model may produce more positive predictions. As a result, the Recall might increase, but the Precision might decrease [33]. In fact, actions aimed at increasing Recall often lower the Precision, and vice versa [13]. A convenient mean of displaying the effect of the decision threshold is the Precision-Recall curve, i.e., a plot where each point represents the couple Precision vs. Recall at a specific decision threshold [22]. A convenient mean of summarising the information in the Precision–Recall curve is the area under the curve (AUC P-R) [22], which takes values between 0 and 1. Being independent on the decision threshold, the AUC PR is considered a more comprehensive indicator of the model performance if compared with Accuracy, Precision and Recall. In general, a large AUC P-R value indicates good performance [33].

7.4 Results and Discussion

Table 7.4 summarises all the results from the examples described in Sect. 7.2. The results from the two approaches used (MLR and DNN) are directly compared to identify the best predictive performance. MLR shows a higher number of higher values in green cells (9) if compared to DNN (6).

However, this overall result cannot convey the message that MLR performs better than DNN as "there ain't no such thing as a free lunch". In fact, in these examples, DNN was applied with default parameters (e.g., number of layers and nodes

Table 7.4 Summary of results from the representative examples of ML application for safety purposes

	Multi Linear Regression					**Deep Neural Network**				
	C: None	C: 1-10	C: 10-100	WDF	AC	C: None	C: 1-10	C: 10-100	WDF	AC
Accuracy	🟩 0.88	🟥 0.87	0.99	🟥 0.82	🟩 0.95	🟥 0.77	🟩 0.88	0.99	🟩 0.83	🟥 0.94
Precision	🟩 0.90	🟩 0.20	0.00	🟥 0.80	🟥 0.94	🟥 0.89	🟥 0.11	0.00	🟩 0.84	🟩 0.93
Recall	🟩 0.98	🟥 0.07	0.00	🟩 0.91	🟩 0.94	🟥 0.85	🟩 0.09	0.00	🟥 0.86	🟥 0.93
PR AUC	🟩 0.95	🟥 0.22	🟩 0.06	0.92	/	🟥 0.87	🟩 0.27	🟥 0.01	/	/

C stands for consequence class, *WDF* stands for wellhead damage frequency, *AC* stands for alarm chattering, and *PR AUC* stands for the area under the precision recall curve. Greed and Red cells respectively show higher and lower values when compared with the other approach suggested by Tensorflow tutorials [13]). In addition, DNN is relatively more sensitive to poor quality of data [24].

Table 7.4 reports all the metrics discussed in Sect. 3.3. If we exclusively focus on accuracy, we notice that the highest value (0.99) is obtained for both MLR and DNN predictions of the consequence class 10–100 fatalities associated with a hazardous substance release. However, accuracy alone is not informative if the problem involves the identification of rare classes, i.e., when the dataset is class imbalanced [14].

Releases of hazardous substances with 10–100 fatalities are (fortunately) rare events as they represent about 1% of the records in the MHIDAS database. In this case, the models have learned that the result will be correct 99% of the times if they predict that this kind of event never occurs. If the cost of a False Negative is higher than the cost of a False Positive (such as the case of a release of hazardous substance with 10–100 fatalities), Recall is the most meaningful metric. In this context, a good model must produce high Recall, while low precision might be considered acceptable and, to a certain extent, conservative.

The prediction of events with a relatively higher frequency and lower consequence (e.g., a release of a hazardous substance with no fatalities, an increase of wellhead damage frequency or a chattering alarm) may instead benefit from higher precision at the expense of the recall value.

For this reason, rather than considering Precision and Recall individually, one may aggregate them into the so-called F-score [6], especially if the area under the precision recall curve indicates the potential of optimising the model by tuning the decision threshold probability. Human contribution would again come into play in the setting of the algorithm parameters, which would inevitably represent a form of subjective calibration. For this reason, the techniques used in the depicted examples require a deep understanding of their benefits, limitations and application boundaries.

This contribution aims to convey the message that AI-based techniques must be considered as tools supporting and not substituting decision making. Awareness and knowledge of these tools' properties by the user are essential to effectively exploit their results. The role of the human as user of these tools is even more central than before. AI should not be intended as a way to replace the human, but only as an improved approach assisting the human. This is compatible with the concept of trustworthy AI by the European Commission [8] promoting explainable AI (XAI), human centrality by means of interpretability, infobesity (overload of information) avoidance and transparency.

Embracing the principles of trustworthy AI and XAI will unlock the vast potential of machine learning in safety management, especially considering emerging variants of the traditional approaches described in this contribution, such as:

- **Transfer learning**, aiming at developing methods to exploit the knowledge gained in one task (i.e., the *source task*) to address a new task (i.e., *target task*).
- **Federated learning**, machine learning technique that trains an algorithm across multiple decentralised servers holding local data samples, without exchanging them.

- **Meta-learning**, focus on the learning model and its optimisation towards new observations, in order to apprehend the emergence of unknown scenarios (e.g., unknown risks [35, 36]).

Machine learning has the potential to be eventually capable of supporting human users as [7] states. However, the author must admit the presence of another important challenge ahead that is yet to be fully overcome: ensuring appropriate safety culture by the user, i.e., foundations and motivations for which such advanced tools would be used. Once again, this challenge brings the discussion back to humans. Risk and safety analysts and managers would potentially have an advantage in the application of digitalised safety management due to their predefined state of mind, but only given their willingness to learn the basics and use of such advanced and promising techniques.

7.5 Conclusion

This contribution has illustrated examples of AI-based prediction used to continuously update the evaluation of the safety level in an industrial system. The examples refer to the impact prediction of a hazardous substance release in chemical industry, the wellhead damage frequency in offshore oil and gas and chattering alarms in ammonia production. The results can and must be read on different levels, carefully considering the available metrics based on the scenario addressed. This shows that we are not (and will not be in a near future) in a "no-brainer" condition in which the responsibility for human and system safety is entirely moved to the machine. At the same time, an understanding of digital solutions will be progressively required to guarantee their effective application. These advanced techniques have the potential to provide reliable support for critical decision making, guiding industry towards more risk-informed and safety-responsible planning.

References

1. AEA technology—Major hazards assessment unit, in *MHIDAS—Major Hazard Incident Data Service* (UK, 2003)
2. K. Aika, L.J. Christiansen, I. Dybkjaer, J.B. Hansen, P.E.H. Nielsen, A. Nielsen, P. Stoltze, K. Tamaru, in *Ammonia: Catalysis and Manufacture* (Springer Science & Business Media, 2012)
3. ANSI/ISA, 2016. ANSI/ISA–18.2–2016 Management of Alarm Systems for the Process Industries. ANSI/ISA.
4. M. Bucelli, N. Paltrinieri, G. Landucci, Integrated risk assessment for oil and gas installations in sensitive areas. Ocean Eng. **150**, 377–390 (2018). https://doi.org/10.1016/J.OCEANENG.2017.12.035
5. H. Chen, T. Moan, H. Verhoeven, Safety of dynamic positioning operations on mobile offshore drilling units. Reliab. Eng. Syst. Saf. **93**, 1072–1090 (2008). https://doi.org/10.1016/J.RESS.2007.04.003

6. N. Chinchor, MUC-4 Evaluation Metrics, in *Proceedings of the 4th Conference on Message Understanding, MUC4'92* (Association for Computational Linguistics, USA, 1992), pp. 22–29. https://doi.org/10.3115/1072064.1072067
7. E.J. Diekmann, Risk analysis: lessons from artificial intelligence. Int. J. Proj. Manag. **10**, 75–80. https://doi.org/10.1016/0263-7863(92)90059-I
8. EC's High Level Expert Group on AI, *Draft Ethics Guidelines for Trustworthy AI* (Belgium, Brussels, 2018)
9. European Parliament and Council, Directive 2012/18/EU of 4 July 2012 on the control of major-accident hazards involving dangerous substances, amending and subsequently repealing Council Directive 96/82/EC—Seveso III. Off. J. Eur. Union 1–37 (2012)
10. European Parliament and Council, Council Directive 82/501/EEC of 24 June 1982 on the major-accident hazards of certain industrial activities. Off. J. Eur. Union 1–18 (1982)
11. I.C. Giddings, in *IMO Guidelines for Vessels with Dynamic Positioning Systems*. Dynamic Positioning Conference (Houston, Texas, U.S., 2013)
12. I.J. Goodfellow, Y. Bengio, A. Courville, *Deep Learning* (The MIT Press, Citeseer, Cambridge, Massachusetts, US, 2016)
13. Google, Classification: Precision and Recall | Machine Learning Crash Course [WWW Document] (2020)
14. Google, Classification: Accuracy | Machine Learning Crash Course [WWW Document] (2020)
15. J.A. Hoxmeier, Typology of database quality factors. Softw. Qual. J. **7**, 179–193 (1998). https://doi.org/10.1023/A:1008923120973
16. F. Hsu, M.S. Campbell, A.J. Hoane Jr., in *Deep Blue System Overview*, Proceedings of the 9th International Conference on Supercomputing (1995), pp. 240–244
17. G. Kasparov, in *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins* (Hachette, UK, 2017)
18. S.R. Kondaveeti, I. Izadi, S.L. Shah, T. Black, in *Graphical Representation of Industrial Alarm Data*, IFAC Proceedings Volumes (IFAC-PapersOnline) (IFAC, 2010). https://doi.org/10.3182/20100831-4-fr-2021.00033
19. G. Landucci, N. Paltrinieri, A methodology for frequency tailorization dedicated to the Oil and Gas sector. Process Saf. Environ. Prot. **104**, 123–141 (2016). https://doi.org/10.1016/j.psep.2016.08.012
20. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. Nature **521**, 436–444 (2015). https://doi.org/10.1038/nature14539
21. A. Matteini, *Human Factors and Dynamic Risk Analysis: A Case-Study in Oil and Gas Drilling* (Italy, Bologna, 2015)
22. K.P. Murphy, *Machine Learning: A Probabilistic Perspective* (MIT Press, Cambridge, Massachusetts, United States, Adaptive Computation and Machine Learning, 2012)
23. K. Øien, I.B. Utne, I.A. Herrera, Building Safety indicators: Part 1—Theoretical foundation. Saf. Sci. **49**, 148–161 (2011). https://doi.org/10.1016/j.ssci.2010.05.012
24. N. Paltrinieri, L. Comfort, G. Reniers, Learning about risk: machine learning for risk assessment. Saf. Sci. **118**, 475–486 (2019). https://doi.org/10.1016/j.ssci.2019.06.001
25. N. Paltrinieri, G. Landucci, W.R. Nelson, S. Hauge, Proactive approaches of dynamic risk assessment based on indicators, in *Dynamic Risk Analysis in the Chemical and Petroleum Industry: Evolution and Interaction with Parallel Disciplines in the Perspective of Industrial Application* (Butterworth-Heinemann, 2016), pp. 63–73. https://doi.org/10.1016/B978-0-12-803765-2.00006-8
26. N. Paltrinieri, S. Massaiu, A. Matteini, Human reliability analysis in the petroleum industry: tutorial and examples, in *Dynamic Risk Analysis in the Chemical and Petroleum Industry: Evolution and Interaction with Parallel Disciplines in the Perspective of Industrial Application* (Butterworth-Heinemann, 2016), pp. 181–192. https://doi.org/10.1016/B978-0-12-803765-2.00015-9
27. N. Paltrinieri, K. Øien, V. Cozzani, Assessment and comparison of two early warning indicator methods in the perspective of prevention of atypical accident scenarios. Reliab. Eng. Syst. Saf. **108** (2012). https://doi.org/10.1016/j.ress.2012.06.017

28. N. Paltrinieri, R. Patriarca, M. Pacevicius, P. Salvo Rossi, *Lessons from Past Hazardous Events: Data Analytics for Severity Prediction*, eds. by P. Baraldi, F. Di Maio, E. Zio, E-Proceedings of the 30th European Safety and Reliability Conference and 15th Probabilistic Safety Assessment and Management Conference (ESREL2020 PSAM15) (Research Publishing, 2020)
29. N. Paltrinieri, R. Patriarca, E. Stefana, F. Brocal, G. Reniers, Meta-learning for safety management. Chem. Eng. Trans. **83** (2020). https://doi.org/10.3303/CET2082029
30. N. Paltrinieri, G. Reniers, Dynamic risk analysis for Seveso sites. J. Loss Prev. Process Ind. **49** (2017). https://doi.org/10.1016/j.jlp.2017.03.023
31. PSA, Trends in risk level in the petroleum activity (RNNP) [WWW Document] (2016). http://www.psa.no/about-rnnp/category911.html
32. A.L. Samuel, Some studies in machine learning using the game of checkers. IBM J. Res. Dev. **3**, 210–229 (1959). https://doi.org/10.1147/rd.33.0210
33. Scikit-learn.org, Precision—Recall [WWW Document] (2020)
34. R. Solini, *Data Analytics for Chemical Process Risk Assessment: Learning Lessons from Past Events Towards Accident Prediction* (Italy, Bologna, 2017)
35. E. Stefana, N. Paltrinieri, ProMetaUS: A proactive meta-learning uncertainty-based framework to select models for Dynamic Risk Management. Saf. Sci. **138**, 105–238 (2021). https://doi.org/10.1016/j.ssci.2021.105238
36. E. Stefana, N. Paltrinieri, *Meta-learning Potential to Assess Uncertainties in Dynamic Risk Management*, eds. by P. Baraldi, F. Di Maio, E. Zio, E-Proceedings of the 30th European Safety and Reliability Conference and 15th Probabilistic Safety Assessment and Management Conference (ESREL2020 PSAM15) (Research Publishing, 2020)
37. D. Svozil, V. Kvasnicka, J. Pospichal, Introduction to multi-layer feed-forward neural networks. Chemom. Intell. Lab. Syst. **39**, 43–62 (1997). https://doi.org/10.1016/S0169-7439(97)00061-0
38. N. Tamascelli, N. Paltrinieri, V. Cozzani, Predicting chattering alarms: A machine Learning approach. Comput. Chem. Eng. **143** (2020). https://doi.org/10.1016/j.compchemeng.2020.107122
39. H. Topsoe, Ammonia | NH3 | Process | Haldor Topsoe [WWW Document] (2020)
40. V. Villa, N. Paltrinieri, F. Khan, V. Cozzani, Towards dynamic risk analysis: A review of the risk assessment approach and its limitations in the chemical process industry. Saf. Sci. **89** (2016). https://doi.org/10.1016/j.ssci.2016.06.002
41. K.S. Wang, F.S. Hsu, P.P. Liu, Modeling the bathtub shape hazard rate function in terms of reliability. Reliab. Eng. Syst. Saf. **75**, 397–406 (2002). https://doi.org/10.1016/S0951-8320(01)00124-7
42. Yara Italia S.p.A, Relazione di riferimento della Yara Italia S.p.A. dello stabilimento di Ferrara (2016)

Chapter 8
Looking at the Safety of AI from a Systems Perspective: Two Healthcare Examples

Mark A. Sujan

Abstract There is much potential and promise for the use of artificial intelligence (AI) in healthcare, e.g., in radiology, mental health, ambulance service triage, sepsis diagnosis and prognosis, patient-facing chatbots, and drug and vaccine development. However, the aspiration of improving the safety and efficiency of health systems by using AI is weakened by a narrow technology focus and by a lack of independent real-world evaluation. It is to be expected that when AI is integrated into health systems, challenges to safety will emerge, some old, and some novel. Examples include design for situation awareness, consideration of workload, automation bias, explanation and trust, support for human–AI teaming, training requirements and the impact on relationships between staff and patients. The use of healthcare AI also raises significant ethical challenges. To address these issues, a systems approach is needed for the design of AI from the outset. Two examples are presented to illustrate these issues: 1. Design of an autonomous infusion pump and 2. Implementation of AI in an ambulance service call centre to detect out-of-hospital cardiac arrest.

Keywords Artificial intelligence · Healthcare · Safety · Systems perspective

8.1 Introduction

There is a lot of excitement about the potential benefits that the use of artificial intelligence[1] (AI) can bring to healthcare. Health systems are struggling with rising costs, staff shortages and burnout, an increasingly elderly population with more complex health needs, and health outcomes that often fall short of expectations.

[1] The term artificial intelligence, in the widest sense, refers to the science and engineering of intelligent computer systems. In this paper, the focus is mostly on AI applications that use machine learning. Machine learning refers to the use of computer algorithms that learn from data through supervised learning, unsupervised learning or reinforcement learning approaches.

