FOOD FOR THOUGHT

Bringing together humanity and technology in context

Future challenges for safety in high-risk industries

Corinne Bieder

Publication coordinated by Caroline Kamaté

n° 2024-04

THEME

The future of safety

Les Cahiers de la sécurité industrielle

FONCSI
Fondation pour une culture de sécurité industrielle

The FonCSI, Foundation for an Industrial Safety Culture, is a public interest research foundation created in 2005, located in France.

The FonCSI finances research projects concerning potentially hazardous industrial activities and their interaction with society, and aims to encourage open dialogue with all stakeholders (regulators, associations and NGOs, local government, researchers, industrial firms, trade unions, etc.).

Its originality is the interdisciplinary nature of its activities, in France and internationally, as well as a strong commitment to innovation and to anticipating tomorrow's issues.

FonCSI's missions

- Identify and highlight new ideas and innovative practices
- Develop and fund research into industrial safety and the management of technological risks
- Contribute to the development of a research community in this area
- Transfer research results to all interested parties

FonCSI's values

- Foresight and innovation
- Knowledge and know-how
- Openness and exchange

FONCSI
Fondation pour une culture
de sécurité industrielle

Foundation for an Industrial Safety Culture
Public interest research foundation

http://www.foncsi.org/

6 allée Emile Monso – CS 22760
31077 Toulouse Cedex 4
France

Telephone:	+33 (0) 532 093 770
X:	@TheFonCSI
Email:	contact@foncsi.org

This document

Title	Bringing together humanity and technology in context
Subtitle	Future challenges for safety in high-risk industries
Keywords	safety; digital; artificial intelligence; human; technology; uncertainty
Author	Corinne Bieder
Publication date	April 2024

This *Cahier de la sécurité industrielle* originates from the FonCSI's Strategic analysis "Work and workers in the 2040's". This project brought together a core group of academic researchers and scientific experts from industrial companies and organizations sponsoring FonCSI. They met about 15 times with the aim of exploring the impacts on the safety of high-hazard organizations of the megatrends our world and societies are undergoing. This document focuses on the role of humans and their relationship to technology.

About the author

Corinne Bieder holds a PhD in Sociology and Management Science, former engineer, she also holds a master's degree in Risk Management and a specialized master's degree in Ergonomics. After working at EDF, Dédale and Airbus, she joined ENAC (the French Civil Aviation University) where she is responsible for the Safety Management research program. She was a member of the FonCSI's strategic analysis scientific core-group and became FonCSI's scientific director in 2024.

To cite this document

Corinne Bieder (2024). *Bringing together humanity and technology in context-Future challenges for safety in high-risk industries.* Number 2024-04 of the series *Les Cahiers de la sécurité industrielle,* Foundation for an industrial safety culture, Toulouse, France.

Free download on: Welcome to FonCSI's website — FonCSI

Summary

The fast pace of evolution of digital technologies is shaking up the conventional high-hazard industry landscape, introducing new challenges for safety. This document focuses on the role played by humans in the 2030-2040 timeframe as regards how safety is managed and governed. The results highlight the impact of the implicit framework adopted to appreciate the respective contributions of humans and digital technologies to the safety of high-hazard industries. Whereas a human-centered framework emphasizes specific human capabilities such as empathy, making sense, judgment, as critical to safety, a technology-centered one focuses on computational power and speed as promises to future safety. None of these frameworks opposing humans and digital technologies seems appropriate to account for real situations where they both coexist and are interrelated in more complex ways than just through man-machine interfaces. Furthermore, they are part of a broader social, political, organizational, and cultural context calling for qualifying absolute statements on Technology and Humanity. More generally, high-risk operations are complex. Thinking in terms of dichotomies (e.g., technology/humans; digital/non-digital) is too simplistic to anticipate the safety challenges ahead of us. Exploring the interrelations between humans and digital technologies includes investigating the context in which they evolve to frame possible future safety challenges in a relevant manner. This means involving diverse perspectives and disciplines to bring together humans and technologies in context and reflect the complex reality.

Keywords: safety; digital; artificial intelligence; human; technology; uncertainty

Foreword

When this strategic analysis started in 2019, we, at FonCSI, internally used to nickname it 'The operator of the future'. It was indeed mainly targeting the impact of anticipated technological and societal evolutions on the future role of frontline operators on the safety of industrial operations. This subject has been indeed dealt with. But it has been largely exceeded.

This was due, obviously, to profound changes in the world itself. In 2019, SARS-COV-2 was still the obscure prerogative of a few exotic bats (if not the secret captive of some 'high-security' lab...). Ukraine was not at war. And the 'energy crisis' had not yet breezed on the general public the gentle premises of the ecological cataclysm that mankind has managed to set up. And of course: Chat GPT did not exist yet.

But a second reason is that FonCSI's strategic analyzes share with other research activities a well-known feature: when successful, they rarely follow the anticipated path. And this analysis was particularly fertile. In her synthesis, Corinne Bieder remarkably captures it. She widens the gaze far beyond the traditional Fitts' list-inspired dichotomy between Humans and Technology. And even beyond the more recent Human-Technology interaction perspective.

