Applied Artificial Intelligence: Risk Mitigation Matters

Norbert Jastroch

Applied Artificial Intelligence: Risk Mitigation Matters

Norbert Jastroch[0000-0002-4046-450X]
MET Communications, 61352 Bad Homburg, Germany
norbert.jastroch@metcommunications.de

Abstract. Digital technology is the main driver of the transformation process that is already on its way and expected to take up speed. Science and engineering are challenged to realize the significant innovation potential while keeping an eye on economic and societal sustainability. Research methodology in science as well as development practice in engineering provide well-established approaches to risk management and mitigation relating to this technological transformation. Artificial intelligence, though, brings in new features to address which this chapter shall help to deal with. As such we take into view machine learning, automated decision making and autonomous systems, and data utilization. We look upon characteristic risks within the application lifecycle, and on functional, societal, and cybersecurity risks. We derive suggestions for an approach to proactive risk management addressing the lifecycle of Artificial Intelligence applications. Along with a preparatory section on terminological clarification regarding artificial intelligence, data, and risk this paper is intended to build awareness of risk mitigation matters and set the scene for the development of accountable risk management approaches.

Keywords: Artificial Intelligence, Risk Management, Risk Mitigation, Data Analytics, Machine Learning, Automated Decision Making, Autonomous Systems

1 Introduction

In the course of the Covid19 pandemic we could make, among others, two observations worth reinforcing. Firstly, when a not well understood danger - like the Sars-CoV-2-virus - disseminates out of control, the individual, societal, and economic cost can be tremendous, and reactive risk management becomes difficult whereas essential. Secondly, the development of proactive risk mitigation means – like Covid19 vaccines – is possible as a break-through innovation within record time, while regulation and well-established scientific methodology are in place calling for due procedures to follow that are aiming at the control of related risks. Both these observations strongly support the concept of proactive risk mitigation and management when it comes to the technological transformation digitization and artificial intelligence are introducing into our lives. This holds true even more if we consider that, other than in the case of Covid19, this

transformation is not emerging out of the blue. We are well advised therefore not to neglect thorough reflection on appropriate ways to hedge potential risks that come along with the application of Artificial Intelligence (AI).

1.1 Digitization – Theoretical Considerations

Digitization, the numerical representation of analogue phenomena that can be observed – and measured – in the natural, technical, economic, or societal environment, can be taken as an approach to address the problem of complexity [15], i. e. to make complex things or situations manageable. In his 'theory of the digital society' Nassehi [15] distinguishes between the digital world and the analogue world. In the analogue world – the reality – data appear always in analogue context, continuously in time and space. In the digital world, however, any such context gets transformed into discrete data sets, where the discontinuity is being introduced via functional criteria, according to the specific purpose of observation, and realized via the measurement of related quantities. Result is the data world, the digital twin of reality[1]. While in the data world there is no generic connection of data, causation of and instruction for the analysis or processing of that data as well as their linking originate from the analogue world. This should be noted carefully as it implies a particular limitation to the perception of the concept of Artificial Intelligence, in the sense that any application of Artificial Intelligence in fact has a real-world, i. e. a human cause.

It is also worth noting that in his analysis of the digital society Nassehi takes the perspective of structural functionalism, where there are the two basic conditions of functioning and non-functioning. He analyzes the society as having been binary coded (e. g. science: true/false, law: right/wrong, economy: pay/no-pay) even well before the technologically understood binary pair of states functioning/non-functioning came into being. However, while the societal binary coding is widely accepted to depend on trust, technological binarity is considered being objective, hence more reliable, supporting the imagination as well as the requirement of full control. This is of importance for the way we address risk in the technological field, particularly with respect to Artificial Intelligence in the digital economy. Technically, an AI application must function according to its design, like a machine, where eventual malfunction can be detected easily. But inherent to Artificial Intelligence are the features of reasoning based upon the analysis of data, learning while in operation, and decision taking with full or partial autonomy. An AI application can work completely in line with its functional specification, the results, however, can be undesired, of doubtful validity, or even false. And that kind of, say 'misfunctioning', is anything else than easy to detect as it is being generated by highly complex algorithms. The need for appropriately extended approaches to the subject of risk in Artificial Intelligence thus becomes obvious.