M. A. Sujan (✉)
Human Factors Everywhere Ltd., London, UK

J.-C. Le Coze and S. Antonsen (eds.), *Safety in the Digital Age*,
SpringerBriefs in Safety Management,
https://doi.org/10.1007/978-3-031-32633-2_8

AI is seen as the next step in addressing these challenges, with the hype so high to prompt leading US digital medicine researcher Eric Topol to compile an amusing list of "outlandish expectations" of AI, such as: the ability to diagnose the undiagnosable; treat the untreatable; predict the unpredictable; classify the unclassifiable; eliminate workflow inefficiencies; and cure cancer [29].

There is certainly a strong appetite among governments around the world to promote the use of AI in healthcare. In the UK, a dedicated body (NHSX[2]) has been set up with a remit to accelerate the digital transformation of the National Health Service (NHS) and to support the development and integration of AI applications into the NHS.

Examples of the potential benefit that AI can bring to healthcare can be found readily in news reports and in the scientific literature. Over 200 AI-based medical devices have already received regulatory approval in Europe and the USA [16], and there are many more AI applications that do not require such approvals (i.e., which fall outside the narrow definition of medical devices). The area of diagnostics is particularly strong with examples including AI applications to support identification of diabetic retinopathy [1], skin cancer [7], and, recently, to distinguish COVID-19 from other types of chest infections [12]. Other developments include, for example, ambulance service triage, sepsis diagnosis and prognosis, patient scheduling, and drug and vaccine development.

While these studies provide encouraging results, the evidence base remains weak for several reasons: the focus of the evaluation is usually on a narrowly defined task; the evaluation is typically undertaken retrospectively by the technology developer, and independent evaluation remains the exception; the number of human participants tends to be small; and prospective trials are still infrequent. Taken together, claims that AI outperforms humans are likely to be overstated given the limitations in the study design, reporting and transparency, and the high risk of study bias [17].

The real challenges for the adoption of AI in healthcare will arise when algorithms are integrated into health systems to deliver a service in collaboration with healthcare professionals as well as other technology. It is at this level of the wider system, where teams of consisting of healthcare professionals and AI applications cooperate and collaborate to provide a service, that safety challenges will need to be addressed [26].

The aim of this paper is to review and highlight some of the safety challenges at the system level relevant to the use of AI in healthcare settings by looking back at what we already know from the extensive research on automation, as well as what novel challenges might need further attention. Two examples are described (1. Design of an autonomous infusion pump and 2. Implementation of AI in an ambulance service call centre to detect out-of-hospital cardiac arrest) to illustrate the types of design and implementation issues that should be considered.

[2] NHSX merged into NHS England's Transformation Directorate in February 2022.

8.2 Challenges Old and New

Many of the challenges with using AI in healthcare are actually very familiar. Back in the 1970s and 1980s, industrial systems saw the widespread introduction of automation to improve efficiency and to reduce failures attributed to human error. This was soon accompanied by research studying failures involving highly automated systems, which highlighted the potential for "automation surprises" [22] and the "ironies of automation" [2]. Problems with automation can arise because people are not actually eliminated from the system, but instead the automation changes the nature of the work that people do [18], often resulting in a set of tasks, which were left over by the developers of the automation. This can make the human role and the interaction between people and automation challenging, e.g., due to lengthy periods of monitoring, the need to respond to abnormal situations under time pressure, and the difficulty of building an understanding of different situations and strategies for their management.

However, modern AI systems (especially those that are increasingly autonomous) also present completely new challenges that were not as relevant in the design of traditional automated systems. AI systems can augment what people do in ways that were not possible when machines simply replaced physical work. The interaction with interconnected AI-based systems could potentially develop more into a relationship between people and the AI, especially where the AI has means of expressing something akin to a personality via its interfaces [11]. Social aspects will become much more relevant, as well as mutual understanding of expected behaviours and norms. We might think of, for example, the seemingly ubiquitous voice-enabled virtual assistants (e.g., Amazon's Alexa or Apple's Siri) that aim to deliver a realistic and natural social interaction experience. Examples from healthcare might include mental health chatbots and assistive robots. These relationships between people and technology, along with social, cultural and ethical aspects have much greater importance for future AI-based systems than for traditional automation. Healthcare professionals, patients and AI will increasingly collaborate as part of the wider clinical system.

The use of healthcare AI also raises ethical challenges on a much wider and more fundamental scale compared with traditional automation. For example, concerns about privacy and data protection have come to the fore, such as the controversy and subsequent litigation around the transfer of 1.6 m identifiable confidential electronic patient records from the Royal Free London NHS Foundation Trust to Google subsidiary Deep Mind in 2015. This data transfer was within the scope of a collaboration to develop a tool to support the identification of patients at risk of developing acute kidney injury (which was subsequently abandoned), but the data sharing agreement did not impose any explicit bounds on the use of these patient records. In addition, wider issues around fairness and impact on different stakeholder groups need to be considered, such as racial bias and disparities in accuracy across different population groups [30]. Many data sets are representative of more affluent health systems, and are, therefore, at risk of disadvantaging ethnic minority and vulnerable groups. Fairness at the health system level can also go beyond issues of bias in training data. For example, the use of AI-based patient-facing symptom checkers paired with remote consultations (such as the UK "GP at hand" service offered by Babylon Health) can potentially disadvantage elderly patients and those with significant healthcare needs by shifting and depleting the budget allocated to primary care: these services are typically attractive to younger, healthier populations, leaving traditional primary care services to care for more complex cases with a significantly reduced budget. It is important to note that addressing such concerns requires a broader range of expertise and a social and political dialogue to advance health equity in the age of AI [23].

Designers of AI and healthcare organisations deploying AI should be aware of these critical considerations at the systems level. Examples from the extensive literature have been summarised in a recent White Paper published by the UK Chartered Institute of Ergonomics and Human Factors and include (not an exhaustive list) [27]:

- **Situation awareness**: design options need to consider how AI can support, rather than erode, people's situation awareness. The Distributed Situation Awareness (DSA) model emphasises the systems perspective on situation awareness [24]. According to this model, situation awareness is distributed around the socio-technical system and is built through interactions between agents both human and non-human (e.g., AI). Understanding the situation awareness requirements of each agent can inform the design of the AI and its integration into the wider system.
- **Workload**: the impact of AI on workload needs to be assessed because AI can both reduce as well as increase workload in certain situations. An example of unintended increase in workload is the introduction of electronic health records, which led to situations where clinicians spend around 40% of their time on data entry [10].
- **Automation bias**: automation bias (or automation-induced complacency) describes the phenomenon that people tend to trust and then start to rely on automated systems uncritically [18]. Studies on automation bias suggest that the accuracy figures of AI applications in isolation do not allow prediction of what will happen in clinical use, when the clinician is confronted with a potentially inaccurate system output [13]. Strategies need to be considered to guard against people relying uncritically on the AI, e.g., the use of explanation and training.
- **Explanation and trust**: explainability and transparency of AI decision making might reduce the potential for automation bias. However, there is limited agreement on how to achieve this. Many approaches focus on providing detailed accounts of how an algorithm operates, i.e., to provide explanation of why a decision was made, for example, by reference to salient features [15]. In order for explanation to be fully useful, and to support building and maintaining trust in AI decision making, efforts need to be put into developing interfaces that enable users to interrogate recommendations and to allow dialogue between the user and the AI.
- **Human–AI teaming**: models of teamwork, e.g., the Big Five model (Salas et al., 2005), can provide insights for the design of behaviours (leadership, mutual

performance monitoring, back-up behaviour, adaptability, and team orientation) and supporting mechanisms (shared mental models, mutual trust, closed-loop communication) to enhance human–AI teaming. The design should consider how human team members can understand the AI's roles and responsibilities, and—more challenging—how the AI can understand the human's roles and responsibilities, e.g., in dynamic AI applications that take over human tasks when people are at risk of being overloaded. It is also important that appropriate mental models are shared across human and AI team members.

- **Training**: people require opportunities to practise and retain their skill sets when AI is introduced, and they need to have a baseline understanding of how the AI works. Maintaining core skills is important to provide healthcare workers with the confidence to override and take over from AI applications. Healthcare workers also need to understand potential weaknesses of the AI and how the safe envelope is defined, maintained or breached.
- **Relationships between staff and patients**: the impact on relationships needs to be considered, e.g., whether staff will be working away from the patient once more and more AI is introduced [28].
- **Privacy and ethical concerns**: at the European level, the High-Level Expert Group on AI published "Ethics Guidelines for Trustworthy AI" [9]. The guidelines are based on a Fundamental Rights Impact Assessment and operationalise ethical principles through seven key requirements: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being; and accountability. These ethical requirements necessitate a thorough understanding of stakeholders and their diverse needs and expectations.

Below, two examples are described to illustrate the type of considerations that follow from this line of systems thinking for the design and use of healthcare AI.

8.3 Example 1: AI-based infusion pumps for IV medication administration

This first illustrative example is taken from a project that studied safety assurance challenges of the use of autonomous infusion pumps (i.e., infusion pumps driven by AI) for intravenous (IV) medication administration in intensive care [25]. The purpose of the research was to make recommendations that could feed into the design and implementation of the autonomous technology in such a way that its use enhances rather than diminishes safety.

It has been estimated that as many as 237 million medication errors occur in England every year, and that these cause over 700 deaths [6]. Intravenous medication preparation and administration are particularly error prone. The introduction of highly automated and ultimately autonomous IV medication management systems might contribute to reducing these error rates by taking over functions previously carried out by clinicians, such as safety cross-checks (e.g., patient identity and prescription), calculating infusion rates and independently adjusting infusion parameters based on the patient's physiology. A large UK-US study found that whether or not infusion technology successfully improves patient safety depended largely on the specific context of implementation within the clinical system [3].

The project was carried out in an English NHS hospital, serving a population of 600,000. The intensive care unit (ICU) within the hospital has 16 beds and cares for 1300 patients annually. Patients on ICU are, by default, very ill. Patients can be on life support machines, such as ventilators, and they typically require a significant number of drugs. Some of these drugs are given intravenously via an infusion pump. The infusion pump controls the flow of the drug. The traditional set-up is that a doctor (or clinician with prescribing privileges) prescribes a drug as part of the patient's treatment plan, and a nurse then needs to prepare the drug syringe, load the infusion pump with the drug syringe and then program the infusion pump to run at the required infusion rate for a specific duration. This is the baseline scenario used for illustration in this paper. A more comprehensive description of the analysis is given in [8].

Interviews with patients, healthcare professionals and individuals with responsibility for procurement, IT integration and training were undertaken, as well as an analysis of existing working practices. The interviews and analysis helped to anticipate and explore potential implications for the design of the AI and the impact on the wider clinical system, such as:

- Will clinicians' skills related to medication administration be affected? This relates to training needs in as far as clinicians require opportunities to practice their core clinical skills after the AI has taken over this task. The ICU had already observed a decrease in manual drug dose calculation skills after so-called smart pumps were brought in, which automated this task.
- What new skills do clinicians require, e.g., how to tell if an autonomous infusion pump is working correctly? This also relates to training, but is concerned with how to manage and supervise an AI system, e.g., how to tell the difference between "good" and "bad" AI performance, what to look out for and how to recognise the limitations of the AI.
- How will the relationship between clinicians be affected? The autonomous system replaces the practice of double checking by a second nurse, which often serves also as an opportunity for teaching and discussion.
- How will the relationship between patients and clinicians be affected? The use of autonomous infusion pumps could provide nurses with more time to spend with patients—or nurses might be spending more time managing and supervising autonomous systems away from the patient's bedside. Patients on ICU form very close bonds with their nurses, and they are anxious that nurses might spend more time away from the bedside. Nurses suggest that the operation of (standard) infusion pumps also provides them with an opportunity to do other things concurrently, e.g., check up on the patient's wellbeing and social/psychological needs.

- How will the autonomous system interact with other systems, e.g., other autonomous infusion pumps or the electronic health record, and what will be the impact on the overall IT infrastructure (e.g., in case of failures)? Lack of interoperability of IT systems is a major problem in clinical settings.
- What is the impact on the medication administration task, e.g., does the autonomous system reduce clinician workload by taking over parts of the task or does it increase workload, e.g., due to monitoring and administration requirements? For example, the AI system requires high-quality data, but electronic patient records are often incomplete. This can be potentially safety-critical unless clinicians spend additional time providing that high-quality data to the AI.
- How does the autonomous system impact clinician situation awareness, if clinicians do not manage infusion settings by themselves any longer? Is the autonomous system able to exchange situation awareness with the clinician? Can clinicians easily tell what the system is doing and what kind of situation awareness it has, e.g., through the use of interfaces that explain behaviour and that allow clinicians to explore what the AI is doing?
- What is the impact on the perception of job roles, e.g., on the nursing role? Will nurses be regarded as autonomous clinicians who manage and supervise autonomous infusion pumps potentially away from the bedside, or will nurses' roles change towards more personal caring tasks with less responsibility and authority for managing medications?

These considerations at the level of the clinical system can support designers of AI applications in defining the operating environment and in understanding relevant interactions with people, other tools and systems, other tasks that might be relevant and the characteristics of the local work environment.

8.4 Example 2: AI to support the recognition of out-of-hospital cardiac arrest

The second example is concerned with the implementation of an AI support system in an NHS ambulance service to improve the recognition rate of and time to recognition of out-of-hospital cardiac arrest (OHCA). Currently in the UK, approximately 30,000 people sustain an out-of-hospital cardiac arrest (OHCA) annually, and survival to hospital discharge ranges from 2.2 to 12% [20]. Early defibrillation within the initial 3–5 min could deliver survival rates of 50–70%, but each minute of delay to defibrillation reduces the probability of survival by 10% [19]. Hence, speedy recognition of OHCA by the ambulance service call handler is crucial to support bystander cardiopulmonary resuscitation and to enable fast paramedic attendance at the scene. However, recognition of OHCA is difficult, because signs can be subtle, and the international evidence demonstrates that around 25% of OHCA are not picked up by call centre operators [5].

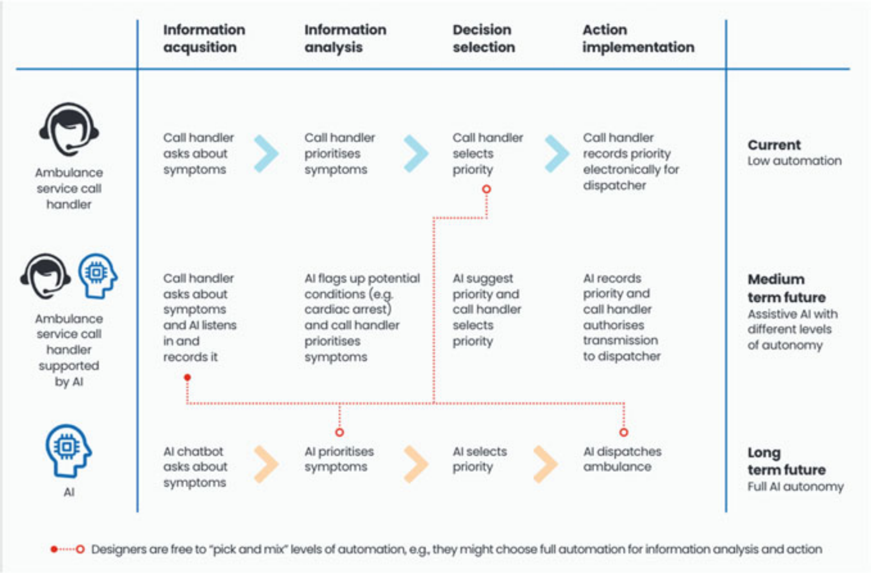

Information acqusition
Information analysis
Decision selection
Action implementation
Ambulance service call handler
Call handler asks about symptoms
Call handler prioritises symptoms
Call handler selects priority
Call handler records priority electronically for dispatcher
Current
Low automation
Ambulance service call handler supported by AI
Call handler asks about symptoms and AI listens in and records it
AI flags up potential conditions (e.g. cardiac arrest) and call handler prioritises symptoms
AI suggest priority and call handler selects priority
AI records priority and call handler authorises transmission to dispatcher
Medium term future
Assistive AI with different levels of autonomy
AI
AI chatbot asks about symptoms
AI prioritises symptoms
AI selects priority
AI dispatches ambulance
Long term future
Full AI autonomy
Designers are free to "pick and mix" levels of automation, e.g. they might choose full automation for information analysis and action

Fig. 8.1 Different levels of automation and interaction in the implementation of AI support for the recognition of out-of-hospital cardiac arrest[3]

An AI system to support ambulance service call handlers in recognising OHCA has been developed by a Danish manufacturer, and initial independent retrospective evaluation with data from Copenhagen produced encouraging results demonstrating that the AI system had higher sensitivity than call handlers (84% vs. 75%), but slightly lower specificity (97% vs. 99%) [5]. However, a subsequent randomised controlled trial of the technology in use found that the AI support did not lead to improved recognition of OHCA [4].

