Artificial Intelligence and/as Risk

Isabel Kusche

Artificial Intelligence (AI) is a notion that has stimulated both the imagination and the engineering efforts of generations. Currently, however, AI predominantly refers to an ensemble of technologies and applications that digitally compute very large and heterogeneous data sets in a way that seems to mimic human intelligence, although it actually works very differently (Datenethikkommission 2019). This ensemble is a subset of technologies based on sophisticated algorithms, i.e. sets of instructions, given to and executed by computers (Buiten 2019: 49). Algorithms are associated with intelligence when they are complex enough to learn, i.e. to modify their own programming in reaction to data (Buiten 2019: 49–50; Datenethikkommission 2019: 273). Such machine-learning algorithms encompass a number of classes, among them artificial neural networks. Although the concept for this class of algorithms was inspired by the structure of neural networks in the brain, they do not actually model a brain or even a part of it. With regard to specific, clearly demarcated tasks related to the identification of patterns in large amounts of data, they perform much better than the human brain thanks to their implementation of advanced statistics to find patterns in data.

Artificial neural networks lend themselves to a great variety of applications, from natural-language processing to autonomous driving. Their utility is based on predictions that are derived from the patterns they find in large amounts of data. Consequently, artificial neural networks are expected to be superior instruments in predicting risks. In mid-August 2020, a search on Google Scholar for the combined keywords "risk prediction" and "machine learning" returned about 19,800 results. A scan of the first 150 results, sorted by relevance as determined by the Google algorithm, indicates a great variety of risks that machine learning algorithms are expected to predict better than previous methods. Various applications in medicine are

especially prominent, ranging from the risk prediction of specific diseases to the risk of particularly severe progressions of a disease and mortality. Other risks addressed with machine learning algorithms include aviation incidents, city traffic accidents, delays in construction projects, problems with supply chains, urban flood risks, driving behavior, criminal behavior, bankruptcy and suicide attempts. Substituting "artificial intelligence" for "machine learning" in the search leads to fewer results (around 10,600) but similar applications.

Most of these applications seem completely benign and extremely useful. Moreover, they seem to carry forward existing practices of risk calculation to render an uncertain future more predictable. At the heart of risk calculation is a promise of tamed contingency. Statistical techniques deal with the occurrence of events. In this context, prediction is not necessarily focused on avoiding future events based on knowledge of past events, but on managing their consequences (Aradau/Blanke 2017: 375–78). Prediction does not focus on concrete situations deemed undesirable or even dangerous; rather, it is about identifying a complex of risk factors, each indicating an increased likelihood that an undesirable event might occur (Makropoulos 1990: 421). In the case of work accidents, for example, risk calculation is not about preventing any specific accident, but about managing the phenomenon of infrequent but regularly occurring accidents. This phenomenon becomes observable only through statistics. The key to managing it is insurance, which is based on predicting the frequency of accidents (O'Malley 2009: 32–33). The specific conditions of insurance may take a number of factors into account, such as employees' skill level, work-specific training or the quality and maintenance of technological equipment. Importantly, these risk factors are abstract and detached from any particular individual who might be in danger or endanger others (Castel 1991). Similarly, crime statistics render an uneven distribution of crime observable, in terms of both socio-demographics of perpetrators and victims and geographical distribution of crime scenes. The knowledge produced in this way then may inform specific interventions directed towards the collectives identified as being at risk.

Risk analysis has not been limited to the prediction of events for which statistical records exist or are feasible in principle. Risk models for extremely rare events such as nuclear power plant failures cannot draw on statistical data. They replace these with expert judgements, which assign probabilities to various events based on past research and experience. Since ma-

chine learning needs large amounts of training data, AI applications would seem less promising for the risk analysis of rare technological or natural hazards. Researchers still expect them to be better in terms of predictive accuracy and/or computational burden, provided they take into account the challenges related to the nature of available data (Guikema 2020). Yet the greatest impact of machine learning on the management of contingency is expected in fields of application where statistics have played a large role in the past.

From the point of view of computer science, machine-learning algorithms are, like statistical procedures, probabilistic. There are also social scientists who see AI applications mainly as a continuation and intensification of contingency management based on statistical knowledge. Yet others point out that the developments actually amount to a break with this governance tradition. Rouvroy and Berns (2013) stress the amassment of data that are processed in search of correlations, resulting in data profiles. Profiles resemble combinations of risk factors in that they constitute knowledge that abstracts from the individual and enables predictions about individuals based on this abstraction. However, de Laat (2019: 322) emphasizes that the combination of massive amounts of data and AI re-introduces a focus on the individual. Conceptually, it would imply the end of insurance, which manages the uncertainty of the individual case (accident yes/no, illness yes/no, joblessness yes/no) by transforming it into a risk that characterizes a collective. Statistics does not reveal individualized risks, but AI is supposed to be able to do just that (Barry/Charpentier 2020). Provided machine-learning algorithms have access to sufficient amounts of individual-level data, they promise to predict for the individual whether they will or will not have an accident, lose their job or commit a crime.