1.2 Artificial Intelligence – Regulative Context

[1] Nassehi uses the term „Verdoppelung" (doubling), which might be misleading, as the data world in fact is not equally as rich as the analogue world, given that it is the result of a functional reduction by which the non-functional features are omitted. The data world in fact is a reductive data model of reality that is determined by the choice oft he functional perspective and the quality of data gathered.

The European Commission in their White Paper on Artificial Intelligence [5] apply an approach to risk in AI focusing on rights protection, safety, and liability[2]. This report distinguishes between AI systems of high and low risk, suggesting different regulative treatment of both categories. A quite detailed elaboration regarding the criteria for high-risk AI as well as the respective regulation of these has been published recently by the European Commission in their proposal of an Artificial Intelligence Act [3]. This proposed act comprises detailed requirements for high-risk AI systems including the obligation to implement risk management and data governance provisions. In any case, while regulation will be of large impact to the development of Artificial Intelligence, risk management beyond mere compliance with legislative rules has major value for the developers and operators of applied AI as it undoubtedly will affect the acceptance, hence successful uptake of such technology.

The view of policy-making institutions like the European Commission, whose purpose, among others, is to set the rules which frame the technological and economic progress for societal benefit, is kind of a view from the outside - where there is also the inside view of those who develop and run AI applications, and the third side view of those being affected, the users. No doubt it is of interest to any of these to see risk issues be thoroughly addressed in advance. Even more so if the suggested voluntary labeling of AI applications (cf. [4], [5]) gets implemented as a means of quality certification. The term coined in this regard is Trustworthy Artificial Intelligence.

Starting point for the generation of trust in Artificial Intelligence are the ethics guidelines elaborated by the High-Level Expert Group appointed by the European Commission, laid down in their report [9], [10]. Seven topics have been identified there and are taken up in [5] as 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
- Accountability

It is crucial, though, to get to operationalization of these requirements, as authors like Stix [16] have pointed out, claiming the necessity to develop actionable principles which are missing so far. Attempts to operationalize these requirements in the application of Artificial Intelligence should include appropriate procedures to assess, monitor, and mitigate risk. This calls for proactive risk management.

[2] "The main risks related to the use of Artificial Intelligence concern the application of rules designed to protect fundamental rights (including personal data and privacy protection and non-discrimination), as well as safety and liability-related issues" [5, p. 10].

2 Terminological Clarification

In [14] we briefly reflected on the meaning of the term Artificial Intelligence as compared to Human Intelligence. Simply put, the difference is the absence of mental processes in AI systems. They work with empirical or statistical data as their input, process these by algorithms, and generate correlation-based results of certain probability. However, there is a wide range of conceptual attributes used when talk is about Artificial Intelligence. The extremes are smart machines – whatever smart may mean there – at the low end, and AI systems with an own legal or even ethical status at the high end. Likewise, the terms data and information often get mixed up in this context, taking data as a kind of a natural resource while ignoring they come into existence only by functional reduction of reality. And similar confusion as to risk and its management can be observed in the public discourse. In this chapter we lay out the meaning of the terms as we apply them subsequently.

2.1 Artificial Intelligence is Unconscious Intelligence

In whatever form applications of Artificial Intelligence appear, they are utilizing software as their very kernel. They are programs, designed to simulate the analogue reality. The question is, do they reach as far as to make the difference between reality and simulation vanishing, which would justify the perception of AI systems being intelligent. Not few protagonists would subscribe to this, basically arguing that if no difference can be observed, there is no difference, or, a little weaker, the difference is no longer relevant. The opposite position is drawing on Artificial Intelligence as unconscious intelligence. Fuchs in [8] presents well elaborated principal considerations distinguishing human from artificial intelligence which in essence are summarized as persons are not programs, and programs are not persons [8, p. 35: 'Personen sind keine Programme. Programme sind keine Personen.']. Besides living vs. not-living, consciousness vs. unconsciousness, and subjectivity vs. no-subjectivity he names self-causation vs. external initiation and reflexivity (understanding) vs. input-output-transformation as the criteria which distinguish human from artificial intelligence.