Again, this reinforces the need for consideration of the wider system when designing and implementing AI technology. Taking a systems perspective can help understand the breadth of design decisions and their potential impact, especially when considering the implementation of the AI tool in a different context (in this case using the technology in a more rural environment as opposed to the urban environment where it was initially tested). For example, the interaction between ambulance service call handler and the AI can be designed to accommodate different levels of support (or levels of automation); see also Fig. 8.1:

- No AI support (current situation).

[3] Figure from the white paper "Human Factors and Ergonomics in Healthcare AI", 2021, reproduced with permission from the UK Chartered Institute of Ergonomics and Human Factors. All rights reserved.

- AI operates autonomously (full automation), e.g., an AI chatbot interacts with the caller, asks for symptoms, prioritises the symptoms and selects call priority, and then dispatches an ambulance according to the call priority
- Several levels of support and interaction in-between, e.g., the call handler leads the conversation with the caller, but the AI picks up symptoms and prioritises these, the call handler makes the call priority decision, and the AI dispatches the ambulance.

Interviews with ambulance service staff and a cognitive task analysis [14] of call handlers' tasks suggested that different types of interaction design might have far-reaching consequences that need to be considered:

- To what extent should the AI communicate to the call handler the reasoning behind its decision making, e.g., should the AI simply pop up an alert suggesting that it recognised a potential cardiac arrest, or should it provide a running commentary on what it considered for that decision?
- How will call handlers know about whether the AI is making good or bad decisions?
- How will the interaction with the AI affect call handler workload? Will looking at AI alerts increase or decrease workload?
- How will the false positive rate of the AI affect call handler trust in the system? Will call handlers disregard the AI input or will they start over-relying on it?
- How will call handlers' skills in recognising cardiac arrested be affected?

In addition to the above questions about the interaction between the AI and the call handler, it is also not clear how the AI best augments what the call handlers do, i.e., how to support call handlers with difficult decisions. For example, the cognitive task analysis identified issues that make OHCA recognition more challenging as well as strategies that call handlers employ to overcome these difficulties:

- there are difficulties in understanding what is being said, e.g., the caller has poor mobile phone reception, does not speak the language or has speech impairments;
- the caller is in a panic and unable to provide a coherent description;
- the caller might be hesitant to provide accurate description of the patient's condition, e.g., a close relative being in shock and denial; or the caller might use ambiguous and contradictory language;
- the caller is hesitant to provide cardiopulmonary resuscitation;
- strategies that call handlers employ include aiming to calm down the caller, asking clarifying questions, listening to background noises (e.g., patient breathing), and using synonyms to describe symptoms (e.g., "is the patient gulping for air like a fish out of the water?" to establish whether the patient is breathing sufficiently).

8.5 Conclusion

The aspiration of using AI to improve the efficiency of health systems and to enhance patient safety requires a transition from the predominant technology-centric focus that contrasts people and AI ("humans vs. machines") towards a systems approach that considers AI as part of the wider health system.

Several lessons can be learned from research and practical experiences with the design and operation of highly automated systems. However, advanced AI systems also present novel challenges around social and relational aspects, and human–AI teaming. Addressing these requires a multidisciplinary approach as well as a broader political and societal dialogue around fairness and values in algorithms. This should be reflected in policies of research funding bodies and regulators, because funding specifications and regulatory frameworks frequently only reflect the technology-centric perspective of AI rather than reinforcing a systems approach.

There is a need to raise awareness of these issues among healthcare stakeholders, because Human Factors and Ergonomics (HF/E) and safety science, which advocate a systems approach, are currently not sufficiently well embedded in health systems.

Acknowledgements This work was supported by the Assuring Autonomy International Programme, a partnership between Lloyd's Register Foundation and the University of York (HF/AI and ASSIST demonstrator grants).

Ethics Statement The study described in Example 1 received institutional approval at the participating NHS hospital as a service evaluation study. Prior to the interview, potential participants received a participant information leaflet. Interviews took place in a meeting room at the hospital (patients and hospital staff), over the telephone or at the business offices of the interview participant. Participation was voluntary, and all participants provided written consent. The study described in Example 2 was approved by the Health Research Authority and Health and Care Research Wales (IRAS reference 21/HCRW/0002).

References

1. M.D. Abràmoff, P.T. Lavin, M. Birch, N. Shah, J.C. Folk, Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. Npj Dig. Med. **1**, 39 (2018)
2. L. Bainbridge, Ironies of automation. Automatica **19**, 775–779 (1983)
3. A. Blandford, P.C. Dykes, B.D. Franklin, D. Furniss, G.H. Galal-Edeen, K.O. Schnock, D.W. Bates, Intravenous Infusion Administration: A comparative study of practices and errors between the United States and England and their implications for patient safety. Drug. Saf. **42**, 1157–1165 (2019)
4. S.N. Blomberg, H.C. Christensen, F. Lippert, A.K. Ersbøll, C. Torp-Petersen, M.R. Sayre, P.J. Kudenchuk, F. Folke, Effect of machine learning on dispatcher recognition of out-of-hospital cardiac arrest during calls to emergency medical services: A randomized clinical trial. JAMA Netw. Open **4**, E2032320–E2032320 (2021)
5. S.N. Blomberg, F. Folke, A.K. Ersbøll, H.C. Christensen, C. Torp-Pedersen, M.R. Sayre, C.R. Counts, F.K. Lippert, Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. Resuscitation **138**, 322–329 (2019)

6. R.A. Elliott, E. Camacho, F. Campbell, D. Jankovic, M.M. St James, E. Kaltenthaler, R. Wong, M.J. Sculpher, R. Faria, in *Prevalence and Economic Burden of Medication Errors in the NHS in England* (Policy Research Unit in Economic Evaluation of Health & Care Interventions, Sheffield, 2018)
7. A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**, 115 (2017)
8. D. Furniss, D. Nelson, I. Habli, S. White, M. Elliott, N. Reynolds, M. Sujan, Using fram to explore sources of performance variability in intravenous infusion administration in ICU: A non-normative approach to systems contradictions. Appl. Ergonom. **86** (2020)
9. High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI* (European Commission, Brussels, 2019)
10. R.G. Hill, L.M. Sears, S.W. Melanson, 4000 Clicks: A Productivity Analysis of Electronic Medical Records in a Community Hospital ED. Am. J. Emerg. Med. **31**, 1591–1594 (2013)
11. International Organization for Standardization 2020, Iso/Tr 9241-810:2020 Ergonomics of Human-System Interaction. Part 810: Robotic, Intelligent and Autonomous Systems. ISO (2020)
12. L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, K. Cao, D. Liu, G. Wang, Q. Xu, X. Fang, S. Zhang, J. Xia, J. Xia, Artificial intelligence distinguishes Covid-19 from community acquired pneumonia on chest CT. Radiology, 200905 (2020)
13. D. Lyell, E. Coiera, Automation bias and verification complexity: A systematic review. J. Am. Med. Inform. Assoc. **24**, 423–431 (2016)
14. L.G. Militello, R.J.B. Hutton, Applied cognitive task analysis (acta): A practitioner's toolkit for understanding cognitive task demands. Ergonomics **41**, 1618–1641 (1998)
15. T. Miller, Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell. **267**, 1–38 (2019)
16. U.J. Muehlematter, P. Daniore, K.N. Vokinger, Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): A comparative analysis. Lanc. Dig. Heal. **3**, E195–E203 (2021)
17. M. Nagendran, Y. Chen, C.A. Lovejoy, A.C. Gordon, M. Komorowski, H. Harvey, E.J. Topol, J.P.A. Ioannidis, G.S. Collins, M. Maruthappu, Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. BMJ **368**, M689 (2020)
18. R. Parasuraman, V. Riley, Humans and automation: Use, misuse, disuse, abuse. Hum. Factors **39**, 230–253 (1997)
19. G. Perkins, A. Handley, R. Koster, M. Castrén, M. Smyth, T. Olasveengen, K. Monsieurs, V. Raffay, J. Gräsner, V. Wenzel, Adult basic life support and automated external defibrillation section collaborators. European resuscitation council guidelines for resuscitation 2015: Section 2. Adult basic life support and automated external defibrillation. Resuscitation. **95**, 81–99 (2015)
20. G.D. Perkins, S.J. Brace-Mcdonnell, The UK out of hospital cardiac arrest outcome (OHCAO) project. BMJ Open **5**, E008736 (2015)
21. E. Salas, D.E. Sims, C.S. Burke, Is there a "big five" in teamwork? Small Group Res. **36**, 555–599 (2005)
22. N.B. Sarter, D.D. Woods, C.E. Billings, Automation surprises, in *Handbook of Human Factors and Ergonomics*, ed. G. Salvendy (Wiley, 1997)
23. L. Sikstrom, M.M. Maslej, K. Hui, Z. Findlay, D.Z. Buchman, S.L. Hill, Conceptualising fairness: Three pillars for medical algorithms and health equity. BMJ Health and Care Informatics **29**, E100459 (2022)
24. N.A. Stanton, R. Stewart, D. Harris, R.J. Houghton, C. Baber, R. Mcmaster, P. Salmon, G. Hoyle, G. Walker, M.S. Young, M. Linsell, R. Dymott, D. Green, Distributed situation awareness in dynamic systems: Theoretical development and application of an ergonomics methodology. Ergonomics **49**, 1288–1311 (2006)

25. M. Sujan, D. Furniss, D. Embrey, M. Elliott, D. Nelson, S. White, I. Habli, N. Reynolds, Critical barriers to safety assurance and regulation of autonomous medical systems, in *29th European Safety And Reliability Conference (ESREL 2019)*, ed M. Beer, E. Zio (CRC Press, Hannover, 2019a)
26. M. Sujan, D. Furniss, K. Grundy, H. Grundy, D. Nelson, M. Elliott, S. White, I. Habli, N. Reynolds, Human factors challenges for the safe use of artificial intelligence in patient care. BMJ Health and Care Informatics. **26**, E100081 (2019b)
27. M. Sujan, R. Pool, P. Salmon, Eight human factors and ergonomics principles for healthcare AI. BMJ Health and Care Informatics (2022a)
28. M. Sujan, S. White, I. Habli, N. Reynolds, Stakeholder perceptions of the safety and assurance of artificial intelligence in healthcare. SSRN (2022b) [Online]. Available from: https://doi.org/10.2139/Ssrn.4000675
29. E. Topol, *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again* (Hachette, New York, 2019)
30. J. Wawira Gichoya, L.G. Mccoy, L.A. Celi, M. Ghassemi, Equity in essence: A call for operationalising fairness in machine learning for healthcare. BMJ Health and Care Informatics **28**, E100289 (2021)

Chapter 9
Normal Cyber-Crises

Sarah Backman

Abstract Despite an increasing scholarly interest in cyber-security issues, the phenomenon of large-scale cyber-crises affecting critical infrastructure is largely unexplored. While some characteristics of its consequence dynamics have been identified—prominently its transboundary features—the underlying conditions that allow such dynamics to unfold have not yet been thoroughly explored. This chapter aims to contribute to bridging this gap by applying the classical theoretical perspectives of Normal Accidents (NA) and High Reliability organisations (HRO) on the sociotechnical systems of modern critical infrastructure. It argues that NA characteristics (the combination of interactive complexity and tight coupling) can be found in multiple layers of critical infrastructure operations (technical, cognitive, organisational and macro). Implications are discussed in terms of its connection to transboundary crisis dynamics.

Keywords Cyber-security · Cyber-crises · Critical infrastructure · High reliability organisation theory

9.1 Introduction

Globalisation and fast-paced technological innovation have given rise to a worldwide digitalisation over the last 20 years. The impact of the internet and cyberspace on modern societies is so vast that some commentators argue that we have entered a new era and “cyber age”, characterised by rapid change [1]. In this development, cyberspace has become an enabler of economic and social prosperity, but also of vulnerability and transnational security concerns. An especially serious concern has been centred around the increasing connectivity and interconnectedness of critical infrastructures. These worries have been enhanced by high-profile cyber-incidents

S. Backman (✉)
Department of Economic History and International Relations, Stockholm University, Stockholm, Sweden
e-mail: sarah.backman@ekohist.su.se

J.-C. Le Coze and S. Antonsen (eds.), *Safety in the Digital Age*,
SpringerBriefs in Safety Management,
https://doi.org/10.1007/978-3-031-32633-2_9

and crises, including the NotPetya attacks in 2017 affecting organisations in over 60 counties including the global transport and logistics firm Maersk, the pharmaceutical giant Merck and the National Bank of Ukraine, the WannaCry attacks on the UK National Health Services in 2017 and the ransomware attack on an American oil pipeline system in 2021.

Meanwhile, research contributions on the phenomenon of large-scale cyber-crises have been surprisingly thin and scattered within the field of crisis management research and beyond, reflecting the state of macro-level cyber-security research in general. Despite an increasing cyber-security interest within research fields dealing with national and international security affairs, attempts to theorise this essentially complex and multidisciplinary issue have been fragmented. As noted by Green [2], this fragmentation is problematic because of the "blind spots" of each perspective. "Thus lawyers have little idea of the technology that they are trying to regulate, strategists do not pay enough heed to the wider ethical and legal implications of acts of interstate cyber-aggression; and computer scientists delineate the intricacies of the technology with little focus on its political and strategic implications" [2, p. 3]. In other words, there is a shortage of comprehensive (although necessarily less detailed) scholarly perspectives on cyber-security issues in general, and cyber-crises in particular.

Although still relatively slim, the current body of literature focusing on cyber-security at a macro-level has illuminated some important aspects of the consequence dynamics of cyber-crises. One is that cyber-crises (as cyber-security issues in general) tend to blur important dichotomies, including internal/external, technical/strategic and civilian/military, making them difficult to analyse, frame and conceptualise [3, 4]. This has resulted in, for example, early national cyber-crisis strategies that differed significantly from each other, despite facing largely the same challenges [5].

Another aspect identified in the literature is that the consequences and response efforts of cyber-crises tend to be characterised by transboundary-ness, to the extent that cyber-crises can be conceptualised specifically as transboundary crises [6, 7]. For example, this manifests in moments of rapid escalation after moments of slow development and a quick increase of involved and affected actors [8]. These consequence dynamics tend to put substantial stress on national crisis management structures that were not built with transboundary crises in mind [9, 10]. Previous cases have highlighted that response challenges that are challenging in any crisis (especially if the crisis shows transboundary features), like performing joint sense-making, coordination of response efforts and effective crisis communication, become even more challenging in cyber-crises. Not least because of the added complexity of a crisis which essentially involves technical matters, but has societal consequences [6, 7].

However, less attention has so far been placed on understanding the underlying conditions that allow this transboundary dynamic to unfold. This chapter begins to address this research gap, drawing upon the classic theoretical perspectives of Normal Accidents (Perrow) and High Reliability organisations [11, 12].

9.2 Normal Accidents and High Reliability Organisations

Since the publication of Charles Perrow's book Normal Accidents in 1984, the idea that tightly coupled complex systems will inevitably cause accidents (making them "normal") has been highly influential across many academic disciplines, especially those concerned with technology, risk, safety and crisis.

The idea of Normal Accidents (NA) is based upon the combination of two system conditions: interactive complexity and tight coupling. System in this sense is loosely defined and can refer to a computer system as well as an organisational system. Interactive complexity of a system starts with a lot of different components. These components can be technical parts (software or hardware, for instance), but they can also refer to procedures or human operators. Within this setting, failures among system components can interact in unexpected ways. Due to the complexity of the system and all its components, few if anyone (designers of the system included) can predict the many ways that failures in different components can interact and the consequences of these interactive failures [13]. In itself, interactive complexity is not a major problem, unless combined with what Perrow refers to as "tight coupling". If a system is complex but not tightly coupled, it means that even if failures interact in unexpected ways within this system, there is enough "slack" within the system to have time to figure out how to do things in other ways (where it is possible to do things/operate the system in other ways). When a system is both complex and tightly coupled, it means that the interactive failures, for some reason, cannot be isolated from each other, and that there is no alternative way of operating. This means that the disturbances within this system will spread quickly and "cascade" [14].

When the integral system characteristics of interactive complexity and tight coupling are present, accidents will inevitably (although perhaps rarely) happen due to multiple and unexpected interaction of failures. According to Perrow, neither new technological solutions nor better organisation can totally undo this dynamic, since the added complexity (either organisational or technological) from these "fixes" will then be part of the possible interactive failures within the system [14]. Decentralisation is required to deal with unexpected interactions in tightly coupled and complex systems. The problem is that systems cannot be decentralised and tightly coupled at the same time and there are strong economic incentives to keep and extend the tight coupling [14, 15].