She takes the strategic analysis' lesson to its very essence and its multidisciplinary, and quasi philosophical implication: safety is an emerging behavior of our complex societies, a social, political, cultural, emotional, and recursive, construct of that behavior. She warns us against the temptation of delusive simplifications. A warning more essential than ever in these times of turbulent uncertainty.

Jean Pariès,
former Scientific Director, ICSI-FonCSI

Contents

Introduction

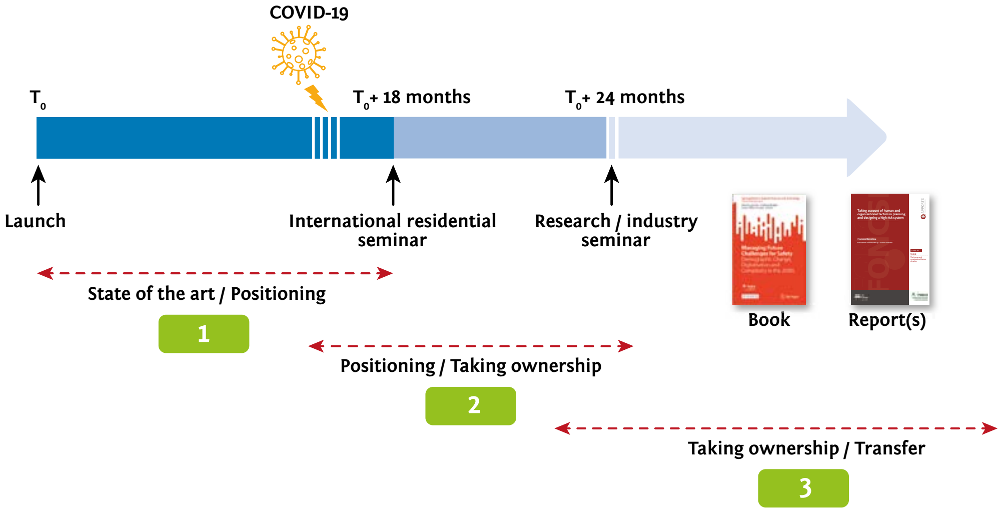
Context

The idea and content of this document mainly emerged from a strategic analysis led by FonCSI (Foundation for an Industrial Safety Culture) on the role of humans in the safety of high-hazard industries in a 2030-2040 timeframe. The objective of this work was not only to consider how megatrends are likely to affect the world and more specifically industry in the decades to come, but also to reflect upon how these trends could affect the role of humans in the safety of tomorrow's industry. Among these trends is the fast pace of development of digital technologies and the many ways in which it affects and may continue to affect the evolution of the world (Okhrimenko, Sovik, Pyankova, & Lukyanova, 2019). In high-hazard industries, some of the questions raised by this technological evolution had already emerged with the development and spread of automation. However, artificial intelligence paves the way to new opportunities, but also raises new issues. While the role of humans in the safety of high-hazard industries is currently commonly acknowledged despite the remaining human, organizational and cultural factors (Roe & Schulman, 2008; Daniellou, Simard, & Boissières, 2011), the advent of a technology purporting to be "intelligent" may challenge it.

The strategic analysis at a glance

The strategic analysis was entitled "Work and workers in the 2040's". Based on the concerns raised by FonCSI's sponsoring organizations (transport and energy industries, regulatory authorities, and other bodies) on this topic, this analysis aimed to generate quality research in a relatively short time and to create a continuum between research, innovation, and industry. As previous FonCSI's strategic analyses, the project comprises 3 main steps (described in the figure below):

1. **The state of the art**. This consists of an assessment of current research with the world's leading experts. Its highlight is a high-level research seminar with international academics identified and invited by the GSAS. This seminar provides most of the scientific material on which the future book relies.
2. **Taking ownership**. The GSAS appropriates the outcome of the international seminar and compares the theoretical results and concepts to how they are actually used in practice in the industry. This step ends with a seminar gathering researchers and industry partners.
3. **Transfer**. FonCSI oversees the dissemination of research results and translation of outcomes into industrial practices through different publications: an academic book and synthetical reports such as the *Cahiers de la sécurité industrielle*. In line with FonCSI's public interest status, all publications are open-access.


COVID-19
T_0
T_0+ 18 months
T_0+ 24 months
Launch
International residential seminar
Research / industry seminar
Book
Report(s)
State of the art / Positioning
1
Positioning / Taking ownership
2
Taking ownership / Transfer
3

Of course, as for many activities at that time, the schedule of the strategic analysis has been disturbed by the COVID-19 crisis.

Beyond these classical steps, FonCSI also organized, in the course of the strategic analysis, a foresight workshop on the future of rail safety. The objective was to stimulate debate between recognized rail experts and to identify ways to anticipate some rail safety issues that could emerge in 2030 and beyond. The guests who participated in this event were :

- François Davenne, Director General of UIC (International Union of Railways);
- Loïc Dorbec, President of AGIFI (French Association of Independent Rail Infrastructure Managers);
- Yann Leriche, CEO at Getlink (Eurotunnel, Europorte, ElecLink and CIFFCO);
- Pierre Messulam, Safety and Security Risk Director of the SNCF Group;
- Dominique Riquet, European Parliament, member of the Transport and Tourism Committee.