The latter two are of practical relevance for applied Artificial Intelligence[3]. Although an AI system can, based on the data input from the exterior and processed by interior algorithms, well initiate some action making the system look as if it causes the action, this still is result of an input-output-transformation following the rules put in externally (and on functional purpose), which the system itself cannot reflect upon, but only execute. Think of autonomous vehicles as an example. The system thus has no responsibility for what it is executing. Responsibility, and in the operational sense liability, can only be ascribed to the ones having designed the rules and those having specified the purpose. That is where transparency, accountability, and human agency and oversight come into play, key requirements of the European Commission for trustworthy Artificial Intelligence.

[3] Their theoretical relevance would be pointed out if we spoke of Algorithmic Intelligence instead of Artificial Intelligence.

2.2 Data and Information

Reduction of complexity through digitization, Nassehi's doubling of the reality to generate the data world by means of functional reduction, digital simulation of analogue processes, or the instantiation of analogue phenomena in digital twins are all bound to generate data. Applied Artificial Intelligence, including data analytics, are making use of that data. Algorithmic processing then results in action, e. g. in robotics, or in decisions, e. g. in autonomous systems, or in derived data sets for subsequent processing, e. g. in embedded systems. The procedural pattern is input-output-transformation of data. The functional effect, though, is the generation and utilization of information, be that the movement of a robot, executing a decision, or detecting a correlation. Only this functional effect is to be held as valuable outcome of the data transformation. The mere transformation itself has no effect unless it feeds into a connection with the environment, be that physical, literal, or logical. To make sense, the data world needs to come into exchange with the real world[4] (cf. [15]).

The question now is whether this exchange is genuinely of technical nature, i. e. is nothing else than the transportation of data on the physical layer. Janich in [12] has presented a well elaborated investigation showing the problems of such a perception. He breaks the question down to the analysis of the conditions for successful communication, the pre-requisite of senseful interaction. These conditions comprise necessarily a process of mutual understanding, and this process is what turns data into information. He lets open whether it is possible to simulate this process technically. It seems to be clear, though, that such simulation could only work on a finite set of functional determinants. Through functional reduction (cf. [15]) such a finite set can be created. But without it, not even the simulation of mutual understanding is possible.

Resuming these considerations, we note that transporting data from one point to another does not imply equal information at both points exists. Thorough use of the terms data and information thus is recommended. In the context of Artificial Intelligence, taking data as the raw material should not mislead to taking it as some natural resource. Instead, informed approaches are needed to generate data from reality, e. g. functional reduction, and informed processes are necessary for their utilization, e. g. intentional algorithms. These informed approaches and processes are human made. A detailed investigation on these processes has been presented in [13].

In essence, applied Artificial Intelligence rests on human causation. And it implies human caused as well as inherent risks.

2.3 Risk Management

The management of risks is a well-established element in business, science, and technology. In the finance industry comprehensive regulative obligations for risk assessment and monitoring govern day-to-day operation. Research as well as application in medicine and pharmacology rest upon pertinent methodological regimes for the control

[4] An approach to deal with this in data business and applied AI is to make use of metadata providing contextual enrichment of raw data.

of undesired effects in diagnostics and treatment. Almost any technological field is ruled by norms, standards, and regulative precautions aiming to enable interoperability and the mitigation of risks connected with their appliances. All but surprising risk management becomes an issue for Artificial Intelligence, too. There are features of AI, however, which raise the need to look upon specifics to consider there. They concern the evaluation of risks difficult or even impossible to anticipate, since AI systems are explicitly expected to generate outcomes not fully determined.

The terms risk, risk management etc. have many specifications. For our purpose in this text those suggested by the International Standards Organization can be used as common ground. We are referring to the publicly available definitions in [11].

Risk there is the effect of uncertainty on objectives of an organization. It is subject to assessment by an overall process including the identification, analysis, and evaluation of risks. Coordinated activities to control and direct an organization regarding risk build risk management. This is governed by criteria of risk to be set in advance. These comprise tangible and intangible uncertainties, measurement of likelihood and consequences, and combination of multiple risks. The issue of the identification of risks, i. e. recognizing and describing them, is particularly relevant in the case of applied AI and poses problems, because often it can be done ex post only. The same applies to risk analysis, i. e. the comprehension of the nature, characteristic, and level of risks. Risk evaluation, finally, is based upon these preparative steps and leads to decide on implementing options for addressing risk, i. e. risk treatment (fig. 1).