While the perspective of NA drew attention and gained popularity both within and outside of academia in the 1990s, critical reactions and perspectives also emerged. One of the prominent stemmed from Todd La Porte and a group of Berkeley researchers. These scholars highlighted the fact that some organisations experience virtually no accidents despite the presence of interactive complexity and tight coupling (referred to as High Reliability organisations, or HROs), thus challenging the idea that organisational "fixes" cannot prevent accidents. A common finding of this research has been that HROs seem to allow flexible and decentralised decision making, have strong external preferences for failure-free operations and invest heavily in reliability improvement, including redundancy and training. The cost of failure is high in these organisations [15]. Bierly et al. [16] argue that HROs in general share two main characteristics, besides interactive complexity and tight coupling, that set them apart from other organisations: catastrophic potential (which increases scrutiny and expectations of accident-free operations) and accountability (linked to clear areas of responsibility, control and expectations of performance) [16].

In more recent contributions of NA application, two main trends can be identified. The first is the notion that despite some differences, the perspectives of NA and HRO can largely be viewed and used as complementary to understand the complex dynamics of accidents and safety in high-risk systems [15, 17]. The second departs from the observation that the world becomes ever more interconnected and complex, with global, multiorganisational, large-scale systems that are managed by a plethora of private and public actors. Thus, contributions within this trend aim to extend the classical theoretical arguments of both NA and HRO beyond technological systems and organisations to the macro-level, or the level of "organisation of organisations". This chapter draws on both trends.

The main argument of this chapter is that the consequence dynamics of large-scale cyber-crises, characterised by their transboundary nature, can be explained by NA-dynamics in several layers of the sociotechnical systems that comprises modern critical infrastructure operation. Through five interviews with senior experts on cyber-security and critical infrastructure and two case examples (the incident involving the Ukraine power grid in 2015 and the Kaseya incident in 2021), this chapter will explore how these dynamics can manifest on several layers of critical infrastructure operation. In doing so, this chapter aims to contribute to bridging the gap between the understanding of the consequence dynamics of cyber-crises and the structural conditions in sociotechnical systems of critical infrastructure that allow them to unfold and cascade as observed in previous research [3, 6, 7]. The interviewees included senior experts from Sweden, the UK and the USA, all with a background of working with national level cyber-security and critical infrastructure protection. Beyond the interview data, this chapter also used media reports and official reports as material.

The analysis in the following empirical sections will be loosely structured around four analytical categories, or layers, of NA application, identified by Le Coze [17]: 1. Technology, 2. Cognition, 3. Organisation and 4. Macro.

9.3 Analysis

9.3.1 Technology

The first category of NA application, as relevant for critical infrastructure operations, is technology. Perrow realised rather early the applicability of NA to the internet, which is essentially composed of technical systems (both hardware and software) with interdependent components in interactive complexity [see for example Perrow

[18]]. However, the extent of digitalisation of society and the development of ICS (Industrial Control Systems) in critical infrastructure has come a long way since then. Today, an analysis of the NA dynamic applied to critical infrastructure calls for a focus on the problem of legacy code, systems and hardware.

A legacy system can be defined as “An information system that may be based on outdated technologies but is critical to day-to-day operations. As enterprises upgrade or change their technologies, they must ensure compatibility with old systems and data formats that are still in use” [19].

In modern critical infrastructure, it is not uncommon to have legacy systems and code more than 15 years old underpinning operations. Some legacy code is written in old and outdated programming languages, like COBOL, which relatively few programmers know today. Moreover, many legacy systems were built without security in mind, making security considerations an “add on” aspect, or afterthought. When the systems underpinning ICS in critical infrastructure are built upon layers upon layers of legacy code, and systems written in a variety of code languages, with numerous add-ons to make them compatible, interactive complexity is continuously built into the system as a whole. Thus, components with the potential to fail, and interact with other failing components in unexpected ways, are continuously added.

Getting rid of legacy code, systems and hardware is often exceedingly expensive, which is one of the factors why many of them operate way beyond their intended lifetime. As one of my interviewees explained: “There is lots of legacy code and legacy hardware out there. I’ve been to places where they can’t find a vendor to replace the hardware any more, and places that buy things from online auction sites, because that’s cheaper than to upgrade to something modern, and they just don’t have the funding” (Interviewee 1).

NA are expected when the characteristic of interactive complexity is paired with tight coupling: an inability to isolate subsystems and interdependent systems from each other, or to stop them. This means that failures will cascade until a major part of the system, or all of it, fails [14]. Previous procedures to decrease the degree of tight coupling in Industrial Control Systems (ICS) of critical infrastructure, such as creating an “air gap” between the system and the internet, are made more difficult as the demand for digital transformation and efficiency of industrial organisations increases. Instead, modern ICS networks may be connected both to third parties and the wider organisation [20].

In the words of one expert commentator: “Legacy systems are often maintained only to ensure function, and their operations are often digitized with upgraded Internet of Things (IoT) functionality for the sole purpose of operability. OT maintenance may fail to consider the IT and cybersecurity perspective, seeking to make changes to improve systems without questioning if those systems remain secure. While these legacy systems may seem helpful after years of use, networked systems’ prolonged exposure to these legacy devices proves time and time again the familiar adage: What can go wrong will go wrong” [21].

Moreover, as one interviewee highlighted, legacy technologies might be maintained because the processes they underpin are too important to risk being disrupted even for a short amount of time: “Many organisations are afraid of swapping out legacy technologies for new ones because they are afraid that it will disrupt the production process or cause it to fail" (Interviewee 4).

9.3.2 Cognition

A key problem connected to the cybersecurity of critical infrastructure operations is that the actual danger characterised by NA characteristics in large-scale systems is not always easily perceived. This difficulty is partly because the components interacting are not only technical but include organisational and human components too, making the complex interactions and interdependencies between components and subsystems more difficult to understand and estimate. In the words of Grabowski and Roberts: "In general, large complex systems are difficult to comprehend as a whole. Therefore, the tendency is to decompose or factor them into smaller subsystems, which can lead to the development of a large number of subsystem interfaces" [22]. One of the interviewees of this study highlighted the difficulty of getting a comprehensive picture of all the subsystems involved in critical infrastructure operations: "many operators and roles are very specialized now. You only understand one small part of the system and worry about that. However, all the parts must be included and compared to achieve a common model and understanding. Some parts may affect or even disturb other parts in unexpected ways. You must build your operation on a comprehensive analysis including supply chain dependencies. But this is currently lacking when it comes to critical infrastructure" (Interviewee 2).

According to the findings of HRO research, commitment to reliability is a key feature of managing the danger of the combination of interactive complexity and tight coupling, and this commitment is connected to a common understanding of the potential of catastrophic consequence [11]. As one interviewee argued, this commitment to reliability does not appear to be widespread when it comes to critical infrastructure operations and cyber: "It seems that when it comes to cybersecurity and digitalization of critical societal functionalities, we have not learned much from the high-risk industries. In those industries, we are happy to let security cost whatever is necessary. This is not the case with cybersecurity yet, despite that healthcare services (for example) could be disrupted nationwide due to a zero-day vulnerability. We are not yet at the point where we allow digitalization to be expensive due to security concerns" (Interviewee 3).

The difficulty of grasping cyber related vulnerability in critical infrastructure can also be enhanced by the fact that some systems (and system components) are critical and high risk, but many are not (and thus would not require strict security and safety measures in accordance with HRO). However, due to interactive complexity, system components that appear to be non-important could potentially be contributing to the failure of a system that is, thus making the distinction between critical and non-critical more challenging.

9.3.3 Organisation

A common finding of both classical HRO and NA research has been to point to the importance of reducing tight coupling in systems through "organisational slack". This can be done by achieving structural flexibility and redundancy, which involves duplication or multiple and independent ways of operating, communicating and making decisions [11].

The 2015 cyber-attack on the Ukraine power grid is an example of how centralisation can make accidents more consequential, but is also an example of how redundancy in organisations can reduce the same. In December 2015, one of Ukraine's power grid providers was taken down by a cyber-attack, leaving 230,000 customers across various areas without electricity for several hours. Through a sophisticated, long-term attack campaign earlier that year, including spear-phishing methods, hackers had succeeded in taking remote total control of the ICS of at least three energy distribution companies. In disrupting power through remotely switching breakers, the attackers also disabled backup power supplies to all but one distribution centre in order to hinder operators from reaching out and giving or receiving information about the evolving situation. Finally, they launched a distributed denial-of-service (DDoS) attack on the customer service centre.

Despite the sophistication and novelty of the attack, the operators were able to restore service within 3–6 h by moving over operations to manual control [23]. The possibility of going manual was highlighted in a later report as an important mitigation mechanism. It also highlighted that those utilities that are more reliant on automation might not be able to do this [24]. An interviewee echoed this: "The reason this attack was not more disruptive was that there were parts of the grid which was not digitalized, which means it was possible to move to manual operating mode" (Interviewee 5).

Applying the NA-dynamics perspective, interactive complexity allowed the hackers to gain total access and control over the ICS of the energy grid. They identified, attacked and exploited many individual technical and human components of the system in unexpected ways, and once they were in, they were able to "cascade" their access due to tight integration of the system [25]. However, this case is also an example of the impact of redundancy and organisational slack on limiting the tight coupling of the system and thus the full effects of the NA dynamics. By being able to operate the grid manually, the tight coupling of the system was reduced, and operational capacity was restored before the situation could develop into a serious, long-lasting energy crisis.

9.3.4 Macro

The interdependent linkages and interactive complexity that surround critical infrastructure services on a macro-level involve a complex ecology of supply chain actors and other critical infrastructure-sectors. These interdependencies may be difficult to analyse and regulate. This affects the ability to detect NA characteristics and possible cascade paths of disruptions. It also makes it difficult to implement coherent regulation across the transnational structures of critical infrastructure. As noted by Grabowski & Roberts: “In large-scale systems, subsystems are often characterized paradoxically by both autonomy and interdependence. At one level, subsystems exist and operate independently of other systems, resources, and interference, and they are often responsible for their own survival, success, and growth. Thus, they appear to be rather autonomous entities, requiring little coordination. At the same time, subsystems are also interdependent” [22].

This dynamic can be exemplified by the tendency of critical infrastructure services to be dependent on supply-chain actors for upholding its digital systems (including legacy systems). As one interviewee explained: “The overall fundamental security of the system will be dependent on the particular company running the legacy technology to be updating its software to be compatible with the latest operating systems from Microsoft and other companies. For example, if the technology only runs off Windows XP and cannot be upgraded to Windows 10, because the company that created it does not support that, or no longer exists, then you have a fundamentally vulnerable system” (Interviewee 4).

The combination of centralisation, interactive complexity and tight coupling in the systems of organisations that underpins the functionality of critical services (through, for example, the reliance on supply chain actors) as well as in the technological systems of critical services is continuously exploited by cyber-threat actors who use this dynamic to launch ransomware and supply-chain attacks. Interactive complexity creates the possibility for a cyber-exploit to spread quickly and unexpectedly, and centralisation in combination with tight coupling enhances the impact of the attack, putting the victim under more pressure to pay the ransom (especially if the victim provides a critical societal service such as energy, water or food distribution).

An example of this can be found in the case of the recent REvil (also known as Sodinokibi) ransomware attack, affecting at least hundreds of businesses worldwide, including the grocery chain COOP in Sweden, in early July 2021. The attackers managed to use a vulnerability in Kaseya’s VSA software to bypass security measures and distribute malware (ransomware) to its customers [26]. Through the centralisation of IT-infrastructure service delivery, the attackers could focus on targeting this one business to get to connected businesses further down the line. In the case of COOP, there was no alternative way of operating without access to their digital payment system (no organisational slack/redundancy), and they had to simply close most of their 800 grocery stores in Sweden until the problem was solved [27].

In other words, REvil used the conditions of centralisation of service providers (making it possible to effectively spread malware to many businesses by exploiting a single supply-chain actor), interactive complexity (using different interactive components to achieve unexpected consequences) and tight coupling (leveraging the victim’s dependency on the ransomed digital systems in order to force them to pay the ransom) to achieve their goals.

9.4 Conclusion

This chapter aimed to contribute to the understanding of the transboundary characteristics of large-scale cyber-crises by suggesting that they can be explained by the existence of NA dynamics (the combination of interactive complexity and tight coupling) in several layers of the sociotechnical systems that support modern critical infrastructure operations. With support from the insights of classical NA and HRO theory, it explored the application of these arguments. Analysing the technical layer, the chapter highlighted the problem of interactive complexity and tight coupling in the sociotechnical systems underpinning critical infrastructure, especially through legacy code, systems and hardware. Analysing the cognitive layer, it pointed to the difficulty of clearly perceiving the danger stemming from NA dynamics in large-scale systems. At the organisational layer, it highlighted the 2015 cyber-attack on a Ukraine power grid as an example of how centralisation can make accidents more consequential, but also an example of the mitigating effect of operational redundancy measures to reduce tight coupling. Finally, analysing the macro-layer, the chapter used the case of the recent REvil attacks to discuss how the macro-NA dynamics including the reliance on supply-chain actors, can be exploited by cyber-threat actors and create cascading consequences and transboundary cyber-crises.

Ethics Statement This work adhered to the research ethics that are stipulated in the "Stockholm University research integrity and ethics policy" that complies with relevant legislation regarding ethical conduct of research. Informed consent was obtained from participants, and all data have been anonymised. Ethics board approval is not required for this kind of study in Sweden.

References

1. O.R. Young, J. Yang, D. Guttman, Meeting cyber age needs for governance in a changing global order. Sustainability. **12**(14) (2020)
2. J.A. Green (ed.), *Cyber Warfare: A Multidisciplinary Analysis*, 1st edn. (Routledge, 2015).
3. S. Boeke, National cyber crisis management: Different European approaches. Governance (2017). https://doi.org/10.1111/gove.12309
4. H. Carrapico, A. Barrinha, The EU as a coherent (cyber) security actor? JCMS: J. Common Market Studies **55**(6), 1254–1272 (2017). https://doi.org/10.1111/jcms.v55.6, https://doi.org/10.1111/jcms.12575
5. J. Collier, Strategies of cyber crisis management: Lessons from the approaches of Estonia and the United Kingdom, In *Ethics and Policies for Cyber Operations: A NATO Cooperative Cyber Defence Centre of Excellence Initiative*, 187–212 (2017).
6. S. Backman, Conceptualizing cyber crises. J. Conting. Cris. Manag. **00**, 1–10 (2020)
7. M.F. Prevezianou, Beyond ones and zeros: Conceptualizing cyber crises. Policy Stud. Organ. **12**, 51–72 (2021)
8. A. Boin, The transboundary crisis: Why we are unprepared and the road ahead. J. Conting. Cris. Manage. **27**(1), 94–99 (2019)
9. A. Boin, M. Rhinard, Managing transboundary crises: What role for the European Union? Int. Stud. Rev. **10**, 1–26 (2008)
10. E. Olsson, Transboundary crisis networks: The challenge of coordination in the face of global threats. Risk Manage. **17**, 91–108 (2015)

11. T. La Porte, High reliability organizations: Unlikely, demanding and at risk. J. Conting. Cris. Manage. **4**(2) (1996)
12. T. La Porte, A strawman speaks up: Comments on the limits of safety. J. Conting. Cris. Manage. **2**(4) (1994)
13. C. Perrow, *Normal Accidents: Living with High-Risk Technologies* (Princeton University Press, 1999)
14. C. Perrow, The limits of safety: The enhancements of a theory of accidents. J. Conting. Cris. Manage. **2**(4) (1994)
15. H. Brown, Keeping the lights on: A comparison of normal accidents and high reliability organizations. IEEE Technol. Soc. Mag. (2018)
16. P. Bierly, S. Gallagher, J.C. Spender, Innovation and learning in high-reliability organizations: A case study of United States and Russian nuclear attack submarines, 1970–2000. IEEE Trans. Eng. Manage. **55**(3) (2008)
17. J.-C. Le Coze, 1984–2014. Normal accidents. Was Charles Perrow right for the wrong reasons? J. Conting. Cris. Manage. **23**(4) (2015)
18. C. Perrow, *The Next Catastrophe: Reducing Our Vulnerabilities to Natural, Industrial, and Terrorist Disasters* (Princeton University Press, 2007). http://www.jstor.org/stable/j.ctt7t4c1
19. Gartner (2021) https://www.gartner.com/en/information-technology/glossary/legacy-application-or-system
20. Walker, Legacy systems in a connected world: Securing critical infrastructure (2020). https://manufacturingglobal.com/technology/legacy-systems-connected-world-securing-critical-infrastructure
21. Schrader, Prevention is the only cure: The danger of legacy systems (2021) https://www.darkreading.com/vulnerabilities---threats/prevention-is-the-only-cure-the-dangers-of-legacy-systems/a/d-id/1341075
22. M. Grabowski, K. Roberts, Risk mitigation in large-scale systems: Lessons from high reliability organizations. California Manage. Rev. **39**(4) (1997)
23. Sternstein, DHS: Cyberattack on the Ukraine power grid could happen here (2016) https://www.nextgov.com/cybersecurity/2016/04/dhs-ukraine-cyberattack-power-grid-could-happen-here/127262/
24. Assante, Confirmation of a coordinated attack on the Ukrainian power grid (2016). https://www.sans.org/blog/confirmation-of-a-coordinated-attack-on-the-ukrainian-power-grid/
25. C. Perrow, *The Next Catastrophe: Reducing Our Vulnerabilities to Natural, Industrial, and Terrorist Disasters* (Princeton University Press, Princeton, NJ [u.a.], 2011)
26. Newman (2021) https://www.wired.com/story/revil-ransomware-supply-chain-technique/
27. SVT (2021) https://www.svt.se/nyheter/inrikes/coop-tvingas-halla-stangt-aven-under-sondagen
28. C. Ansell, A. Boin, A. Keller, Managing transboundary crises: Identifying the building blocks of an effective response system. J. Conting. Cris. Manage. **18**(4), 195–207 (2010)
29. P. Pawlak, A. Missiroli, Introduction: Trends, patterns, and challenges for international cooperation in cyberspace. Eur. Foreign Aff. Rev. **24**(2), 125–134 (2019)
30. J.P. Hauet, P. Bock, R. Foley, R. Françoise, Ukrainian power grids cyberattack (2017). https://www.isa.org/intech/20170406/