The discussions were moderated by Pierre-Franck Chevet, President of IFP Energies Nouvelles and Jean Pariès, former Scientific Director of FonCSI and ICSI.

The strategic analysis scientific group

The project was led by the Scientific Group for Strategic Analysis (GSAS) of FonCSI, and coordinated by Caroline Kamaté. The latter consists of a permanent core committee of researchers who participate in all the strategic analyzes carried out by FonCSI:

- René Amalberti, FonCSI, France
- Corinne Bieder, FonCSI/ENAC, France
- Hervé Laroche, ESCP Business School, France
- Jean Pariès, FonCSI/ICSI, France
- Jesús Villena López, Ergotec, Spain

On this particular topic, the committee was reinforced by experts from FonCSI sponsoring organizations, widely recognized in the field of security and risks:

- Florence Reuzeau, Airbus, France
- Raluca Ciobanu, EDF, France
- Laurent Cebulski & Bruno Dember, EPSF, France
- Franck Ollivier, Eurovia, France
- Nicolas Engler & Thierry Escaffre, GRTgaz, France
- Dounia Tazi, ICSI, France
- Alexandre Largier & Tania Navarro Rodriguez, IRSN, France
- Stella Duvenci-Langa & Cyril Cappi, SNCF, France
- Raphaël Waxin, TotalEnergies, France

The international experts

For this strategic analysis, the disruptions linked to the Covid 19 crisis prevented the holding of a residential academic seminar. The seven internationally renowned experts who are listed below therefore presented their work at a remote seminar in November 2020:

- John Allspaw, Adaptive Capacity Labs, USA;
- Stian Antonsen, Norwegian University of Science and Technology (NTNU), Norway;
- Michael Baram, Boston University, USA;
- Flore Barcellini, Cnam, France;
- Gérard de Boisboissel, Saint-Cyr Military Academy Research Centre, France ;
- Steven Shorrock, Eurocontrol, United Kingdom and France
- Akira Tosé, Niigata University, Japan.

The main objective of this seminar was to contrast points of view, induce an emulating debate and propose areas for improvement. Its final aim was to come up with a collective book, published in the open access series 'SpringerBriefs in Safety Management' (Laroche, H. ; Bieder, C. ;Villena-Lopez, J. (Eds), 2022).

This document: structure and goals

This document more specifically addresses the following question: how can humanity and technology be brought together to ensure the safety of high-hazard industries in the decades to come?

It starts by exploring a fundamental question raised by a technology tending to become more human, namely: what makes us human? There is an abundant literature on what makes us human compared to other species, but reflecting on the role of humans in the safety of tomorrow's high-hazard industry raises a somewhat different question, namely: what makes us human, and is thus unique in contributing to safety, compared to new digital technologies?

Trying to provide an answer to this question raises another one which is then addressed in the document: on what grounds are the unique capabilities of humans and/or technologies appreciated?

Yet, in daily life, and even more so in high-hazard industries, humans and technologies do not exist separately from one another. The document therefore continues with a description of how humanity and technology are interrelated, and a discussion on how to bring them together and envisage them together in a consistent way from operations to global organizational, societal, and even philosophical levels.

Further reading

We refer the interested reader to the bibliography section at the end of the document, and in particular to the other publications of the strategic analysis on "The contribution of human work to safety in high-risk industries by 2040":

- the book published in open access by Springer in October 2022;
- the summary of the railway workshop published in May 2021, freely downloadable from the FonCSI website (document in French language).

SpringerBriefs in Applied Sciences and Technology
Safety Management
Hervé Laroche · Corinne Bieder · Jesús Villena-López Editors
Managing Future Challenges for Safety
Demographic Change, Digitalisation and Complexity in the 2030s
FONCSI
OPEN ACCESS
Springer

1

A tendency to oppose digital technologies and humans

Trying to identify unique human capabilities or characteristics is not new. During the Enlightenment, a "*general tendency took hold to define human beings as being outside of nature, and Enlightenment rationalism led to a vision of a technologically dominated environment*" (Williams, 2020). Anthropologists have also made a great effort, and continue to do so, in investigating the question of what makes us human, especially in comparison to apes (see for example (Pollard, 2009)).

With the progress in medicine and biotechnologies and the advent of the "posthuman" age, the same question was raised from a different perspective. In the attempt to compare humans with other species or elements, technological progress, exacerbated in recent years by the development of digital technologies, seems to have generated another reference against which humans try to compare themselves. Norman (2014) underlines that the evolution of society has promoted a machine-centered perspective and calls for reversing this machine-centered point of view and adopting a human-centered point of view. In that respect, he continues the tradition of opposing humans and machines as if it were essential to take sides and think of them separately.