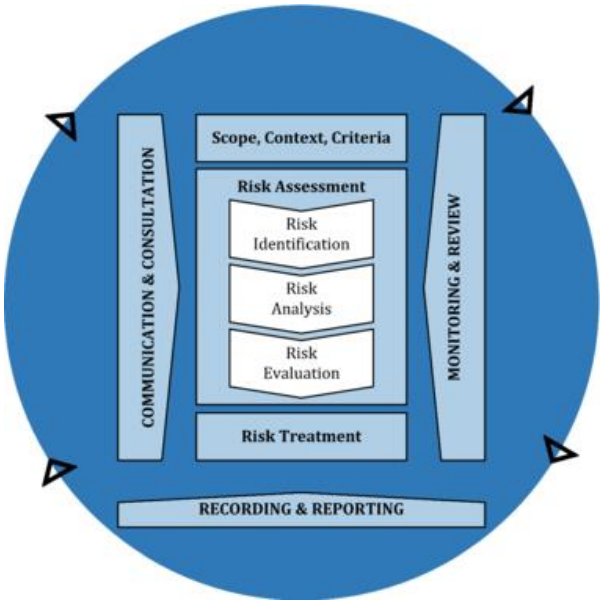

Figure 1[5] : Risk management process

[5] Source: https://www.iso.org/obp/graphics/std/iso_std_iso_31000_ed-2_v1_en/fig_4.png Reproduced with permission DIN Deutsches Institut für Normung e. V.

If there is uncertainty, planning becomes hard. The outcome of a technical application that is functionally closed, e. g. a machine, can in general be forecast even as to cases of malfunctioning. The uncertainty here comes down to possible malfunctioning. Empirical analysis of its operation will enable the assessment of the probability of malfunctioning. Applied AI systems, however, are considered functionally open to certain extent. Not only malfunctioning, but also what we have called misfunctioning are sources of uncertainty. This affects both the tasks of risk identification and analysis. An appropriate comprehension of the functional openness of AI applications therefore is needed. We address this in the following section.

3 Applied Artificial Intelligence

The distinction between applied AI and "smart" computing is not easy to make. In first order, it can be related to the level of openness, i. e. how far processing outcomes are not completely determined by processing algorithms. This may vary largely, depending on functional features implemented in an AI application such as Machine Learning, Automated Decision Making, and Data Utilization. Here is a brief characterization of these features.

3.1 Machine Learning

A machine is said to be learning when algorithms use empirical data sets for the detection of certain structures or relations within these data by an iterative process, where a subsequent iteration generates an improved representation of the pattern of structures/relations as compared to the previous iteration in a probabilistic sense, with improvement being calculated using stochastic models. When such a learning process is implemented in an application, it can operate (cf. Bartneck et al. [1])

- unsupervised, which basically is to explore clusters (like the position of specific marks in e. g. analysis of images), or
- supervised, meaning the iteration process uses a classifier for a target category generating a classification of the data set with respect to that category (think of facial recognition) - this classifier is to be trained by test data, or by
- reinforcement, where iteration is to proceed towards a desired goal (think of the movement of an autonomous vehicle) by feeding in and using additional data, e. g. delivered by sensors.

In essence, in machine learning proceeding to the next step depends on a probabilistic calculus. Selection is made by comparison according to higher similarity. It works in an environment that is functionally modelled and digitized. Modelling and digitization imply the used data sets being finite, even though highly complex, and providing reduced representation, as compared to reality. However, within this limitation, machine learning as part of an AI system enables a certain form of autonomous functioning or auto-control. The outcomes of such AI systems (the classification of a facial image according to specified attributes, the driving or flying course of an autonomous vehicle, etc.) are not necessarily straightforward. There is still room left for variance. They are open, although in a limited sense, including the possibility of misfunctioning.