Chapter 10
Information Security Behaviour in an Organisation Providing Critical Infrastructure: A Pre-post Study of Efforts to Improve Information Security Culture

T.-O. Nævestad, J. Hovland Honerud, and S. Frislid Meyer

Abstract The study examines whether information security behaviour (ISB) in an organisation providing critical infrastructure improved after systematic efforts to improve information security culture (ISC) through the implementation of an information security management system (ISMS). The data are based on quantitative surveys before ($N = 323$) and after ($N = 446$) efforts to improve ISC in the organisation. Qualitative interviews were also conducted before ($N = 22$) and after ($N = 12$). The study finds that the organisation has managed to improve its ISC through systematic efforts over a two-year period (2014–2016), and that this also has led to improvements in ISB among the personnel in the organisation. Multivariate regression analyses indicate that ISC is the most important variable influencing ISB, while ISMS measures is the most important variables influencing ISC. Thus, our results indicate that it is important to work with ISMS and ISC to increase IS in our increasingly digitalised society, especially in organisations providing critical infrastructure.

Keywords Information security management system · Information security culture · Critical infrastructure

T.-O. Nævestad (✉) · S. F. Meyer
Institute of Transport Economics, Oslo, Norway
e-mail: Tor-Olav.Naevestad@toi.no

J. H. Honerud
University College of Southeast Norway, Kongsberg, Norway

J.-C. Le Coze and S. Antonsen (eds.), *Safety in the Digital Age*,
SpringerBriefs in Safety Management,
https://doi.org/10.1007/978-3-031-32633-2_10

10.1 Introduction

10.1.1 Background

One of the key aspects of increased digitalisation of society's functions, especially those related to critical infrastructure, is that it introduces new vulnerabilities, indicating the need for new types of protection. An important insight in this respect is the critical importance of human and cultural factors for the security level of critical infrastructure. According to Lim et al. [1], security is becoming more challenging in today's business because people are both a cause of information security (IS) incidents as well as a key part of the protection from them. In this context, physical and technological measures provide an insufficient strategy for protection. Several studies indicate the critical importance of information security behaviour (ISB) for IS in organisations, suggesting that these behaviours often reflect more general patterns of information security culture (ISC) in the organisations [1–3].

We define ISBs as behaviours that are relevant to IS. IS is often defined as protection against breaches of confidentiality, integrity and accessibility. This applies to information that is oral, written or electronic. Confidentiality refers to ensuring that only those who are authorised to access information, accesses it. Integrity refers to protecting the accuracy and entirety of information and processing methods. Accessibility refers to ensuring that authorised users have access to the information and associated equipment when necessary. We define ISC as shared and information security relevant ways of thinking or acting that are (re)created through the joint negotiation of people in social settings.

According to Chen et al. [2], previous studies have paid little attention to the important influence of ISC on ISB. Additionally, Chen et al. [2] state that there is little research on the relationship between comprehensive efforts to manage IS and ISC. We may also refer to such efforts as an information security management system (ISMS), which defines policies and procedures to ensure, manage, control and continuously improve IS in an organisation. One of the most prevalent ISMS is the ISO 27001 standard, which involves systematic efforts to ensure confidentiality, integrity and accessibility.

10.1.2 Aims

In this chapter, we address the research gap identified by Chen et al. [2] by studying whether the implementation of an ISO 27001 compliant ISMS (and additional measures) has led to changes in ISC and subsequently IS in an organisation providing critical infrastructure.

The aims of the study are to:

(1) describe the organisation's efforts to improve ISC through the implementation of an ISMS;
(2) compare ISC in the organisation before (2014) and after (2016) the efforts to improve ISC, and examine factors influencing ISC in the organisation;
(3) compare ISB in the organisation before (2014) and after (2016) efforts to improve organisational ISC, and examine factors influencing ISB in the organisation.

We compare implementation and effect in the six departments of the organisation.

10.1.3 Previous Research

10.1.3.1 Information Security Management System

An ISMS consists of formal routines and measures that enable the organisation to work systematically to avoid breaches of confidentiality, integrity and accessibility, by establishing formal safety policies and goals, establishing important roles and responsibilities, conducting risk analyses systematically gathering information on incidents and dangers, developing countermeasures, monitoring the effects of these and adjusting measures if necessary (cf. Mitsch et al. [4]).

Other key elements in an ISMS are role descriptions with responsibilities, reporting systems, risk assessments, security training, security procedures, etc. The security policy states security goals and how these are to be achieved via ISMS. The procedures for achieving the goals are documented, along with who is responsible for doing what.

Although the implementation of an ISMS, e.g., ISO 27001, is often cited as an important way of establishing an ISC, there seems to be few studies which have actually examined this relationship (cf. [2]). The relationship between management systems and culture is, however, well established in the research on safety culture [5].

10.1.3.2 Information Security Culture

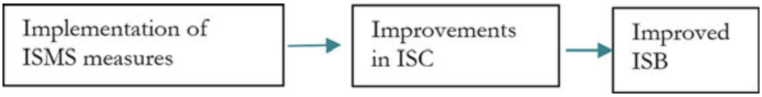
While an ISMS refers to the formal aspects of IS management ("how things should be done"), as it is described in procedures, manuals, etc., the informal aspects of IS management generally refer to ISC ("how things are actually done") [6]. ISC is often studied using various concepts and models of organisational culture [1]. Although Ruighaver et al. [7] note that the organisational security culture concept has gained recognition, they also underline the lack of consensus on definitions and concepts (cf. [8]). Additionally, they assert that in spite of a large amount of research on organisational security and how it should be improved, this research only focuses on certain aspects of security, and not how these aspects can be analysed as part of a larger organisational culture [7]. Based on this understanding, they choose to draw on organisational culture research in their analysis of ISC. This approach is similar to that applied by scholars studying organisational safety culture, who analyse safety culture as a focused and safety-relevant aspect of the larger organisational culture (e.g., [6]). Based on this, we may also analyse ISC as "security-relevant" aspects of the larger organisational culture, defined and conceptualised using models of organisational culture. When studied qualitatively, ISC refers to common frames of reference that form the basis for interpretations of actions, hazards and our own identity, and which motivate and legitimise behaviour that affects IS (cf. [6, 9]). Such common frames of reference arise through interaction in groups. When studied quantitatively, ISC is measured as the way IS is valued in the organisation by managers and employees and their perceptions of the ISMS, or "the way things actually are done" when it comes to IS.

10.1.3.3 Information Security Behaviour

Lim et al. [1] cite several studies indicating that IS problems in organisations have been linked to employee behaviour. These behaviours are typically different types of violations and non-compliance with IS procedures, indicating that it is not sufficient to have a formal system in place, if it is not supported by the ISC [1]. This indicates the importance of viewing IS behaviours as part of a larger ISC context.

10.1.3.4 Theoretical Model and Hypotheses

Based on previous research, we have developed the following theoretical model and hypotheses:



(1) **Hypothesis 1**: implementation of ISMS measures will lead to improvements in ISC;
(2) **Hypothesis 2**: the departments with the best ISMS implementation will have the best ISC improvements;
(3) **Hypothesis 3**: improvement in ISC has led to improvement in ISB.

10.2 Methods

10.2.1 Qualitative Interviews

We used a semistructured and relatively open interview guide, focusing on security work in the organisation since 2014. The interviews were built up around the following main topics: (a) IS measures since 2014 and ISMS implementation, (b) follow-up of the 2014 ISC survey and (c) managers and employees' perceptions of IS rules and measures.

10.2.2 Quantitative Survey

10.2.2.1 Survey Items

The survey contains a set of background questions (e.g., gender, age, experience, education). The survey also includes questions measuring ISMS measures and implementation, e.g., questions about each department's follow-up of the survey results in 2014 and information about specific IS issues over the past two years (e.g., passwords, security policy for mobile units, policies for strangers in the premises) (cf. Sect. 3.1.2).

In this study, we choose to reformulate one of the few existing universal organisational safety culture scales, the GAIN scale [10] for safety culture, into an organisational security culture scale. The questionnaire contains 24 questions concerning, e.g., respondents' perceptions of management's and employees' focus on information security, reporting of information security issues. Respondents can rate the questions from 1 (totally disagree) to 5 (totally agree). Thus, a security culture index with a minimum value of 24 (1 × 24) and a maximum value of 120 (5 × 24) can be compared across companies and sectors.

We have one question measuring ISB: "When I am asked for information, I always think carefully about whether the information can be used for purposes other than its intended purpose".

10.2.3 Samples

We compare the results of two surveys done over a period of just over two years; the first in the spring of 2014 and the second in the autumn of 2016. Response rates are provided in Table 10.1.

Table 10.1 Response rates

Department	Respondents	Employees	Response rate (2016)	Response rate (2014)
1	84	112	75%	79%
2	26	31	84%	–
2	62	85	73%	69%
4	38	54	70%	70%
4	115	162	71%	46%
6	84	109	77%	24%
7	13	17	76%	92%
8	24	28	86%	113%
Total	446	600	74%	56%

10.3 Results

10.3.1 IS Management System Implementation

10.3.1.1 Qualitative Results

The study organisation is a provider of critical infrastructure in Norway. As a provider of critical infrastructure, it is obliged to follow the requirements of the Safety Act (“Sikkerhetsloven”) when it comes to preventive safety work, which includes safety analyses, securing objects, IS and safety drill.

The study organisation decided to map and analyse its ISC in 2014, due to its legal obligations, work activities and engagement. The organisation used the measurement of ISC in 2014 as an indicator of the IS level in the organisation and as a basis for identification of critical areas (related to attitudes, knowledge, practices) in need of improvement. Based on this, future goals for improvement were established, both at a general level and at a more specific level. A score of 87 or higher on the ISC index was established as a general goal for all departments. The 2014 survey identified needs when it comes to, e.g., increasing IS knowledge, attitudes and engagement among the personnel.

Several measures were taken to improve IS. Department managers were given the task of presenting the results from the 2014 ISC survey to their employees, discuss the results and measures that could be implemented to improve the status on the specific challenges discussed. A number of actions were taken: digital and on-site intrusion test, strengthened physical access control in the facilities, access card pin codes, new password policy, VPN dashboard and stronger fire wall policy (including strengthened fire wall and two-factor authentication), new routines for use of access card and visitor registration, internal training on password handling, handling of physical documents in field operations, photography and social media, office and document access in-house and information on possible consequences of negligent information leaks relating to security critical objects.

Second, the organisation started to establish a basic safety organisation in 2014, describing roles and responsibilities, as well as principles and guidelines related to IS. This was developed in accordance with the ISO 27001 principles. The implemented ISMS is particularly linked to the administrative department's responsibilities, the role of security coordinators and systematic risk analyses on all processes and goals relating to IS. A security coordinator was appointed, with a special responsibility to provide systematic training of the personnel in IS issues, to coordinate and train security coordinators in each department and to further develop IS protocols and instructions. The organisation reviewed all information related to security critical objects, crisis plans, security certifications and critical ICT systems, defining information into three categories: (a) open, (b) internal, (c) sensitive and graded. Additionally, new policies were developed for personal information, graded information, acceptable use of information and security policy for mobile units. The organisation developed a security declaration for employees to sign, documenting that they had received all the required training and information. A new system for recording and dealing with non-conformities was developed. The organisation also started to arrange an annual security month, and information security became a mandatory theme in each manager meeting. In spite of all these measures, interviewees agreed that challenges remained and that there were needs for improvement and maturing of the ISMS.

10.3.1.2 Quantitative Results

We asked the respondents three questions about the follow-up of the ISC measurement in 2014. We introduced the questions with the text: "We want to know a little about what actions your immediate supervisor has taken in your department/section after the evaluation of the ISC in 2014".

> The head of my department/section has gone through the results of the evaluation with us
>
> My department/section has taken steps to improve the ISC (e.g., focus on passwords and security-critical information) based on the results of the evaluation
>
> I am satisfied with how my department/section has worked on IS over the past two years

We also asked the respondents three questions related to whether they have received useful information over the past two years that has increased their knowledge and awareness of IS. These questions measure training in IS:

Table 10.2 Mean scores on the indexes (min:3, max:15) measuring follow-up of the 2014 ISC survey and training in IS, per department

Department	Follow-up index	Training index
1	12.4	12.9
2	10.9	12.4
3	11.0	12.3
4	11.4	12.5
5	11.9	13.1
6	13.8	14.3
Total	11.7	12.7

> During the past two years I have received useful information (e.g., from security coordinator, manager, intranet) about what a secure password is
>
> During the past two years, I have received information (e.g., from security coordinator, manager, intranet) that made me more aware of strangers in our premises
>
> During the past two years, I have received information (e.g., from the security coordinator, manager, intranet) that has given me more insight into what security-critical information is

We made two indexes based on these six questions (cf. Table 10.2). The first on follow-up, the second on training. All the questions include six response options: 1 = totally disagree and 5 = totally agree and 6 = have no knowledge about this. When we created the indexes, we removed the sixth response option.

Table 10.2 indicates the highest levels of follow-up in department 6, 1 and 5, and the highest levels of training in department 6, 5 and 1.

10.3.2 Improvements in ISC

We have combined the 24 statements with five response options on the five different aspects of IS in an ISC index. The indexes for the departments correspond to the average scores for the respondents. The minimum score is 24 (24 × 1), and the maximum score is 120 (24 × 5) (Table 10.3). Cronbach's Alpha for the 24 questions in the index was 0.913 in 2014, which means very good agreement between the questions and that the index is very good.

We see that the Department 6 (again) had the highest score in 2016 and 2014, followed by Department 2 and Department 1. All the departments saw an improvement on the ISC index in 2016, especially Department 1, which increased by 12 points on the index. The average improvement for all the departments from 2014 to 2016 was 9 points. This change is statistically significant at the 1% level. The

Table 10.3 Department scores on the ISC index (min: 24, max 120 points) in 2014 and 2016

Department	2014	2016
1	77	89
2	82	89
3	80	82
4	75	84
5	76	87
6	87	95
Total	78	87

differences between the departments are significant at the 1% level, both in 2014 and 2016.

GAIN [10] defines different types of culture, based on the scores of the index. The limits for "positive culture" range from 88 to 120 points on the GAIN index. The moderate culture scale goes from 47 points to a maximum of 87 points, and scores below 46 points correspond to a poor culture. If we are to transfer the GAIN scale values from safety culture to ISC, we see that none of the departments had a positive ISC in 2014. However, in 2016 we find that Department 6, Department 2 and Department 1 were within the part of the scale that we refer to as a positive culture.

10.3.2.1 Which Factors Influence ISC?

In Table 10.4, we examine the variables influencing respondents' ISC. We include variables measuring ISMS implementation and background variables.

Table 10.4 indicates two main results. The first is that the follow-up index is the strongest predictor of respondents' ISC. We see that the department variable ceases

Table 10.4 Linear regression of factors influencing respondents' scores on the ISC index in 2016. Standardized beta coeffisients.

Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Age	0.100***	0.181**	0.125	0.111	0.032	0.034
Education (University = 2)		−0.105	−0.055	−0.056	−0.013	−0.052
Seniority			0.136	0.109	0.067	0.058
Department (Dep. 6 = 2)				0.226***	0.035	0.005
Follow-up index					0.754***	0.645***
Training index						0.214***
Adjusted R^2	0.034	0.040	0.047	0.093	0.611	0.641

Dependent variable: ISC standardised beta coefficients
*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

to contribute significantly from Model 4 to Model 5, when the follow up variable is included. This indicates that this variable is related to follow-up, i.e., that this department has a higher score on the ISC index, due to the follow-up of the 2014 survey of ISC. The second main result is that the training index also contributes significantly and positively, indicating that IS training is related to a higher score on the ISC index. The Adjusted R^2 value in Model 6 is 0.641, indicating that the model explains 64% of the variation in the dependent variable.

10.3.3 Improvements in Information Security Behaviour

Table 10.5 shows results on the question: "When I am asked for information, I always think carefully about whether the information can be used for purposes other than its intended purpose", in 2014 and 2016.

Table 10.4 indicates the highest level of improvement in Departments 6 and 5, followed by Departments 4 and 1.

10.3.3.1 Which Factors Influence Information Security Behaviours?

In Table 10.6, we examine the variables influencing respondents' ISB. We include variables measuring ISC, IS knowledge and background variables.

Table 10.6 indicates two main results. The first is that the ISC index is a strong and significant predictor of the ISB of the respondents. We see that the department variable ceases to contribute significantly from Model 4 to Model 5, when the ISC index is included. This indicates that this variable is related to ISC (i.e., that the ISC score is higher in this department). The second main result is that respondents' IS knowledge also is a strong and significant predictor of their ISB. Knowledge is measured as the degree of agreement with the statement: "I am well aware of which kind of information that is sensitive and security graded". The Adjusted R^2 value in Model 6 is 0.207, indicating that the model explains 21% of the variation in the dependent variable.

Table 10.5 ISB scores in 2014 and 2016

Department	2014	2016
1	3.8	4.2
2	3.9	4.2
3	3.9	4.0
4	3.8	4.1
5	3.8	4.2
6	3.8	4.4
Total	3.8	4.1

Table 10.6 Linear regression of factors influencing respondents' ISB in 2016. Standardized beta coeffisients.

Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Age	0.045	0.024	0.016	0.009	−0.006	−0.008
Education (University = 2)		−0.132*	−0.126*	−0.118*	−0.084	−0.058
Seniority			0.017	0.017	−0.035	−0.044
Department (Dep. 6 = 2)				0.104*	0.025	0.008
ISC index					0.358***	0.233***
Knowledge						0.314***
Adjusted R^2	−0.001	0.012	0.009	0.016	0.129	0.207

Dependent variable: ISB standardised beta coefficients
*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

10.4 Discussion

10.4.1 The Implementation of ISMS

The first aim of the study was to describe the efforts to improve information security culture through the implementation of an ISMS. The study organisation has implemented several efforts to manage IS following the first measurement of ISC in 2014. In accordance with the continuous improvement approach inherent in ISMS, the organisation used the 2014 measurement as a baseline for future improvement, treating ISC as a broad indicator of IS in the organisation. Improvement goals were agreed upon, which were going to be followed up in a new measurement. In the meantime, several organisational measures were implemented and/or improved, e.g., related to the security coordinator, training, risk analyses and procedures. These were developed in accordance with ISO 27001 principles. As a consequence, interviewees in 2016 agreed that they had established a basic "security organisation" through the implemented measures.

10.4.2 How Can We Explain the Improvements in Information Security Culture?

The second aim of the study was to compare information security culture in the organisation before (2014) and after (2016) efforts to improve organisational information security culture, and examine factors influencing ISC. Our analyses indicate a 12% increase in the score for ISC in the organisation, which is statistically significant at the

1% level. Multivariate analyses indicate that variables measuring ISMS implementation were the most important predictors of ISC. This is in accordance with Hypothesis 1, stating that implementation of ISMS measures will lead to improvements in ISC. The follow-up index, measuring the follow-up of the 2014 measurement of ISC in each department was the strongest predictor of the respondents' ISC. This indicates the importance of such group-wise processes of continuous improvement, when it comes to developing ISC. The rates of improvement in ISC also varied between the different departments; ranging from 2.5% to 15% improvement. In line with Hypothesis 2, results indicate that the departments with the best implementation had the best improvements in ISC. This especially applies to department 6. According to Chen et al. [2], there is little research on the relationship between comprehensive IS programs and ISC. In this study, we contribute to this knowledge gap by studying how the ISMS measures of the study organisation have contributed to improvements in ISC and subsequently ISB.

10.4.3 How Can We Explain the Improvements in Information Security Behaviours?

The third aim of the study was to compare information security behaviour in the organisation before (2014) and after (2016) efforts to improve organisational information security culture, and examine factors influencing ISB. Our analyses indicate an 8% increase in the average score in the examined ISB from 2014 to 2016, which is statistically significant at the 1% level. These changes were also significant when controlled for factors like the age and education of the respondents. Multivariate analyses indicate that the ISC index and IS knowledge were the most important predictors of ISB. This result is in line with Hypothesis 3, stating that improvements in ISC will lead to improvement in ISB. This is in accordance with previous research indicating a relationship between ISC and ISB [1–3]. The rates of improvements in behaviour varied among the different departments in the organisation, ranging from 3% improvement in Department 3 to 16% in Department 6, again indicating the importance of effective ISMS implementation for ISC and ISB results. The ISC index measures questions related to management and employee focus on IS, IS training, etc. Our results indicate that this is positively related to our measure of ISB, which is related to confidentiality: "always thinking carefully about whether requested information can be used for purposes other than its intended purpose".

10.4.4 Safety Culture Versus Security Culture

We used a modified organisational safety culture scale [10] to measure ISB. The scale was chosen as the research on organisational safety culture seems to have been through many of the challenges that the organisational security culture research now is facing (cf. Ruighaver et al. [7]). At the same time, the research on organisational safety culture seems to have matured a bit more conceptually and methodologically, as it has employed the culture perspective for a few more years than the field of security research. We therefore draw on the experiences of the safety culture literature.

There are however several important differences between safety culture and security culture and the applications of the concepts. First, the difference between sharp end and blunt end is harder to see in the field of security. All members of the organisation are in one sense in the security sharp end, as they are users of information, equipment and facilities. Additionally, the results of serious security incidents and non-conformities may remain unseen and unnoticed for a long time. The same applies to latent system failures, which may be exploited undetected for long periods by third parties. Security incidents are generally not physical accidents with immediate damages or injuries. A consequence of this is that it may be even more difficult to define a state of "security" (e.g., as the absence of serious security incidents) than it is to define a state of "safety" (e.g., as the absence of physical accidents).[1] Thus, security is a more abstract state than safety, making security assessments and preventive efforts more challenging.

It could also be mentioned that in the field of security, the distinction between the private domain and work domain is blurred, as employees use digital workplace equipment (e.g., phones, tablets, computers) in their leisure time. Thus, they must also act according to their ISMS and ISC after working hours and in their private spheres. Thus, in contrast to the field of safety, the field of security requires employees to be always "at work". This has interesting implications for ISC management: it stretches into both the professional and the private sphere.

Another difference is related to intent: safety culture mainly concerns prevention of incidents related to combinations of technological, (***unintentional***) human and organisational risk factors, while security culture concerns prevention of incidents related to combinations of technological, (***intentional***) human and organisational risk factors. As the human component in the security field often deals with intentional actors with hostile intentions, the possibilities for failure (security breaches) are greater. In the case of safety, human risk factors are typically related to unintended errors, mistakes and violations, combined with technological and organisational weaknesses. In the case of security, these risk factors at the "victim end" are combined with the creativity, expertise and imagination of intentional human actors at the "offender end". An implication of this is that security management also is related to crime prevention.

[1] This may also be the case with latent organisational or technological safety failures, which may exist undetected until they interact with active failures in ways that lead to accidents. Given the abstract character of security, preventing unwanted security incidents may have more to learn from research on safety culture in complex technological systems, where technological failures may act in unseen, unanticipated and incomprehensible manners due to "interactive complexity" (cf. Perrow [15]). Previous research indicates that cultural management is particularly important in these settings (cf. Weick et al. [16]).

Finally, the most important similarity between safety and security management is that physical and organisational measures are important for increasing both safety and security, but as long as human actors use these systems and relate to the physical measures, the state of safety and security is contingent on the behaviours and subsequently the culture of these actors. This indicates the crucial importance of ISC for IS and the importance of organisational safety culture for safety.

10.5 Conclusion

The study finds that the organisation providing critical infrastructure has managed to improve its ISC through systematic ISMS efforts over a two-year period (2014–2016), and that this also has led to improvements in ISB among the personnel. Multivariate analyses indicate that ISC is the most important variable influencing ISB. Respondents' knowledge about IS was also an important variable in the analyses. Thus, our results indicate that it is important to work with organisational ISC and knowledge to increase IS in our increasingly digitalised society, especially in organisations providing critical infrastructure.

Acknowledgements we are thankful to the contact persons, our respondents and interviewees in the study organisation. Results of the study have also been presented in Nævestad et al. [11, 12].

Ethics Statement The methods for data collection in the present project are in accordance with the ethics policies and requirements of the Institute of Transport Economics and the Norwegian Centre for Research Data (NSD). NSD assists researchers with research ethics of data gathering, data analysis and issues of methodology. The study was reported to NSD. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements. Oral informed consent was obtained from participants, and all data have been anonymised.

References

1. J.S. Lim, S. Chang, S. Maynard, A. Ahmad, Exploring the relationship between organizational culture and information security culture, in *Proceedings of the 7th Australian Information Security Management Conference*, 1–3 December 2017 (Perth, Western Australia, 2017)
2. Y.K. Chen, K. Ramamurthy, K.-W. Wen, Impacts of comprehensive information security programs on information security culture. J. Comp. Inform. Syst. **55**(3), 11–19 (2015)
3. A.R. Nasir, A. Arshah, M.R.A. Hamid, S. Fahmy, An analysis on the dimensions of information security culture concept: a review. J. Inform. Secur. Appl. **44**(2019), 12–22 (2019)
4. M. Mirtsch, J. Pohlisch, K. Blind, *International Diffusion of the Information Security Management System Standard ISO/IEC 27001: Exploring the Role of Culture.* Research Papers, p. 88 (2020)
5. T.-O.I.S. Nævestad, K. Hesjevoll, S.A. Ranestad, Strategies regulatory authorities can use to influence safety culture in organizations: Lessons based on experiences from three sectors. Saf. Sci. **118**(2019), 409–423 (2019)

6. S. Antonsen, The relationship between culture and safety on offshore supply vessels. Saf. Sci. **47**(8), 1118–1128 (2009)
7. A.B. Ruighaver, S.B. Maynard, S. Chang, Organisational security culture: Extending the end-user perspective. Comp Secur. **26**, 56–62 (2007)
8. P. Chia, S. Maynard, A.B. Ruighaver, in Understanding organizational security culture, in *Sixth Pacific Asia Conference on Information Systems*, 2–3 September 2002 (Tokyo, Japan, 2002)
9. T.-O. Nævestad, Cultures, crises and campaigns: Examining the role of safety culture in the management of hazards in a high risk industry. Ph.D. dissertation, Centre for Technology, Innovation and Culture, Faculty of Social Sciences, University of Oslo (2010)
10. GAIN (Global Aviation Network), in *Operator's Flight Safety Handbook* (2001). http://flightsafety.org/files/OFSH_english.pdf
11. T.O. Nævestad, J. Hovland Honerud, S. Frislid Meyer, Organizational information security culture in critical infrastructure: developing and testing a scale and its relationships to other measures of information security, in *Safety and Reliability—Safe Societies in a Changing World, Proceedings of ESREL 2018,* June 17–21, 2018, Trondheim, Norway, ed. by S. Haugen, A. Barros, C. van Gulijk, T. Kongsvik, J.E. Vinnem (CRC Press, London, 2018a)
12. T.-O. Nævestad, J. Hovland Honerud, S. Frislid Meyer, How can we explain improvements in organizational information security culture in an organization providing critical infrastructure?, in *Safety and Reliability—Safe Societies in a Changing World, Proceedings of ESREL 2018,* June 17–21, 2018, Trondheim, Norway, ed. by S. Haugen, A. Barros, C. van Gulijk, T. Kongsvik, J.E. Vinnem (CRC Press, London, 2018b)
13. K.J. Knapp, T.E. Marshall, R.K. Rainer, F.N. Ford, Information security management's effect on culture and policy. Inf. Manag. Comput. Secur. **14**(1), 24–36 (2006)
14. E.H. Schein, *Organizational Culture and Leadership*, 3rd edn. (Jossey-Bass, San Francisco, 2004)
15. C. Perrow, *Normal Accidents: Living with High-Risk Technologies* (Basic Books, New York, 1984)
16. K.E. Weick, Organizational culture as a source of high reliability. Calif. Manage. Rev. **24**(2), 112–127 (1987)

Chapter 11
AI at Work, Working with AI. First Lessons from Real Use Cases

Yann Ferguson

Abstract This chapter deals with the transformations of employment and work associated with recent developments in artificial intelligence. It proposes a classification based on five figures of the worker: replaced, dominated, augmented, divided and rehumanised. This taxonomy is illustrated by use cases from the catalogue of the Global Partnership on AI, a multistakeholder initiative which aims to bridge the gap between theory and practice on AI. We conclude by highlighting three shifts in the forms of work engagement that are likely to impact safety issues: the distancing of the object of work, the work on the machine itself and the reconfiguration of professional identity.

Keywords AI applications · Employment · Skill management · Professional identities

11.1 Artificial Intelligence at Work: Five Workers Stories

Over the past decade, the impact of AI on the future of work has been the subject of much research by academics, governments, experts, non-governmental organisations, professional federations, international organisations, philosophers, essayists and others. It is not the intention here to list them all. However, from a worker's point of view, we can organise the anticipated effects of AI into five categories [13].

The replaced worker: AI systems will massively replace workers and destroy jobs. Several studies or essays tend to show that many jobs will disappear (more than 40%), with the machine performing tasks more efficiently and at lower cost than humans [6, 7, 12, 14]. Adopting a "job-based approach", they estimate that many occupations are at high risk of automation. Other studies prefer a "task-based approach" [5] focusing

The author takes full responsibility for the statements made in this article.

Y. Ferguson (✉)
ICAM, Toulouse, France
e-mail: yann.ferguson@icam.fr

J.-C. Le Coze and S. Antonsen (eds.), *Safety in the Digital Age*,
SpringerBriefs in Safety Management,
https://doi.org/10.1007/978-3-031-32633-2_11

on the complementarities between automation and labour.[1] From this perspective, AI will destroy few jobs (around 10% depending on the country) but will transform many occupations (around 50%).

The dominated worker: AI systems will dominate workers by reducing their empowerment. Beyond the "technological singularity" hypothesis, many studies are concerned about the effects of AI on workers' autonomy, due to the development of an "algocracy". But the dominated worker does not only result from active forms of domination. It can also result from the worker's passivity in the way they interact with the system: overconfidence, the contentment effect (being satisfied with a relatively satisfactory solution obtained without effort) and overcautiousness can disengage the worker, reduce their expertise and consequently increase their dependence on the system.

The augmented worker: Workers' empowerment is strengthened by AI. Combined with AI, the enhanced human being reaches a level of performance normally unattainable, thanks to a good partnership between man and machine, with man bringing his true added value [18, 19].

The divided worker: "winner-takes-all-economy", the polarisation of labour. Many studies suggest that AI may polarise the labour market. On the one hand, an "aristocracy of intelligence" with a high level of complementarity with artificial intelligence occupy highly qualified and stimulating jobs. On the other hand, workers in low-skilled jobs have precarious and uninteresting work [1, 2, 4, 8, 16, 17] (Graham and Woodcock 2019).

The rehumanised worker: workers focus on properly human skills. The automation of tasks and trades could be an opportunity for the "de-automation" of human work. It would allow the development of human capacities: creativity, manual dexterity, abstract thinking, problem solving, adaptability, emotional intelligence [10, 19].

However, in recent years, several companies have started to integrate AI into their organisations and professions, thus moving away from speculative approaches.

11.2 From Stories to Real Cases: What Working with AI Could Mean

Building a Collection of Real Use-Cases of AI Systems in the Workplace

The Global Partnership on Artificial Intelligence (GPAI) has decided to launch the creation of a global catalogue of real-world AI use cases at work.[2] By exploring the

[1] According to these authors, substitutable tasks are routine tasks, both manual and cognitive, meaning that there is a limited number of tasks that can be defined with the explicit rules of a program. Conversely, for non-routine, more complex tasks, computer capital is more complementary than substitutable for the worker.