While reflecting about the role of humans in the safety of high-hazard industries in the future, the unique human capabilities were also extensively mentioned. Most of these were similar to the ones put forward at the time of the introduction of automated systems but others were introduced on different grounds. Shorrock (2022) while analyzing micronarratives from healthcare practitioners following the COVID 19 pandemic, highlights as examples the following capabilities required to cope with rather unpredictable situations: "*ability to anticipate potential challenges required imagination and a deep understanding of the realities of everyday work*" or "*collective intelligence via inclusive collaboration and open communication*" for preventing harm both to patients and healthcare workers. The author also cites previous works on these unique capabilities: Cook (2020), "*Human practitioners are the adaptable element of complex systems*", and "*The system continues to function because it contains so many redundancies and because people can make it function, despite the presence of many flaws*" or Dekker (2015, p. vi), "*people as the source of diversity, insight, creativity, and wisdom about safety, not as sources of risk that undermine an otherwise safe system.*" Anticipation, imagination, collective intelligence, adaptability, diversity, insight, creativity, and wisdom appear as human-specific characteristics that are critical to safety. Boucher (2019), cited by Antonsen (2022) adds to this list of human unique capabilities by mentioning what is missing from AI technology: "*It is not alive, it does not have a consciousness, and it is completely uncapable of understanding, creativity and empathy*". The main aspects that make these capabilities critical to safety include the many flaws present in systems, but also their complexity and unpredictability.

In contrast, several unique machine-capabilities are put forward by technologists, the first of which being the information-processing capacity. Others are also mentioned depending on the context, such as the ability to perform physically difficult, hazardous, or repetitive, boring tasks. In high-hazard activities, what is also highlighted is the error-free performance, the reaction time, or the availability for operations 24 hours a day (De Boisboissel, 2022), in other words, the reliability and consistency of the performance and its reproducibility.

Although the two points of view seem irreconcilable, dichotomies tend to simplify reality. Reaching beyond this limited understanding of humanity and technology and general statements about humans on the one side and machines on the other, might require making the link between the unique capabilities of humans or machines respectively and the contexts in which they are relevant or needed. Building on the example developed by Norman (2014) on distractibility, it is important not to be distractible when performing certain tasks and yet, distractibility is also what can allow new events in the surrounding environment to be noticed, which might be just as critical, if not more, especially to the safety of high-hazard operations. Making explicit the context and referential against which capabilities are appreciated is critical for taking the reflection further.

2

The implicit underlying appreciation framework

Among the consequences of the machine-centered point of view pushed by society, Norman (2014) highlights that machines become the reference against which everything is evaluated. Humans are no exception, being judged on “artificial, mechanical merits”. Common illustrations are the comparison of AI and human information-processing capacity or the point that machines make no mistakes (or more specifically that humans make mistakes as opposed to machines).

Conversely, human capabilities such as empathy, adaptability, creativity, wisdom, and their absence in machines so far, are established on human-specific merits. It is no surprise that by using different referential, one comes to different conclusions, positions, and proposals.

Antonsen (2022) introduces some nuance in how to appreciate the capabilities of humans and algorithms by making explicit the scope of use. He distinguishes between two types of tasks. The first type is those where there is little variability such as simple actions and recurring decisions, where algorithms represent “*the static pre-programming of expert knowledge*”. For these tasks, the author confirms that this first wave of AI technology, driven by humans, can alleviate routine tasks performed by humans as well as other more complex tasks. For the second type of tasks, where right or wrong solutions cannot always be easily identified, or handled through simple rules and coding, the conclusions are not so obvious despite the second wave of data-driven AI algorithms, building on machine learning and neural networks. Although AI includes the word intelligence, the author calls for a distinction between different types of intelligence. The first is specialized, narrow intelligence, which refers to the perfect performance of a particular task such as face recognition or obeying a voice command. Antonsen reminds us that this kind of AI “*can only apply intelligence to the specific problems it is programmed for*”. The second is physical intelligence, which links mental and motor skills. The author highlights that for this type of intelligence, the effort put into AI and the performance it achieves is disproportionate with what can be achieved by any human, taking the example of robots “running” on an uneven terrain. He emphasizes that “*while our information-processing capacity in narrow domains may be limited, our action repertoire in the physical world is not particularly limited, since we can improvise with whatever tools and information we have at hand.*”[1] (Antonsen, 2022). The final category, common-sense intelligence, refers to “*the ability to interpret and understand virtually any situation and learn how to act in the situation*”. This involves a number of capabilities such as contextual understanding of a social situation, sense making or “*making good guesses in rare situations, based on incomplete information*”. Despite the accelerating development of AI, these capabilities, which are deemed critical to safety by the author, are not yet accessible to algorithms and remain human unique ones, at least so far.

Distinguishing between these various tasks and types of intelligence allows Antonsen to qualify general statements such as the superiority of AI information processing capacity over that of humans by insisting on the specific form of intelligence it applies to, namely “*on narrowly defined areas, where there is sufficient data and the similar situations occur over and over again*” (Antonsen, 2022). In these cases, replacing humans by machines may have significant value, even more so when it protects humans, such as soldiers for example, from exposure to hazardous situations like on the battlefield or allows for a permanent and continuous presence, as developed by De Boisboissel (2022).

However, real operational situations involve more than these areas. Getting back to what the task, or more generally the situation requires is necessary to avoid reaching hasty and approximative conclusions about humans and digital technologies.