3.2 Automated Decision Making and Autonomous Systems

Machine learning is an essential prerequisite for autonomous systems, as Wahlster [18] pointed out. In the industrial context, such systems combine Artificial Intelligence, involving machine learning for auto-controlled decision making, and interaction with the environment through sensors and actuators. Sensors observe the environment and measure relevant parameters to feed them into internal processing for the purpose of deciding upon subsequent operation. Actuators perform resulting actions.

Autonomous systems are expected to operate goal-oriented in a basically open, while limited, space of possible action. This is evident in the case of, e. g., an autopilot of an aircraft, or a robot in a manufacturing shopfloor. And they are bound to become part of AI applications in the medical sector, supporting e. g. diagnostics, or in the finance industry, e. g. as robo-advisors for investments. Common to all these are interaction with the exterior, governed by the flow of data, and the intentional establishment of an internal experience base built from empirical data gathered through operation. Quality of data thus becomes fundamental, as does the functional fit of the algorithms being employed.

Misfunctioning, in the sense of the generation of undesirable outcomes although there are no functional or data faults, is an inherent issue with any system that involves automated decision making. Wahlster [18] describes an approach to resolve this by introducing human supervision for non-standard events occurring during operation. Such semi-autonomous systems need well-organized mechanisms for the exchange and transfer of control between human agent and the automated system. They comply with the principle of human agency and oversight suggested by the European Commission as a key requirement to trustworthy Artificial Intelligence.

3.3 Data Utilization and Uncertainty

Data used as input to AI systems typically are subject to a processing cycle aiming to secure appropriate data quality. A useful illustration of data transformation along the lifecycle of an AI system is shown in figure 2.

Figure 2: Data transformation along AI Lifecycle development stages[6]

The basic steps in this process are initial data collection, data pre-processing including cleaning and transformation into a numerical data set, reduction to a functionally

[6] Source: [6]. Permission granted.

relevant data set, and training of the AI system according to a machine learning model (see 3.1 above) to generate the augmented data set.

As mentioned in 2.3, one expectation related to an Artificial Intelligence application is that it shall generate results which are reliable, as they are based upon objective data and imply no human bias. However, as pointed out in 3.1 above, the machine learning algorithms involved use probabilistic calculation. Judgements in the iterative learning process are made under uncertainty. The problems of judging under uncertainty have been established since long, going back to the work of Tversky and Kahneman [17]. In our context here, particular attention relates to the issues of representativeness and bias of input data used for judging. Tversky and Kahneman showed that they can substantially affect the reliability of judgement. This applies not only to judgement made by a human, but likewise to judgement as an inherent function in machine learning. For that reason, a trained AI application in general is subject to subsequent testing and evaluation.

It should be noted, however, that this leads to the question of quality and representativeness of testing resp. evaluation datasets. There are basically two approaches to address this issue, usage of empirical datasets, or standardized ones that get agreed in advance. While the first one is applicable ex post observing the operation of a system, the latter requires ex ante specification. The frequent claim of Artificial Intelligence is, while being based on objective data instead of subjectively estimated evaluation, to reduce or even avoid sources of flaw of this kind. It deserves thorough consideration yet in any case. Various examples have been reported where AI applications, e. g. for the recognition of specific patterns like mood or emotion in face images or videos, let show significant bias or incoherence when tested with real life datasets.

4 Risk Categories

Along with the entering of digital business models into the economic and societal ecosystem over the past years, implications showed up which gave reason to public debate about risks involved. It is not in the scope of this text to discuss these in detail. We note here, though, that digitization appears to bring about more than just technical risks, in particular societal risks and those related to cybersecurity. Artificial Intelligence will add specifics which we are concentrating on, as they are expected to affect the acceptance of respective applications, hence their business potential. The subsequent considerations are the result of initial reflections, but we do not claim they represent the complete picture. Our first attention is given to the lifecycle aspect of applied AI.

4.1 Risk along the Application Lifecycle

When the features of machine learning and/or autonomous operation are included in an application, it will become dependent on an empirical experience base introduced during design and implementation, while this is bound to develop during operation. The application is subject to changing their base of operation, meaning its mode of operation and even the outcomes may change. These changes and any risks they might imply cannot in any case be taken care of in advance “by design”.