[2] The GPAI is a multistakeholder initiative which aims to bridge the gap between theory and practice on AI by supporting cutting-edge research and applied activities on AI-related priorities. Built around a shared commitment to the OECD Recommendation on Artificial Intelligence, GPAI brings

state-of-the-art and capabilities of AI in the workplace, the Future of Work Working Group seeks to provide critical technical analysis that will contribute to the collective understanding of

- how AI can be used in the workplace to empower workers and increase productivity, how workers and employers can prepare for the future of work;
- how job quality, inclusiveness and health and safety can be preserved.

Since September 2020, we have started to collect stories from different actors in AI (providers, CEO, managers and end(users) who are involved in its implementation at work and organisations, in order to:

- better understand the motivations of those who integrate AI into organisations and work;
- better understand how AI is deployed in the field;
- highlight the issues and social effects of AI integration;
- highlight the convergences and divergences in the feedback according to the nature of the respondents;
- highlight "good practices" from the field that could outline a method for implementing AI.

The answers to the questions refer to a specific professional application of AI (and not to an AI system in general). Indeed, many questions relate to uses, organisational and social contexts or design methods. After the first year of research, the catalogue consists of 150 use cases, spread over 12 countries.[3]

11.2.1 Five Workers Stories Put to Test of Real World

The replaced worker: In almost all the cases studied, AI systems are not intended to automate an entire process or task. Rather, the goal is to improve the performance of the human worker. In this sense, respondents emphasise the notion of a "decision-making tool" and that the final decision is always human. "*There is the human-in-the-loop and human makes the final decision. The AI alerts and recommends only*" (Private—SME—FinTech—Machine Learning, NLP—Automation of the surveillance). The reasons are not ethical but are related to "probabilistic" or "empirical" AI. AI systems built on this type of algorithm provide only probabilities based on limited knowledge of the environment and contexts. Humans can provide this information and correct possible errors in order to make the decision. Therefore,

together engaged minds and expertise from science, industry, civil society, governments, international organizations and academia to foster international cooperation. With their research and practical projects across various sectors and disciplines, the four working groups and an ad hoc subgroup are initially focusing on five themes: Responsible AI; Data governance; The future of work; Innovation and commercialization; AI and pandemic response.

[3] The full report is available here: https://gpai.ai/projects/future-of-work/ai-at-work-observation-platform/ai-observatory-at-the-workplace.pdf.

the value generated by AI systems in organisations would not come from increasing the organisation's control over its human resources by automating work. It would not come from increasing the power of the organisation through machines or processes. The value created by AI systems would come from trust in human work. "*There are two approaches to AI: one which values the worker, one where he is excluded, because he is fragile and limited. Either we build trust, or we build control. This gives a moral compass on a path that can be paved with rupture. Putting the human at the centre is an incantatory discourse…it must not be said, it must be done. University engineering departments must open up more to the humanities, question their political responsibility*" (Non-profit—SME—Start-up on Augmented intelligence—Computer vision with standard and specific methodologies—Increasing the speed of fault analysis). According to the respondents, current AI is nothing more and nothing less than a human decision support tool. In this sense, they believe that AI's capacities are highly overestimated, which simultaneously generates irrational fears and hopes and, sometimes, frustration. "*It is more about having a machine plus human system, it is better both for efficiency and acceptability. In the end you will always keep a human in the process, mainly because of the high amount of spending decision in case there is a problem. You have someone to complain to if anything goes wrong. The difference is that AI makes different mistakes than humans, and sometimes they also seem stupid ones. There are some people that imagine a magic wand and they have impossible expectations but in the end the experts' knowledge is needed, and everything is aligned*" (Private—SME—Energetic—Defect detection, failure detection—AI for image processing and defect detection on industrial structures. Qualification of defects on wind turbines). Almost ten years after Frey and Osborne's first prediction, it is hard to say that AI has caused a massive wave of task automation.

The dominated worker: Few use cases are mature enough to observe algocracy situations. However, three tendencies emerge:

- Algorithmic management situations in warehouses. Voice order solutions called "*voice picking*" totally control the picker. They receive their instructions via a headset, and dialogue directly with the information system through a microphone and voice recognition software. The voice software solution interfaces with the warehouse management system or directly with the sales management system. The promoters of this solution praise the productivity gains, the reduction of the error rate, the reliability of the organisation, the elimination of the hand-held order support: the user works hands and eyes free. "*Negative results predominate because qualification and work experience are no longer necessary because the AI takes over,* explains a German trade unionist. *Only a short period of training is required to use the system. With regard to the quality of the work, there was a simplification (the system speaks all languages, knows all calculations, processes and products); the AI thus led to a de-qualification of the workers because no previous knowledge was required. There is no further development of the workers because it is not necessary and also impossible, e.g., the workers unlearn calculating and product knowledge. The result can be described as the dumbing down of* *the workers. They act like machines*" (Private—SME—Food Industry—Perception/audio processing—Increase pickers' productivity in a warehouse with a pick by voice-system).

- A growing influence of processes on practices: The more prescribed the work, the better the applications perform. "*The information that our tool delivers is that which the methods have previously structured. The more our clients have filled in their processes, the more complete our system is. When they don't have these series of instructions, the first step is to help them formalize them"* (Private—SME—Software development—Natural Language Processing—Optimising machine usage through speech or text interaction).
- Governance by numbers [24]: When AI systems communicate through numbers, indicators, probabilities, workers have difficulties to position themselves. Some people trust them absolutely. The presumed efficiency of artificial intelligence is in this sense the most recent formalisation of this old dream of "harmony through calculation" where mathematics is the key to the intelligibility of the world. In other situations, workers would like to intervene, but they do not know how: "*what was different compared to other tools is that we know more or less how it works, whereas here, there was a real lack of clarity about the results, how the tool obtained a result*" (Public—Big firm—Public Administration—Machine learning—Identifying errors).

The augmented worker: In the absence of total automation of a task, current AI systems are more like decision-support systems. Total automation by AI is blocked by the impossibility to guarantee the performance levels of a system. AI is not certifiable, an imperative condition for the automation of a critical industrial process. We can distinguish four forms of augmentation of the worker by AI:

- Augmentation-remediation: AI allows the worker to do what he doesn't know how to do. "*This AI application addresses a cybersecurity problem: too many documents produced by different departments. Humans are no longer able to tag them. This is a data governance problem. Because of the evolution of professional practices, one can be identified anywhere in the world with confidential data. We need to secure data outside its traditional security perimeter*" (Private—Big firm—AI development for cybersecurity—Web service—Securing data outside its traditional security perimeter). AI systems can also be used to compensate for the limitations of some workers: "*It stops doing this activity manually in Excel and to start recording the data with a voice recognition system. Saving time, reduction of errors, greater control are three examples of benefits. We can tell if we are within the tolerance levels, within the standards. If not, the operator has to say why. The academic level of the workers is low with poor literacy. The use of voice is empowering*" (Private—SME—Software—NLP—Facilitate quality).
- Augmentation-rationalisation: AI allows workers of different skill levels to reach a more homogeneous result. AI system "*aims to accumulate knowledge so a young worker or newly assigned technician can handle work that requires the knowledge of a highly skilled technician (highly skilled, can operate specialized equipment,* etc.*) This AI application supports the quality of work equivalent to that of highly*

skilled workers by incorporating the tacit knowledge of highly skilled workers into AI" (Private—Big firm—General construction—Deep learning—Raising the skills of young workers).

- Augmentation-delegation: AI relieves the worker from low value-added tasks and refocuses on high value-added tasks. "*The AI system is very good at detecting welding anomalies, but much less good at qualifying these anomalies. The value of the worker has shifted from detecting problems to qualifying problems*" (Private—Big Firm—Computer Vision—Object Recognition).
- Augmentation-cooperation: The association of the worker and the AI produces a new performance. A human/AI association from which would emerge a worker equipped with new capacities, a "synthesis of the best of man and machine": "*The main idea is to look for the bottleneck in calculation programs, where computation times take longer, and replace that part of the code by a digital twin. There is a compromise to be made between precision and time saving. Some people want more speed than accuracy, and others the opposite. Sometimes it is better to know results in 1 h for instance instead of 3 weeks*" (Private—SME—Data sciences—Deep Learning—Digital software twins to increase the speed of calculations).

The divided worker: The systems can effectively feed a polarisation of work by generalising the expertise of the most competent and experienced workers. But the impact of AI systems on human expertise is heterogeneous.

- A shift in value that can reinforce the status of the business expert: AI systems, by shifting human work to high value-added tasks, strengthen the position of experts. "*It changes the organisation with automation of the handling phase and refocusing on the reading and statistical analysis of the results. This was possible because the operators were experienced*" (Non-profit—Big firm—Agri-food—Cobot, Visual recognition, Adaptive learning for movements—Cobot that increases worker productivity by refocusing them on high value-added tasks).
- An association "novice + AI system" that can weaken the status of the business expert. With AI systems, business experts become less indispensable in the long term, after the design and training of the AI system which becomes more autonomous. It strengthens managers' positions. "*We know that the people who do this have a very high added value, but that's not reassuring, they don't want the adjustment to rest on them. It's a critical operation that we do regularly. [...] The intelligence was in the machine, the managers wanted to be able to put anyone on the task. The person was the hand of the application*" (Private, Big Firm-Aircraft industry- Door positioning assistance system).

The rehumanised worker: The most striking "rehumanisation" effect is the automation of repetitive tasks considered as having little added value, such as answering emails. "*In this case, we had a huge social issue. The simple processing of e-mails represented 6–8 h of activity per day. Many wanted to change jobs. Our system now has 94% successful email routing. Now they can focus on their job, accounting analysis. They still answer emails for two hours a day. But these are dedicated, complex, interesting requests that require their expertise*" (Private, Big

Firm-Energy- chatbot for accountants). Cobots can also relieve workers of tasks that generate musculoskeletal disorders: "*Automation of repetitive tasks with high mental load, reduction of musculoskeletal disorders: the operators put the products to be tested in boxes, the cobot recovers, scans, checks in its database that the product "on hand" is the one to be tested. It opens the product and duplicates the protocol of an operator until the end of the preparation phase. The analysis phase remains human*".

Beyond the five stories, what emerges strongly from our survey is above all how, in their current phase of development, the professional applications of AI are profoundly shaped by work and workers. In the majority of cases studied, AI systems consist of generalising expertise. Thus, as was previously the case for expert systems, the actors of the profession are essential to design and improve AI systems. "*The companion has a very important role because we rely on him to educate AI; without his feedback, we are blind. I try to remain humble because I fundamentally believe in the intrinsic value of professions. I'm talking more about "increased intelligence" than AI, and that's what understanding is all about. You can't work without domain experts. That's why all the big AI companies are recruiting trade experts. Unsupervised learning bricks are specific and redundant, always start with an expert system approach*" (Non-profit—SME—Start-up on Augmented intelligence—Computer vision with standard and specific methodologies—Increasing the speed of fault analysis).

11.3 Discussion: AI, Organisation, Workers and Safety

Machine learning is the current main approach of what is called "empirical AI". This means that it does not produce deterministic or certifiable results like classical machines, but works on the basis of statistics from which it derives correlations. These correlations establish probabilities. In high-hazard organisations, these probabilities will have to coexist with a very normative culture. It will be necessary, for example, to estimate the value of a prediction in a structured environment. It is well known that workers do not always comply with prescriptions, and that procedural deviations are essential for the proper functioning of organisations. However, on the one hand, AI can reinforce the prescribed work (algocracy), on the other hand, its functioning remains empirical. Moreover, the balances that workers will be able to find in their interactions with machines will be unstable, as machines will keep improving. In the end, one of the challenges for safety will be to manage situations of paradoxical injunctions: the worker will be asked to trust an AI system while controlling it and assuming responsibility for the whole.

Considering workers, three shifts seem to be particularly structuring. These three shifts converge to reconfigure the forms of engagement in work:

- Shift 1: the object of work could be put at a distance. The worker will do less and supervise more programs and machines. What will be the impact of this distancing on workers' consideration of safety?

- Shift 2: These programs and machines will then become fully fledged objects of the activity. AI applications need to be trained, completed and corrected. Workers could become less expert in the situations they have to solve than in the machines that solve the situations. How can we organise this work on AI so that it optimises safety?
- Shift 3: these two shifts will impact the construction of the self at work, professional identity, and the recognition of singularity. What will be the place of safety in this identity reconfiguration?

From the point of view of safety, we need to understand how these new forms of subjective commitment will contribute positively or negatively. The example of the automation of aircraft piloting provides elements of an answer: it has generally made flights safer, but it has also increased human error, particularly by depriving pilots of the sensations and perception of flight [9]. In response, aircraft manufacturers are working on two completely different avenues: increasing automation and improving the relationship between humans and machines [22]. Security could be faced with the same kind of questions in the near future.

Ethics Statement Informed consent was obtained from all informants interviewed for this work, and their identity has been anonymised. Ethics board approval is not required for this type of study in France.

References

1. E. Anthes, The shape of work to come. *Nature* **550**, 316–319 (2017)
2. M. Arntz, T. Gregory, U. Zierahn, in The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis. OECD Social, Employment and Migration Working Papers, No. 189 (OECD Publishing, Paris, 2016)
3. D. Autor, F. Levy, J.-M. Urnane, The skill content of recent technological change: an empirical exploration. Quart. J. Econom. **118**(4), 1279–1333 (2003)
4. D. Autor, in The Polarization for Employment and Earnings. Center for American Progress (2010)
5. D. Autor, Why are there still so many jobs? The history and future of workplace automation. J. Econom. Perspect. **29**(3), 3–30 (2015)
6. C.B. Frey, M. Osborne, The future of employment: How susceptible are jobs to computerisation? Technol. Forecast. Soc. Change **114**(issue C), 254–280 (2017)
7. C.B. Frey, in *The Technology Trap: Capital, Labor and Power in the Age of Automation* (Princeton University Press, 2019)
8. E. Brynjolsson, A. McAfee, in *The Second Machine Age* (Norton & Company, 2014)
9. N. Carr, in *The Glass Cage: Automation and Us* (Norton & Company, New York, 2014)
10. B. Christian, in *The Most Human Human: What Artificial Intelligence Teaches Us About Being Alive* (Doubleday, 2011)
11. C. Collectif, Éthique de la recherche en apprentissage machine, in *Rapport de recherché* (Cerna; Allistene, 2017)
12. S.W. Elliot, in *Computers and the Future of Skill Demand* (OECD Publishing, Paris, 2017)
13. Y. Ferguson, Une intelligence artificielle au travail—Cinq histoires d'Hommes, in *L'intelligence artificielle dans toutes ses dimensions*, ed. by B. Barraud. (L'Harmattan, Paris, 2020), pp. 129–144

14. M. Ford, in *Rise of The Robots* (Perseus Books Group, USA, 2015)
15. S. Head, in *Mindless: Why Smarter Machines are Making Dumber Humans?* (Basic Books, New York, 2014)
16. I. Goldin, The second Renaissance. Nature **550**, 327–329 (2017)
17. N. Jaimovich, H. Siu, in *The Trend is the Cycle: Job Polarization and Jobless Recoveries.* NBER Working Paper, 18334 (2012)
18. F. Levy, R. Murnane, in *Dancing With Robots: Human Skills for Computerized Work Next* (2013)
19. Mission Villani, in *Donner un sens à l'Intelligence Artificielle* (2018)
20. OECD, in *Artificial Intelligence in Society* (2019)
21. OPECST, in *Pour une Intelligence Artificielle maîtrisée, utile et démystifiée* (2017)
22. V. Scardigli, in *Un anthropologue chez les automates* (PUF, Paris, 2001)
23. G. Simondo, Du mode d'existence des objets techniques (Aubier, Paris, 2012)
24. A. Supiot, La gouvernance par les nombres, Cours au Collège de France (2012–2014) (Fayard, Paris, 2015)

Chapter 12
Safety in the Digital Age—Sociotechnical Challenges

Stian Antonsen and Jean-Christophe Le Coze

Abstract This chapter describes some of the recurring themes that emerged from the contributions in this book, as well as from the workshop in which the contributions were presented and discussed. The themes are in one way or another related to the term "sociotechnical" and thus point to problems (old and new) that are linked to the relationship between the social and technological dimensions of organisations. The chapter provides a brief explanation of the history and current use of the term "sociotechnical" before discussing three sociotechnical issues that we believe are important for dealing with safety in the digital age.

Keywords AI · Technological transformation · Sociotechnical perspective · Risk migration · Governance

12.1 Introduction

The premise underlying this book was that many societies are currently undergoing a process of social and technological transformation. More specifically, we take as our point of departure that there is a significant increase in data harvested from both humans and technology, alongside increasing capabilities and ambitions to utilise these data in developing algorithms for the automation of work, machine learning and a new generation of AI [1, 2]. These developments also take place within high-risk industries such as healthcare (Sujan) and offshore drilling (Paltrinieri), in addition to cyber-space which is becoming a source of societal vulnerability and transnational concern (Backman).