[1] "We" and "Our" is a reference to humans in general.

3

Bringing together humanity and technology in high-hazard industries

3.1 Operational questions

In high-hazard industries, the unique capabilities of both humans and machines are relevant and called upon at the sharp end due to the complexity of situations and their potential consequences. Noticing things in the surrounding environments and being able to interpret them as requiring a change in performance mode or course of action is as critical as doing a repetitive task in a reliable way. The existence of uncertainties in operations is widely acknowledged even though it does not necessarily translate into formal safety management strategies. Some of these uncertainties are related to the design of technology itself due to the incompleteness of the underlying knowledge and the unavoidable limitations of the model of the world used as a reference, as underlined by Downer (2011; 2020). With the development of digital technologies and more specifically of AI algorithms capable of learning and progressively changing their "behaviours", the uncertainties, through the emergence of new ones, are even more blatant since more obviously created by the technology itself. As a consequence, operating safely with these (new) uncertainties requires more than ever the "common-sense" intelligence as introduced by Antonsen (2022) and unique human capabilities such as adaptability, creativity or wisdom on safety.

More generally, although partly autonomous, self-learning AI algorithms technologies still require human intervention in operations[2]. Even in advanced technological domains such as the military, where the merits of robots are widely acknowledged for reasons of exposure, any military action requires a human leader as well as operators. The reasons for that are reiterated by De Boisboissel (2022) as follows: "*The leader is the human key of any military action. He gives meaning to it, he stays responsible of the maneuvering and the conduct of the war and he adapts "en conduite" depending on events*". "*The operator has the closest situational awareness, and, if machine can make probabilistic calculations, probability does not take into account the complexity of military situations on the battlefield which require human analysis*". However, rather than opposing humans and machines, operations bring them together as collaborating parties.

Trust appears as a major challenge raised by this collaboration in high-hazard activities due to the potential safety-related impacts of operations. De Boisboissel (2022) underlines the importance of trust for the military to use self-learning systems. These systems reach beyond the traditional issue of trust in technology raised at the dawn of automation. Trust in adaptive algorithms not only involves the design/development of the algorithms, but also their training and what it entails in terms of data.[3] Transparency is commonly advanced as a way to build trust. "*Adaptive and self-learning systems must be able to explain their reasoning and decisions to human operators in a transparent and understandable manner*" (De Boisboissel, 2022). Antonsen (2022) calls it a "transparency paradox" to be solved before sophisticated AI is introduced into high-hazard industries' safety critical processes. "*When algorithms learn, they become actors in safety management. It is a basic principle of HROs and most advice on safety management that important decisions and actions are checked and double-checked.*" (Antonsen, 2022).

Besides the trust in the algorithms themselves, the massive use of digital technologies gives rise to new types of vulnerabilities, especially cyber-attacks. This aspect is developed in the first Cahier originated from the strategic analysis that synthesizes the impacts of megatrends on the future of safety in high-hazard industries (SASG "Operator of the future", 2023).

[2] Keeping digital technologies working requires a significant yet invisible human effort in the background.

[3] Algorithm development and training are other occasions bringing together humans (data scientists in this case) and digital technologies as addressed in a later section.

3.2 Organizational and design questions

The advent of second-wave AI not only raises questions about the collaboration between humans and these technologies during operations, but also at more organizational levels. It induces organizational changes or extends the scope of design, thereby affecting the interrelation between humans/organizations and advanced digital technologies.

3.2.1 Multiplication and diversification of stakeholders requiring increased coordination

Self-learning algorithms involve coordination not only between digital systems or humans and digital systems, but also between organizations, and thus between humans working closer to the blunt end than to the sharp end. A first set of organizations to coordinate are those contributing to the operation of the high-hazard systems. There is a multiplication and diversification of stakeholders supporting operations with newcomers such as software developers[4], communication services providers or data providers. The fragmentation of organizations that started decades ago has been amplified not only in number but also in diversity. High-hazard industries become more open due to their reliance on these new activities and associated organizations that are not part of their industry per se. In this respect, the spreading of digital technologies leads to the blurring of industrial boundaries. Thereby, safety becomes a concern of diverse importance among the organizations involved in operations. Besides the increased need for coordination, the multiplication of organizations involved in operations might lead to atomization and possible loss or transfer of responsibilities. The situation can become even more complex. Indeed, some algorithms are freely available from the internet without any identified organization behind them. At yet another level, when capabilities (e.g., algorithms or communication satellites) from another country are used to operate a high-hazard system, geopolitical conditions might directly impact the operation of the system. These vulnerabilities come on top of cyber-risks increased by the openness of advanced digital systems including in particular sensors, transmission capabilities, algorithms or packages sometimes available to anyone.

Another challenge with the introduction of specialized digital systems and organizations in the scope of high-hazard industries is that of competencies. As noted by one of the railway experts involved in the workshop mentioned in the methodology section, these days, digital systems are developed by people who are unfamiliar with the railway system and its physical characteristics. The disconnect between the high-hazard and the digital industries disembodies the design of the AI part of technological systems. It might be even more critical that the competences needed within the high-risk organizations to check and validate/approve the digital systems developed by digital expert organizations might not be available with the possibility of this being delegated to a third party (involving an additional stakeholder). This situation again raises the question of responsibility and accountability[5], which might become even more complex in some cases where the algorithms are developed in countries where the regulation differs from that applicable in their country of implementation.