This is not just hypothetical, in fact there are examples. It occurred when a chatbot for writing comments in a chatroom under auto-control, i. e. without being surveyed by a reflecting human actor, ran out of the range of what is being the accepted way of communicating publicly. While it was rather testing than a released application, it revealed, however, the limitation of auto-controlled machine learning. This raises the need for risk considerations being applied not only to the design of AI systems, but also to subsequent phases of its lifecycle. Risk responses should be thought of proactively for the development as well as the operational phase.

4.2 Technical Risks

Like any technical device, an AI application can function or not. Malfunction is an inherent risk. There are ways to its anticipation by implementing appropriate response means in advance. Engineering in general, and software engineering in particular, are used to cover this kind of risk management approach. Sources of risk can be false input, failures in output, and model flaws. The standard way is to address them during design and development by implementing precautional procedures. For the operational mode, maintenance and repair are the routines of choice. Taking it as an applied AI example close to the low end, the cases of fatal problems with automated operation control in aircraft in the recent past turned out to be a combination of false input and an inappropriate software procedure in the system.

In the case of applied AI of more general nature, an additional type of risk can be named, we call it misfunction. It can be result of deficient, while not false input, caused by e. g. insufficient data quality. A facial recognition application can produce false identification because the image data fed in are lacking precision or accuracy. Misfunction can also be the effect of an algorithmic flaw. Algorithms used may be designed such that even in error-free operation they can lead to biased evaluations in the end, although input data are not deficient in the sense above but bear characteristic features which in the process of algorithmic classification generate inappropriate tracks. Examples have been reported a. o. in the field of criminal disposition investigation.

As third type of risk we consider emerging phenomena. Think of the chatbot example mentioned above. Neither an issue of false or deficient data input or output, nor the result of an algorithmic or model flaw, that application of AI featuring machine learning and autonomous operation in a real-life environment produced outcome which was unforeseeable, while not intended. It just emerged. Obviously, this kind of risk is utmost difficult to deal with, during development as well as operation of an AI system. In terms of risk management, here is a major challenge.

4.3 Societal Risks

The easiest-to-grasp risks in this category are those related to regulation including legislation. Providers of AI applications must comply with regulative rules. Breaching them is an obvious business risk. Their potential violation is a risk for the society and individual users concerned. As there is no uniform regulative regime in place internationally, addressing this risk type requires much effort regarding the design of the system architecture during development and operation.

The technical risk of misfunction goes hand in hand with the societal risk of undesired implications. Frequently debated is the danger of bias in the context of AI applied to assess personal attributes. This may concern mainly individuals. Of broader societal impact yet are implications caused by model design which rests on personalization. Examples are echo-chambers in the field of social media or news services constraining the availability of information to a person, thus bearing the potential of e. g. political manipulation. The occurrence of such effects has been brought about by various studies. With respect to the commercial sector, the implications of personalization through filter-bubbles are widely discussed. Arguments draw on the non-transparency to users which increases the information asymmetry between suppliers and consumers and thus affects market efficiency.

A major societal risk category then relates to the connectedness e. g. of digitized infrastructures. Thinking of smart metering in the sector of public energy utilities, where intelligent management is expected to contribute to efficiency improvements to the supply and use of energy, the embedding of AI applications into complex grids of devices poses the risk of potential flaws propagating through the network, thus endangering resilience. As soon as such digital networks get equipped with embedded AI components, the functional openness of these components becomes a source of risk.

The fourth type of risk with societal relevance we take into view is the very basic phenomenon of ethical dilemma. In the context of the autonomous car a reoccurring subject of debate is if and how an artificial system can be designed to operate when it comes to decide between alternatives that both result in damage to a human. While there is good reason to doubt that a satisfying general resolution to this dilemma can be found, the need to address it in practical situations cannot be denied. Apparently, it shows up as a problem with any autonomously operating physical AI application. Catastrophic external events, though not very likely, can happen, and the way an artificial system is designed to deal with them will be a major issue of acceptance[7].

Within the purpose of this text, we put focus on these four risk types. A wider political view is provided in [7]. That report is supposed to support further elaboration of the societal risk category.