Such potential does not necessarily mean that the technology will actually be used in that way. Firstly, individual-level data are more readily available in some cases than in others. In the case of insurance, for example, it is easier to gather large amounts of individual-level data deemed pertinent to the risk of having a car accident than to the risk of developing cancer. Driving behavior is a clearly circumscribed activity, many aspects of which can be measured by telematics devices built into cars themselves. By contrast, health-related behavior is much more diverse and less easily captured, although people who opt to use health apps also provide a plethora of individual-level information.

Secondly, established organizational structures and practices may block or dilute the application of AI for individual risk assessment. Barry and

Charpentier (2020) demonstrate that car insurance products have (so far) changed much less than the availability of relevant behavioral data and machine learning might suggest. Although new variables based on such data are added to existing classifications of drivers, which become more refined as a result, classification itself is maintained. In other words, risks are not individualized. This is perhaps not surprising, considering that fully-individualized risk would amount to predicting individual accidents with high certainty, thus undermining the fundamental concept of insurance. At the very least it would lead to very high and thus unaffordable insurance rates for high-risk individuals and challenge the business model of insurance companies (ibid: 9).

Thirdly, however, potential applications may not be implemented due to profound concerns about their possible wider consequences. That means AI itself would be considered (too) risky. It is this perspective on AI applications as a possible risk that the following sections will discuss. Their aim is to explore how the prominence of thinking in terms of risks does not only feed into the promise of AI but also informs and delimits reflections about the (un-)desirability of AI applications.

Risky AI

The 2020 European Commission's White Paper on Artificial Intelligence states that AI "can also do harm. This harm might be both material (safety and health of individuals, including loss of life, damage to property) and immaterial (loss of privacy, limitations to the right of freedom of expression, human dignity, discrimination for instance in access to employment), and can relate to a wide variety of risks" (European Commission 2020: 10). In recent years, strands of research and public activism have converged towards highlighting ways in which machine-learning applications pose risks – to individuals, certain groups or even democratic society as we know it. At a time when advances in machine learning constantly seem to open up new possible applications, the uncertainty about its social implications has grown.

Anxieties and fears are not focused on safety with regard to possible technological disasters, like the 1984 chemical release at Bhopal or the 1986 Chernobyl nuclear catastrophe, or the possible normalcy of accidents (Perrow 1984). Nor do they concern unforeseen impacts on biological systems and

the well-being of individuals or ecologies when new substances or modified organisms are released into the environment (Wynne 2001). This does not mean that there are no safety concerns when it comes to applications of AI. The possibility of malfunction is in fact an ongoing concern for engineers and computer scientists using AI applications. For example, the so-called deep learning of artificial neural networks is about identifying patterns in training data and then using the trained model to classify unfamiliar data. However, the way in which a model arrives at its classifications is completely different from human cognition and thus not (easily) interpretable. This opens up the possibility of misclassification in cases that a human would find self-evident. When such errors occur, engineers face particular difficulties in understanding and controlling them because they are the result of a type of data-processing that defies human logic. Moreover, small changes to input data that humans are unable to recognize may alter the output completely (Campolo/Crawford 2020). Malicious attackers could make use of such adversarial examples to deliberately cause malfunctions, for example in the computer vision and collision avoidance systems of autonomous cars, resulting in accidents and loss of lives (Garfinkel 2017).

Accordingly, the European Commission's White Paper on Artificial Intelligence demands clear safety provisions in connection with AI applications, which are deemed necessary to protect not only individuals from harm but also businesses from legal uncertainty about liability (European Commission 2020: 12). This poses particular challenges, not only because autonomous behavior is the promise and selling point of certain AI systems. Most AI systems also use at least some components and software code not created by the AI developers themselves but drawn instead from open-source software libraries. The complex interactions of different software and hardware components can render the detection of failures and malfunctions as well as their attribution to specific components and thus to producers or creators extremely difficult (Scherer 2015: 369–73).