3.2.2 Data-based representation of the world

In addition to the degree of familiarity of software developers with operations, the design of self-learning algorithms involves an aspect that did not exist in previous technologies, namely data. As Antonsen (2022) reminds us, a first characteristic of self-learning algorithms is that they can only be developed in areas where data can be collected, and in sufficient quantity. A second one is what data will it rely on? Last, the outcome of these algorithms will heavily depend on the training data set. As an example, a Natural Language Processing algorithm trained on general data (text in this case) available from the web might not lead to the same outputs as the same algorithm trained on domain specific (e.g., railway, nuclear, aviation) text. Antonsen (2022) reminds us, citing Parmiggiani et al. (2021), "*data is rarely "discovered" as objective facts and analyzed as such – it is both selected and prepared before it is available for analysis*". As such, the chosen data shape the algorithms or introduce biases as stated by Antonsen (2022). "*In safety-critical contexts, we can't afford to overlook the fact that data will be biased, labelling contains bias, and that learning algorithms can create their own set of bias. Therefore, there is no reason to believe that AI removes human fallibility. It replaces one form of human fallibility with another.*"

As a consequence, in the same way as humans build into the technology their limited model of the world, data scientists build into AI algorithms their data-restricted representation of the world through both the data available/collected in sufficient quantity and the data selected to train the algorithms. AI algorithms in turn produce outcomes used by people other than designers or developers who build their representations

[4] A huge number of flaws in algorithms are tracked and corrected in real-time by myriads of software developers.

[5] These aspects will be further addressed in this section 3.3.

on these premises without even being aware of them. The transparency of algorithms as a condition to build trust in self-learning systems might therefore encompass not only what algorithms are doing but also on what grounds they are doing it.

3.3 Societal questions

3.3.1 Regulation and governance

When it comes to high-hazard industries, the issue of trust in and transparency of AI algorithms does not only affect operators or leaders in the operational context but also the whole society possibly affected by the potential safety consequences of an operational accident. As such, they raise governance issues. These issues are even more urgent given that traditional regulatory and oversight approaches do not apply to self-learning algorithms or more generally to systems with behaviors evolving over time or that are unpredictable. The control model historically underpinning systems certification (one of the pillars of safety governance) disregarded the variability of human performance despite numerous illustrations and documentations of it. Certification relied (and still does) on probabilistic safety analyses and testing deemed sufficient to anticipate the safety level of technological devices and their use. With the common acknowledgment that the behavior of self-learning algorithms cannot be predicted, the limitations of traditional certification approaches become clear. Ironically, while these approaches were deemed sufficient to address the behavior of first-line operators, they are considered inappropriate for certifying technologies that are becoming more "human", especially in the sense that their performance is no longer deterministic or even probabilistic. Besides certification, it is the whole governance approach that is to be revisited. According to Antonsen (2022), "*regulators and supervisory authorities will never allow critical decision-making processes to change unsupervised. The governance of self-learning algorithms requires regulation, audit tools and competence, something which is not in place, and it is hard to see how this can be done, at least within a prescriptive regulatory regime*".

Self-learning algorithms have made it even more obvious that we need to acknowledge uncertainty and learn to live with it. The COVID 19 pandemic was also an occasion to experience it the hard way. However, it provided the opportunity to envisage alternative approaches to governance as described by an anesthesiologist-intensivist interviewed by Shorrock (2022): "*For many professionals, it has created a touching sense of humility, both among frontline actors and managers. I believe that this humility has facilitated communication and the emergence of a shared governance between caregivers and administrators where I've been working.*" Although this experience refers to the governance within organizations, similar phenomena have occurred during ad hoc exchanges between regulators and the regulated. In aviation for example, the unprecedented experience of having pilots unable to comply with the required number of flights or undergo the required recurrent training to keep their license valid led to unprecedented collaboration between the different parties to find ways forward reaching beyond the current prescriptive approach. On a wider scale, that of the governance of science, Jasanoff (2007) goes one step further by calling for an even more inclusive approach to live with uncertainty and ignorance, by involving citizens as well, not only to find solutions but to frame the problem in the first place.