4.4 Cybersecurity

Cybersecurity has been gaining increasing relevance in the past decades. Adversarial attacks on the IT systems of public authorities, companies, institutions in science or public infrastructure are being reported in ever growing numbers. Criminal, political, or economic interest appear to be the drivers. With Artificial Intelligence on the rise, it is safe to expect growth will continue. The threat is twofold. AI technology can provide novel instruments for intentional attack. And AI systems in operation will be of

[7] Machine ethicists have been discussing that kind of problems since long. In his comprehensive reflections on Limitations and Risks of Machine Ethics, Brundage [2] in essence concludes that „…there is more to successful ethical behaviour than having a good algorithm“ (ibd., p. 268). He draws upon, a. o., two lines of argument: knowledge limitation (real-life ethical decision-making is heuristic, because it is impossible to consider the whole space of possibilities) and computational limitation (computing the implications in a given situation can be intractable because of the number of agents, the time horizon, or the actions involved).

particularly high interest to attack as they are largely non-transparent and, at the same time, of high economic or strategic value. This risk category therefore deserves to be considered separately, although there is much overlapping with the technical and the societal categories. Even more so, as there is still a significant lack of awareness.

The European Union Agency for Cybersecurity ENISA recently presented their report on AI Cybersecurity Challenges [6]. This illustration of the threat landscape for Artificial Intelligence can be taken as a point of reference for risk management activities in applied AI. Within the threat modelling methodology provided there the identification of risks is key, in terms of threat and vulnerability. Our focus in this text is on three risk types: attack, accident, and outage. Other than most of the risk types considered in 4.2 and 4.3, they typically have external causes. Any approach to risk mitigation therefore must include thorough reflection upon the environment an AI application shall operate in.

Further to the considerations in [6], we draw attention to a few specific aspects. With respect to attack by external action or accident by environmental impact especially the interaction of an AI system with the exterior, be it logical or physical, is of concern. Logical program interfaces as well as physical sensors and actuators need to be included in the process of risk identification (and of course then in risk management activities). In fact, this must be done with bi-directional perspective, i. e. inbound and outbound. Not only can a threat be imported into an AI application, but also can one be unintendedly exported to its environment, as far as the application operates autonomously or involves automated decision making.

Likewise, the potential impact of an autonomous system on its environment in the case of physical failure like loss of power can become dangerous. In a system of coupled autonomous units the management of which rests upon communication between the units, the outage of one can result in severe problems for the entire system. Such kind of risk is known in the context of connected systems in general, but it is of significantly higher impact when there are autonomous units operating.

5 Toward Actionable Risk Management

The reflections described so far make us suggest a path to the development of actionable risk management in applied AI which starts from the risk identification matrix in figure 3.

As a recommendation, the development and operation of an AI application should in general comprise an exercise to build awareness of the risks involved. And especially if that risks are estimated to be of high likelihood or represent a large damage potential. Categories and types of risks shown in the table should be assessed as to their relevance in the phase of system development or operation or both as first step. Then the specification of the risks as such – events to be avoided – relating to their source can be defined more specifically. Actors involved can be identified, as well as logical or physical procedures, and coupling interfaces. Final step should deal with exploring possible

responses and their introduction into the development process, resp. their preparation for the operative lifecycle.

Risk category	Risk type	Source	Response Develop-ment	Response Operation
Technical				
	Malfunction	False input or output failure		
		Model flaw		
	Misfunction	Deficient input		
		Algorithm flaw		
	Emerging phenomena	Model design		
Societal				
	Regulative breach	System architecture		
	Undesired implications	Model design		
	Flaw propagation	Embeddedness		
	Ethical dilemma	Catastrophic event		
Cybersecurity				
	Attack	External action		
	Accident	Environmental impact		
	Outage	Physical failure		

Figure 3: Risk identification matrix.