Yet different from cases like nuclear energy or genetic engineering, there are concerns apart from health hazards and the survival of human beings or other living creatures. Wider societal implications of AI ensuing from applications that may operate perfectly in terms of reliability and safety take center stage. The European Commission's White Paper is explicit in this regard when it declares "AI based on European values and rules" (European Commission 2020: 3) as the goal of a common European approach to AI. That implies the technological feasibility of AI based on other, or possi-

bly no, values. The explicit reference to values acknowledges that "AI is not merely another utility that needs to be regulated only once it is mature; it is a powerful force that is reshaping our lives, our interactions, and our environments" (Cath et al. 2018: 508).

The recognition that values are at stake when AI is designed and applied is reflected in the fact that the most prominent symptom of uncertainty about its societal consequences is the call for AI ethics and an abundance of science- or industry-led ethical guidelines (Hagendorff 2020). Calls for ethical, trustworthy, responsible or beneficial AI all appeal to ethical principles to delineate what AI applications should do and what they should not do (Jobin et al. 2019: 392–94). Comparing AI to other technologies perceived as risky, the ubiquitous recourse to ethics may come as a surprise. Critics see it as an attempt to avoid government regulation and as lacking any significant impact on the individuals and companies developing and applying AI (Greene et al. 2019; Hagendorff 2020). Yet it may also indicate that the risks of AI are seen as different from and as less well-defined than the risks posed by previous technologies.

Jobin et al. (2019) identify transparency, justice and fairness, responsibility and accountability as well as privacy as the most common ethical values and principles to which private companies, governments, academic and professional bodies and intergovernmental organizations refer in their documents on AI ethics. Non-maleficence, including the general call for safety and security, is also an ethical principle towards which many statements and guidelines on AI ethics converge (ibid: 394). However, this apparent convergence is contradicted by the many differences in how these principles are interpreted in the documents and which measures are proposed to realize them (ibid: 396). More precisely, questions of implementation and oversight are typically not even addressed. In another review of AI ethics guidelines, Hagendorff (2020) notes that technical solutions are increasingly available to satisfy certain interpretations of principles such as accountability, privacy and fairness. 'Explainable AI', 'differential privacy' and tools for bias mitigation are all technical approaches to render AI applications compatible with corresponding values. Yet other values, for example sustainability or the integrity of democratic political processes, are conspicuously absent from AI ethics guidelines, possibly because technological fixes seem out of reach (ibid: 104–5).

Greene et al. (2019) focus on seven high-profile statements on AI ethics and their common underlying assumptions that connect values, ethics and

technologies. They highlight that these statements frame ethical design and oversight as matters for primarily technical, and secondarily legal, experts (ibid: 14). Although the statements recognize the role of values in designing AI applications, they never question the development of such applications in principle, thus implying a technological determinism that inevitably leads to more and more AI (ibid: 15–16). Uncertainty about particular consequences of AI thus contrasts with certainty about the necessity and inevitability of AI in general. The overall narrative regarding possible negative effects of AI that "[p]oor ethics lead to bad designs, which produce harmful outcomes" (ibid: 17) also suggests that ethics guidelines can reduce (or even eliminate) the uncertainty about consequences of AI.

The distinction between ethical and unethical AI is thus linked to the possibility of risks, but the exact nature of this link remains unclear. There are a number of "potential risks, such as opaque decision-making, gender-based or other kinds of discrimination, intrusion in our private lives or being used for criminal purposes" (European Commission 2020: 1) that AI entails. Obviously, unethical AI would ignore these risks and therefore likely lead to rights violations. It is less clear, however, whether ethical AI is expected to eliminate, minimize or just somewhat reduce these risks.

Again, the comparison to concerns about safety is instructive. Safety is itself a value that is somewhat more specified than non-maleficence as a guiding ethical principle. Yet there is no completely safe technology. Accidents happen, which was precisely the reason why the instrument of insurance, based on statistical calculation of risks, developed (O'Malley 2009). Without the possibility to calculate risks or rationally assess them in some other way (Collier 2008), the notion of risk is hardly more than an empty signifier for the many things that can go wrong.

In sum, expectations of ethics appear to be very high when it comes to dealing with the risks of AI. Disappointment seems almost inevitable. Ethics is presented as a soft version of legal regulation although it cannot ensure compliance and consequently is no functional equivalent to law (Rességuier/Rodrigues 2020). In line with the focus on ethical principles, the implied notion of AI risks is very broad. It denotes potential for the realization of unwanted, negative consequences (Tierney 1999: 217), which remain unspecified but whose negativity would derive from a violation of values.