Capturing the multi-facets of the problem and its contextual specificities might be even more critical at a time where a majority of societies are concerned with other major issues such as security or climate change. As advanced by one of the participants in the railway workshop, the question of the financial resources supporting this technological revolution is key. Is it reasonable to accumulate expensive technologies with respect to societal stakes and the increasing concerns for climate change? More generally, revisiting the governance of digital technologies, and especially in high-hazard industries in our case, might require reaching beyond not only the traditional regulator-regulated actors, but also the sole issue of safety. As raised by Jasanoff (2007): "*Is it sufficient, for instance, to assess technology's consequences, or must we also seek to evaluate its aims?*"

3.3.2 Ethics

One of the aspects that would need to be addressed in a new governance approach would be ethics. As underlined by De Boisboissel (2022) in a military context, "*the consideration of ethical issues in the development, maintenance and execution of software becomes an imperative induced by the autonomy of robotics systems*". Other high-hazard industries/activities also share this concern, which was explicitly mentioned during the railway workshop, in association with accountability issues. More generally, there is a wealth of literature on AI ethics. According to Siau & Wang (2020) "*Building ethical AI is an enormously complex and challenging task*". Preliminary questions could be is it possible at all or is it even desirable? Feelings are mixed about putting the opposition between humans and machines back on the scene. According to De Boisboissel (2022), "*a machine is by nature amoral. The human remains the one and only moral agent, therefore the only one with responsibility.*" Citing Lambert (2020), the author adds that "*ethical reasoning requires situational awareness and consciousness, which are uniquely human characteristics*" (*ibid*). The main challenge therefore becomes how to make AI and humans collaborate to make ethical decisions in critical situations? Conversely, trying to build ethics into machines is rather about trying to make machines become more human. Such an approach, falling into a replacement perspective[6] (in contrast with a collaborative one), implicitly considers that it is feasible provided sufficient resources are dedicated to it. These are different understandings of the problem that it would be worth refining through more contextualized and inclusive reflections, coming back to a higher-level common problem, e. g. how to make ethical decisions? What is an ethical decision?

3.3.3 Legal framework

The issue of accountability underlies more or less explicitly almost all the questions raised so far at all levels, whether operational, organizational & design related, or societal. Therefore, the legal framework is central to the reflections bringing humanity and technology together, even more so with the capabilities of self-learning algorithms and the aforementioned multiplication of stakeholders with the introduction of these technologies.

However, as mentioned during the railway workshop, we manage to enhance the safety level at the cost of increased complexity and fragmentation involving a loss or transfer of responsibility. The legal framework, as it exists in Europe, cannot address uncertainty and fragmentation as it evolves. It is very hard to provide a legal framework for technological progress. The law cannot keep pace with progress in technology. The lawmaker tries to define a framework, but at the end of the day, there is a judge. The perfect system doesn't exist. The idea there is a legislative framework that would clarify all the responsibilities is an illusion. The eye of the judge and that of society are the ones that matter in the end. This is another call for more inclusive and contextualized reflections about AI.

3.3.4 Societal and philosophical questions

The use of AI in high-hazard industries, but also well beyond, questions the relationship between humans and technology at all levels, including a philosophical one. The subject is not new. In 1954, Heidegger (1954) was already advancing that "*we have a technological understanding of ourselves and the world*". However, the question is raised in different terms with the further step taken by self-learning technological artefacts towards "humanization" with the use of expressions such as artificial "intelligence" or "ethical" AI. "*The world in which we live, after all, is increasingly populated not only by human beings but also by technological artefacts that help to shape the ways we live our lives*" (Verbeek, 2005). With the advent of advanced AI and autonomous systems, a different question arises in relation not to the way we live our lives but rather to the way we could lose it: to what extent can the safety of humans and the environment be managed by technology?

As stated by Williams (2020), with the progress of machines, "*technology itself came to be seen as deterministic, "an autonomous, transhuman force in social affairs." (…) we mentally distinguish ourselves from technology (…). And where does that leave us? Are we at the mercy of, or in control of, both technology and nature, or of one over the other, or of neither?*".

Norman (2014) sees the relationship between humans and machines "*as much [as] a social problem as a technological one*". The author claims that "*it is primarily our social structures that determine both the direction that technology takes and its impact upon our lives*" (Preface). It seems that with digitalization, not only has the society adopted a technology-centered point of view, but is also to some extent limiting its representation and understanding of the world to what can be sensored or apprehended through data (in massive quantity). This turn seems to rely on some kind of myth or magic. As highlighted by Antonsen (2022) "*there is a form of epistemic uncertainty built into models and algorithms that gain a form of objectivity because they are seemingly untouched by human fallibility of judgement, while in fact they are not.*"

[6] This replacement perspective mainly applies to the operational phase. The role of humans tends to be reinforced in the design, development and maintenance/correction of AI systems.

Conclusions: Towards an inclusive and contextualized reflection on humans, AI, and safety

With the accelerating advances in digital technologies, especially in AI, new questions are being raised about the role of humans in the safety of high-hazard industries in the coming decades. Machines are seen as more and more human, notably with the use of words like "intelligence" and their unique capabilities are commonly compared in opposition to human weaknesses. Yet, going further into the referential used to characterize humans or machines, or to the actual capabilities of machines beyond the socially constructed myth leads to qualified definite judgments on their respective capabilities.