Apparently, this approach is neither exhaustive nor complete in fitting with the manifold of AI applications that may come up in the future. It is rather the utilization of what has been put up during past efforts to realize the potential of AI in science, technology, economy, and administration. Yet we are convinced that there is significant rationale for the introduction of proactive risk mitigation to this field beyond regulative obligation. It can help to direct future progress toward what is beneficial for business, society, and individuals. It can become a decisive enabler for the building of trust in AI. And it will not downgrade the innovative potential AI can bring to the digital transformation as it supports the acceptance of applied AI. Like with any transformation, scepticism and even resistance are being encountered and must be overcome. Openminded consideration of chances and risks, of opportunities and constraints pave the way to affirmative uptake. And avoidance of potential damage and the related cost at last contribute to the profitability of businesses involved.

The three pillars of the digital transformation – Data, Artificial Intelligence, and Robotics – are subject to risk considerations each to a different extent and under specific perspectives. This will surely extend further with increasing implementation of applied Artificial Intelligence. The realization of its potential rests upon the technical excellence employed. Proactive risk mitigation can significantly contribute to this excellence. This paper describes a starting point for the development of actionable risk management. Further work shall be dedicated to the empirical elaboration and exemplification of the approach. This will include the collection and investigation of AI use cases from different fields of application. Based upon these, a Risk Management Framework will be suggested by filling the gaps in fig. 3 and further operationalizing the approach. Furthermore, it is intended to establish a Risk Management Library and in the longer run a Risk Management Experience base that can be used as a reference for the development of new systems of applied Artificial Intelligence.

References

1. Bartneck, C., Lütge, C., Wagner, A., Welsh, S.: What is AI? In: An Introduction to Ethics in Robotics and AI. Springer Briefs in Ethics. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-51110-4_2
2. Brundage, M.: Limitations and Risks of Machine Ethics. *Journal of Experimental and Theoretical Artificial Intelligence* **26**(3*)*, pp. 355 – 372 (2014). https://doi.org/10.1080/0952813X.2014.895108
3. European Commission: Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence. Brussels (2021/04/21)
4. European Commission: Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics. Brussels (2020)
5. European Commission: White Paper on Artificial Intelligence. A European approach to excellence and trust. Brussels (2020)
6. Malatras, A., Dede, G. (eds.): AI Cybersecurity Challenges – Threat Landscape for Artificial Intelligence. Report of the European Union Agency for Cybersecurity (ENISA) (2020). DOI 10.2824/238222
7. MSI-NET: Algorithms and Human Rights. Council of Europe Study. Strasbourg (2018)
8. Fuchs, Th.: *Verteidigung des Menschen – Grundfragen einer verkörperten Anthropologie.* Suhrkamp, Berlin (2020)
9. High Level Expert Group on Artificial Intelligence (AI HLEG): Ethics Guidelines for Trustworthy AI. European Commission, Brussels (2019)
10. High Level Expert Group on Artificial Intelligence (AI HLEG): Policy and Investment Recommendations for Trustworthy AI. European Commission, Brussels (2019)
11. International Standards Organization: ISO 31000:2018(en): Risk management — Guidelines, accessed 03/18/2021.
12. Janich, P.: *Was ist Information?* Suhrkamp, Frankfurt a. M. (2006)
13. Jastroch, N.: The Information Age – Remarks on Basics and Implications. In: *Multimedia – Innovating Telecommunications '99*. Friedrichsdorf (1999). ISBN 3-00-004825-1
14. Jastroch, N.: Trusted Artificial Intelligence: On the Use of Private Data. In: Nyffenegger, F., Rios, J., Rivest, L., Bouras, A. (eds.): Product Lifecycle Management Enabling

Smart X. IFIP AICT, Vol. 594, pp. 659 – 670. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-62807-9_52
15. Nassehi, A.: *Muster – Theorie der digitalen Gesellschaft.* C. H. Beck, München (2019)
16. Stix, C.: Actionable Principles for Artificial Intelligence Policy: Three Pathways. *Sci Eng Ethics* **27**, 15 (2021). https://doi.org/10.1007/s11948-020-00277-3
17. Tversky, A., Kahneman, D.: Judgement under Uncertainty: Heuristics and Biases. *Science* **185:** 1124 – 1131 (1974)
18. Wahlster, W.: Künstliche Intelligenz als Grundlage autonomer Systeme. *Informatik Spektrum* **40**, 409 – 418 (2017). https://doi.org/10.1007/s00287-017-1049-y