S. Antonsen
NTNU Social Research, Trondheim, Norway
e-mail: stian.antonsen@samforsk.no

J.-C. Le Coze (✉)
Ineris, Verneuil-en-Halatte, France
e-mail: jean-christophe.lecoze@ineris.fr

J.-C. Le Coze and S. Antonsen (eds.), *Safety in the Digital Age*,
SpringerBriefs in Safety Management,
https://doi.org/10.1007/978-3-031-32633-2_12

The call for contributions described these technological trends and posed questions regarding their implications for work, organisations, businesses and regulation. In this respect, it is no surprise to find sociotechnical challenges as recurring themes in the chapters of this volume. Nevertheless, it is striking how the chapters touch upon similar issues regarding the way digital technology produces inscriptions for social life and vice versa, despite the chapters starting from very different perspectives, methods and cases of study. This provides weight to a claim that a strictly technology-centric view runs the risk of misrepresenting both the challenges and opportunities involved in introducing a wide variety of digital tools. The same goes for a strictly human-centric view, given that high-risk systems are usually high-technology systems with a high level of automation. Exploring safety issues in a digital age thus involves a sociotechnical perspective, implying that a sociotechnical lens has something important to offer. A remaining question is, however, what does it mean to adopt a sociotechnical perspective? Before discussing the recurring themes of the book, this question needs a brief consideration.

12.1.1 What is a "Sociotechnical Perspective"?

The literature on sociotechnical systems dates back to research on work design and organisation development at the Tavistock Institute in the 1950s, and subsequent action research projects in Britain, Norway, Australia and the USA over the following decades [3].

The core idea of the sociotechnical approach is that the technical and social systems of work organisations need to be seen in close relation, and hence, should not be designed and developed in separation. From a sociotechnical perspective, the effectiveness of work systems emerges from the match between the requirements of the social and technological systems, often described as consisting of four broad classes of variables: structure, people, technology and tasks [4]. There will not be "one best way" to design a sociotechnical system, but specific analyses need to be undertaken to find ways to organise activities that are tailored to the properties of the technology involved, while also addressing workers' needs for, e.g., autonomy, task variation or interpersonal interaction.

While the specific methods and techniques of sociotechnical improvement are not widely used today, their spirit, the expression itself and the general logic of a sociotechnical approach are very much present in fields like ergonomics, human–machine interface design and cognitive system engineering. What constitutes a sociotechnical perspective can also be argued to have changed over the years. In its origins, it referred to the alignment or joint optimisation of a social and technical system. For instance, within the literature on sociomateriality, it is more common to treat the relationship between the social and material (including technology) as a matter of *entanglement*, thereby pointing to a need to understand the way one involves inscriptions in the other. Similar arguments have been made with reference to high-risk systems, e.g., Le Coze [5] referring to the term "sociotechnical" as an idea that it is virtually *impossible* to distinguish the technological from the “non”-technological when it comes to understanding high-risk systems. Approaching one without a relationship to the other is likely to be misleading if the objective is to analyse risk, whether it is through proactive risk assessments or accident investigations (ibid.).

Haavik [6] provides further insight into this perspective on sociotechnical analysis. He argued that the classical organisational perspectives on safety tend to “*treat sociotechnical systems as complex systems made up of factors belonging to the well-defined realms of humans, technologies and organizations*” [6]. Inspired by Latour [7, 8], Haavik presents empirical case studies illustrating two arguments: 1) the system components (e.g., technical systems) can gain their properties from their relations to other components, and 2) technical systems are boundless in the sense that they are not easily demarcated, neither from social components nor from other technical systems. In this perspective, it is more relevant to explain sociotechnical systems as relationships between heterogenous actors, and that “*the properties of the actors are results of the relations, not* vice versa” [6]. In this perspective, assessing “the social” and “the technical” components of safety will be, at best, half of the process of a sociotechnical analysis. The key to understanding the system dynamics that are involved in the production of unwanted outcomes (how the system “works” in different situational contexts) lies in understanding how system components may influence and shape each other.

The chapters in this volume indicate that a sociotechnical perspective on safety is probably more relevant than ever. They also illustrate that a sociotechnical view should not be restricted to the initial scope described by its pioneers like Trist [3]. It should encompass a wider spectrum of empirical scrutiny and theoretical reach as promoted, first, by broad, or multilevel analysis of situations and second, by a greater emphasis on the digitally mediated practices considering their pervasiveness across activities in safety–critical systems. These two options consist in respectively, considering a wider range of actors (and institutions) than the micro- or mesoframing of Trist allows, and second, granting technology a higher level of agency and power in shaping social realities than so far introduced. The pace and pervasiveness of technological innovation is not only increasing, but has developed far beyond the question of alignment between a social and a technological subsystem in organisations. Digital technology not only mediates passive representations of reality, but also takes on *roles* in production processes and work environments by performing tasks, distributing work, making interpretations of current situations, predictions of future situations and providing both advice and decision making. As such, digital technology *actively* shapes human perception and activity, and its integration into human activities goes beyond the traditional conception of a mere tool. With this perspective on sociotechnical systems, we now turn to examples of more specific research challenges related to safety in a digital age.

12.2 Sociotechnical Challenges

12.2.1 Where is "The System"? The Migration of Risk

Several of the chapters indicate that digitalisation involves changes in the type of actors that provide input that is in one way or another critical to the real-time reliability of systems: for making decisions and adjustments in normal operations, detecting anomalies and weak signals of danger, dealing with disturbances and crises, and for restoration of system operations after failure. Importantly, features that are added to make a system work in new ways, also mean that it can fail in new ways, involving new actors in both successes and failures. The following are examples drawn from the chapters in this volume:

- If there is growth in modelling and simulation science as a form of generic meta-science, then the properties of input data and the assumptions of model-makers and analysts become more critical, as illustrated by Demortain.
- If software becomes more critical, then the practices of software developers become critical, including their navigation in the four interrelated trade-offs described by Roe and Fortmann-Roe.
- If information security can be compromised by the actions of administrative support staff, then the information security culture and behaviours of this staff will also matter for the overall integrity of the system (Nævestad et al.).
- If wearable technology and other IoT devices do in fact gather personal data, the recipients and processors of such data become potential actors in, e.g., the organisation of safety–critical work and accident investigations (Caron; Guillaume).
- If digitalisation does indeed introduce new ways of failing through tighter couplings and increased complexity, this can give rise to a new species of crises, as argued by Backmann.

These examples point to a fundamental question for safety research: Where do we draw the lines around "the system" we study when we aim to describe, analyse and ultimately improve conditions for safety? A wide variety of extra- and intraorganisational actors, e.g., software engineers, computer scientists, model makers, HR staff, all seem to be part of the sociotechnical challenge, but do we really account for them as part of the high-risk system? When posing such questions, complexity becomes not only a word, referring to the number of system components and the interaction between them, but a multilevel phenomenon in need of interpretation. Moreover, it requires understanding system relations, in addition to the properties of each system component. For instance, the information security culture and behaviour of administrative staff is not safety–critical in itself—its criticality depends on its relations to other sociotechnical elements that can be more directly related to an unwanted outcome.

One way of approaching such questions is by framing them as a matter of migration of risk in systems that can be both polycentric and "borderless" [9]. While outsourcing relationships have been around for decades, digital value chains have a potential for becoming so long and involving such a heterogenous network of actors providing critical input that it becomes virtually impossible to draw the line between the inside and outside of the systems at risk [10]. Assessing and managing risk in such a landscape involves viewing a sociotechnical system not as a clearly defined and static entity, but rather as a changing network of human and technological actors.

In such a conceptualisation, it becomes increasingly hard to maintain the traditional division between the "sharp" and "blunt" end of industrial organisations. In a digitalised sociotechnical system, professional communities can both monitor and operate technical systems without being in the vicinity of the physical production processes. While this is by no means a new problem for safety research, knowledge of the operational context in which, e.g., software is entangled, becomes a matter of increased importance. Moreover, reconsidering the division between the sharp and blunt ends of organisations may imply reconsidering the system's control strategies. For instance, it might involve a form of drift along the centralisation/decentralisation axis that is key to both Normal Accident Theory and HRO research. Digitalisation can make a system more decentralised by bringing in new roles taking on responsibilities as "reliability professionals" involved in maintaining system states and recognising and interpreting anomalies [11].

Addressing the issues described here will require a level of granularity in the analysis that enables both the identification and understanding of new and changing relations that are enabled by digitalisation. One research challenge for safety research in the digital age is thus one of "moving closer" to sociotechnical relationships in order to assess the specifics of such relationships.

12.2.2 The Relations of Rationalities

The chapters from Caron and Guillaume illustrate the classic issues of friction between technical administrative rationalisation processes, and the need for professional autonomy of both individual employees and a workers' collective as a whole (e.g., [12, 13]). The desire to maintain an "intimate space" of privacy, both in physical and digital terms, is in many ways part of a power struggle intrinsic to the relationship between employers and employees. At the same time, having digital technology embedded on workers' bodies (e.g., smart wearables) or tools (e.g., smart vehicles) opens new avenues for control in this relationship. This is an important aspect of the Industrial Internet of Things (I-IoT)—the connectedness of clothes, tools and machinery are the most recent and widespread versions of "smart machines" that gather and send data not only about themselves, but about their users and their work [14]. In this way, the embeddedness of technology in the social is not only a matter of technology-technology or human–technology relationships—it exerts power over human–human relationships and is thus a powerful influence in the social sphere of organisations.

Moreover, the introduction of such technology is sometimes intended to increase safety or security, as shown by Caron's and Guillaume's chapters, illustrating the dual nature of this technology: At the same time as it is aimed at protecting workers' physical safety or security, it can be interpreted as an invasion of their privacy in the workplace [15]. This bears resemblance to the concept of securitisation (or *safetyisation*), where expansion of the power of already powerful actors becomes legitimate and justified in the name of safety. Whether such expansion of power is intentional or not, its future path of development is unpredictable. It can be seen as a conquering process, where a technological logic increasingly colonises the social sphere [16]. In this logic, workers are not only employees responsible for performing their tasks, they are also the sources of data fuelling algorithms that monitor and manage work. In addition to a potential for a general dehumanising of the social sphere of work, it contains implications for safety. If we accept the relevance of safety culture and the importance of employees being empowered to have a strong voice in concerns over safety, diminishing room for a workers' collective and an upgrading of the power of big data analysis could raise serious concerns.

Our intention is not to paint a dystopian image of a future of work where humans are reduced to fuel for technology, and there is no technological determinism involved in this line of reasoning. However, as researchers in safety and risk, it is our role to highlight potentially negative future implications of choices and changes that are made today. The ability to do so depends on being both able and willing to zoom out from the specific empirical observations and ask "what is this a case of?". While we are not calling for all safety research efforts to connect to more generalised macro-implications, we do believe that sociotechnical interconnectivity involves an integration of different forms of risk and an increased probability that what constitutes a solution within one risk framing can present problems in other frames. It might be that it will no longer be sufficient to conduct considerations of risk through specialised and compartmentalised approaches.

12.2.3 The Big Picture Versus Empirical Specificities ("Moving Closer" and "Zooming Out")

As the previous section illustrates, discussions about safety in the digital age tend to mix "small" and specific empirical observations with a "big" picture containing macro-trends and potential futures. Somehow, these two levels of analysis seem to be closely connected. How do we as safety researchers deal with such differences in scale?

The overarching diagnoses of what is going on in the digital age, and what the future might look like, often refer to trends and extrapolations, where the use of different technologies in different contexts are subsumed under the same headings. Debates concerning privacy or AI, for instance, are in essence both ethical and principal and constitute dilemmas for the long-term evolution of societies and the values societies are based on. Here, the "sociotechnical" comes in the form of a macro-oriented, mutually constitutive relationship between technology and society.

At the same time as the future of privacy and the control of algorithms present macro-level, "wicked" problems, digitalisation presents an ongoing flow of concrete cases, with different technologies involved, in different sectors, under different regulations and with different risks involved. This calls for a more case-by-case-oriented study and management of safety, security and reliability in the digital age, as argued by Roe and Fortmann-Roe in this volume. While this involves moving closer to the short- and mid-term singularities of specific challenges, it does not mean a detachment from the discourses of the big picture zooming out on the long-term implications. On the contrary, the case-by-case approach is both an instantiation of challenges belonging to the big picture and an opportunity to inform the big picture with more nuance, differentiation and precision with regard to the stakes and opportunities involved.

This a matter of attaining sufficient granularity to be able to grasp the workings of concrete high-risk systems—their work processes, technical tools, precluded events, and sources of brittleness and resilience—while at the same time aiming to recast them as cases of larger and more principal issues. Importantly, this rather tall order of reframing of scale applies not only to safety researchers but also to technology developers.

We are not implying that company managers or computer scientists and software engineers are evil or malevolent. Neither do we expect them to obtain degrees in Science and Technology Studies to deal with potential unwanted side effects of technology such as illegitimate surveillance and breaches of privacy. What we should expect, however, are governance structures and educating systems supplying them with requirements and competence to perform responsible research and innovation. One way of doing this is to also zoom out from the details of the positive potential of specific technologies under development and critically examine the potential side effects of poor or malevolent use of the technology developed. This form of recasting would probably serve to shatter the myth of technology being objective and neutral once and for all.

12.3 Looking Forward

The introduction to this volume drew up a wide landscape of changes and challenges to provide a backdrop for the discussion of some of the pressing issues associated with digitalisation involving algorithms and machine learning for the safe performance of high-risk systems. Needless to say, this book covers only fragments of the debates, challenges and opportunities associated with safety in a digital world. Despite this, the old and new challenges that are identified through the chapters, illustrate the importance of recognising digitalisation as involving transformation processes that are of a genuine sociotechnical nature. In the digital age, the distinct separation of "the technical" and "the social" components of organisations becomes increasingly problematic. The intertwining of technological and human actors and processes makes it more relevant to explain sociotechnical systems by means of the relationships between heterogenous actors rather than as separate components.

Although the remaining questions remain numerous and challenging, the research community on safety and security has probably never been more relevant in addressing both the small-scale issues associated with particular systems, and the larger and fundamental implications for societal risk governance.

References

1. NSTC, *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*. Executive Office of the President (2016a). Retrieved in October 2019 at https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf
2. NSTC, *Preparing for the Future of AI*. Executive Office of the President (2016b). Retrieved in October 2019 at https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf
3. E. Trist, The Evolution of Socio-Technical Systems. (Ontario Ontario Ministry of Labour, Ontario Quality of Working Life Centre, Toronto, 1981)
4. R.P. Bostrom, J.S. Heinen, MIS problems and failures: a socio-technical perspective, Part II: the application of socio-technical theory. MIS Quart. **1**(4), 11 (1977b)
5. J.C. Le Coze, Risk management: sociotechnological risks and disasters, in *Handbook of Risk Studies*, ed. by A. Burgess, O. Zin, A. Aalemanno (Taylor and Francis, London, 2016)
6. T.K. Haavik, On components and relations in sociotechnical systems. J. Contingen. Cris. Manage. **19**(2), 99–109 (2011)
7. B. Latour, *We Have Never Been Modern* (Harvester Wheatsheaf, New York, 1993)
8. B. Latour, *Pandora's Hope: Essays on the Reality of Science Studies* (Harvard University Press, Cambridge, MA, 1999)
9. D. Smith, M. Fischbacher, The changing nature of risk and risk management: The challenge of borders, uncertainty and resilience. Risk Manage. **11**, 1–12 (2009)
10. NOU 2015: 13, *Digital sårbarhet—sikkert samfunn [Digital vulnerability—secure society]*. White paper on digital vulnerability (in Norwegian). (Ministry of Justice and Emergency Preparedness, Norway, 2015)
11. E. Roe, P.R. Schulman, *High Reliability Management: Operating on the Edge* (Stanford University Press, 2008)
12. M. Crozier, *The Bureaucratic Phenomenon* (University of Chicago Press, 1964)
13. S. Lysgaard, *Arbeiderkollektivet* (Universitetsforlaget, 2001)
14. S. Zuboff, in *The Age Of The Smart Machine: The Future Of Work And Power* (Basic Books, 1988)
15. S. Antonsen, P.G. Almklov, Revisiting the issue of power in safety research, in *Safety Science Research: Evolution, Challenges and New Directions*, ed. by J.-C. Le Coze (CRC Press, 2020)
16. J. Røyrvik, A. Berntsen, Verden som teknologi: Allerede alltid erobret. *Norsk Antroplogisk Tidsskrift* **33**(1), (2022)