Challenges arise when humanity and technology are thought of as being apart from one another on all scales and when one wants to make humans more machine-like and machines more human-like. Both have unique capabilities and cannot reasonably be replaced by the other in all respects. Both exist in today's world at all scales, but what might require increased attention is making them fruitfully exist together. This might require considering them for what they are and thinking and designing their complementarity and articulation to avoid fostering/nurturing competition. It also means contextualizing the reflections rather than remaining at very general levels. Distinguishing between different scales, different types of activities, but also between different work situations, organizations, or socio-economic, legal, and political environments is necessary to reflect about the conditions needed to make the collaboration between humans and AI technologies relevant to the particular context at stake. The use of sophisticated AI in high-hazard industries might be different from that in non-critical activities. As stated by Antonsen (2022), "*we should rethink the way we conceptualize and study the relationship between the human and technological agent of safety*". Just as ergonomists suggest starting from work situations and involve operators to design new systems, getting back to real situations and involving the experts of these situations might lead to more realistic and fruitful debates around the relationship between humans and machines in the operational field and better-grounded decisions as to their articulation. Yet, this relationship is also to be thought about and discussed beyond operations at higher levels, especially organizational, societal, and philosophical, in order to develop a broader reflection on the relationship between humanity and technology. Indeed, these levels are not disconnected from one another and combining elements from all these levels and diverse perspectives could help to go beyond the current simplistic dichotomy between humanity and technology and some of the challenges that derive from it.

References

Antonsen, S. (2022). Between Natural and Artificial Intelligence□Digital Sustainability in High-Risk Industries. In H. Laroche, C. Bieder, & J. Villena-López (Eds.), *Managing Future Challenges for Safety* (pp. 41-50). Cham: Springer. doi:https://doi.org/10.1007/978-3-031-07805-7_5

Boucher, P. (2019). *How artificial intelligence works.* European Parliament. Retrieved from https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/634420/EPRS_BRI(2019)634420_EN.pdf

Cook, R. I. (2020). How complex systems fail. *HindSight, 31*, pp. 13-16.

Daniellou, F., Simard, M., & Boissières, I. (2011). *Human and organizational factors of safety: a state of the art.* Toulouse, France: Foundation for an Industrial Safety Culture. Retrieved from https://www.foncsi.org/en/publications/collections/industrial-safety-cahiers/human-organizational-factors-of-safety

De Boisboissel, G. (2022). Evolution in the Way of Waging War for Combatants and Military Leaders. In H. Laroche, C. Bieder, & J. Villena-López (Eds.), *Managing Future Challenges for Safety* (pp. 13-24). Cham: Springer. Retrieved from https://doi.org/10.1007/978-3-031-07805-7_2

Dekker, S. (2015). *Safety differently: Human factors for a new era* (Second ed.). CRC Press.

Downer, J. (2011). «737-Cabriolet": the limits of knowledge and the sociology of inevitable failure. *American Journal of Sociology, 117*(3), pp. 725-762.

Downer, J. (2020). On ignorance and apocalypse: A brief introduction to "epistemic accidents". In J.-C. Le Coze (Ed.), *Safety Science Research: Evolution, Challenges and New Directions.* Boca Raton: CRC Press.

Lambert, D. (2020). *Que penser de...? La robotique et l'intelligence artificielle.* Fidélité.

Laroche, H. ; Bieder, C. ;Villena-Lopez, J. (Eds). (2022). *Managing Future Challenges for Safety.* Cham: Springer. Retrieved from https://link.springer.com/book/10.1007/978-3-031-07805-7

Norman, D. (2014). *Things that make us smart: Defending human attributes in the age of the machine.* Diversion Books.

Okhrimenko, I., Sovik, I., Pyankova, S., & Lukyanova, A. (2019). Digital transformation of the socio-economic system: prospects for digitalization in society. *Revista Espacios, 40*(38).

Pollard, K. (2009). What makes us human? *Scientific American, 300*(5), pp. 44-49.

Roe, E., & Schulman, P. (2008). *High Reliability Management: Operating on the Edge.* Stanford, CA: Stanford University Press.

SASG "Operator of the future". (2023). *Industrial safety in a changing world – Human, digital, new organizations: 10 key points for 2040.* Toulouse: Foundation for a safety culture. doi:10.57071/240fut

Shorrock, S. (2022). Adaptive Imagination at Work in Health Care. In H. Laroche, C. Bieder, & J. Villena-López (Eds.), *Managing Future Challenges for Safety* (pp. 95-104). Cham: Springer. doi:https://doi.org/10.1007/978-3-031-07805-7_12

Siau, K., & Wang, W. (2020). Artificial intelligence (AI) ethics: ethics of AI and ethical AI. *Journal of Database Management, 31*(2), pp. 74-87.

Williams, J. (2020). Humanity, Technology, and Nature. *Icon, 25*(2), pp. 8-28.

Reproduction of this document

A PDF of this document (and other publications in the *Cahiers de la sécurité industrielle* collection) can be downloaded from FonCSI's website, www.foncsi.org.

FONCSI
Fondation pour une culture
de sécurité industrielle

Foundation for an Industrial Safety Culture
Public interest research foundation

http://www.foncsi.org/

6 allée Emile Monso – CS 22760
31077 Toulouse Cedex 4
France

Telephone: +33 (0) 532 093 770
X: @TheFonCSI
Email: contact@foncsi.org

ISSN 2100-3874

FONCSI
Fondation pour une culture
de sécurité industrielle

6 allée Émile Monso
ZAC du Palays - CS 22 760
31 077 Toulouse cedex 4

www.foncsi.org