Quantification and Standardization

Any attempt at a risk-based regulation of AI applications would require a level of concretization that at least renders risk estimates feasible. Considering the close link between the concept of risk, statistical thinking and thus quantification, it is perhaps no surprise that one approach to rendering ethical principles concrete and actionable is the search for quantifiable measures of how well AI systems perform in this regard. However, artificial neural networks already pose challenges to attempts to quantify the uncertainty with which they make their predictions (Begoli et al. 2019). There are uncertainties regarding the training data, i.e. how well they represent the phenomenon in question and how accurate the data have been labelled. There are further uncertainties about the chosen AI model, namely how well it fits the purpose, in terms of performance characteristics and model limitations. There are also uncertainties with regard to what appropriate tests of the trained model should look like (Begoli et al. 2019: 21; Buiten 2019: 51–53). While interpretability of models would minimize many of these epistemic uncertainties, there is a trade-off with performance in terms of predictive accuracy.

These difficulties notwithstanding, computer scientists have made various proposals to not only measure performance in the narrow sense, but also characteristics such as fairness, i.e. the extent to which models produce equal, non-discriminatory outcomes for different groups. The more technical of such proposals do not even address the huge gap between fairness as an ethical principle or abstract value and the proposed quantitative measure (e.g. Speicher et al. 2018). Others stress that judgements about fairness are always context-dependent trade-offs with other values and principles; they limit the possible benefit of fairness measurement to providing a baseline of quantitative evidence that may enter into the deliberation of courts about specific cases (Wachter et al. 2020). Normative questions are thereby delegated to the judiciary and the legal framework already in place.

However, if the notion of immaterial risks is taken seriously, such risks might extend beyond provisions within the existing legal framework, which implies that the judiciary might also be incapable of bridging the gap between normative principle and the minimization or mitigation of harm. In fact, it is the anticipation of this possibility that triggered the demand for ethical guidelines in the first place. A particularly ambitious attempt towards a concretization of such guidelines comes from the AI Ethics Im-

pact Group (2020), led by the Bertelsmann Stiftung and VDE (Verband der Elektrotechnik Elektronik Informationstechnik). Their approach to an operationalization of AI ethics is two-pronged. On the one hand, they propose how to specify values and translate them into observables, resulting in a context-independent AI Ethics label for AI systems. On the other hand, they propose classifying the application context of AI systems in terms of a risk matrix, the idea being that riskier application contexts require higher ethical standards.

The proposal highlights both the rationale and the limits of addressing uncertainty about consequences of AI in terms of values and risks. The context-independent ethics rating is based on an incremental translation of abstract values into criteria, indicators and finally observables. It requires that values are first defined, then rendered measurable and balanceable. As the proposal emphasizes, it is "not possible to logically deduce criteria or indicators from values but to argue for these in deliberative processes" (AI Ethics Impact Group 2020: 17). The proposal abstains from specifying the participants in such deliberations but suggest that these "normative decisions should be made in a scientific and technically informed context" (ibid: 16). The deliberations are supposed to result in a clear picture of "an AI system's ethical characteristics" (ibid: 31).

For the complementary classification of application contexts, the proposal foresees regulators as the decision-makers (ibid: 39), at the same time calling for "the participation of stakeholders with a broad, interdisciplinary perspective" (ibid: 37). The authors propose a risk matrix that distinguishes classes of applications depending on how much potential harm an AI system could cause and how much those negatively affected depend on the AI system. As the proposal readily admits, the two dimensions are hardly separate since "correlations between the dimensions arise depending on the weight of individual aspects in the internal composition" (ibid: 37); furthermore, both dimensions demand value judgements and in particular the resolution of value conflicts. Moreover, there may be thresholds beyond which even very low-probability risks are deemed unacceptable (ibid: 37).

In sum, the notion of risk underlying this proposal has little to do with risk as a calculative technique. The repeated call for stakeholder involvement and the emphasis on deliberation and values indicate that the potential harm of AI applications cannot be predicted based on calculations. Yet the ambition is to arrive at a somewhat standardized assessment of AI that goes

beyond case-by-case decisions about which applications are acceptable and which are not.

Lesson Learnt?

The emphasis on the role of values and the need for ethical principles amounts to a dream come true for many social scientists focusing on risk assessment and risk management. It seems that the approach to risks of AI has been greatly informed by analyses of the shortcomings that troubled and derailed earlier attempts to assess and manage technological risks. About 20 years ago, Wynne (2001: 446) was still showing evidence of a categorical distinction between risk concerns and ethical concerns in relation to genetically-modified organisms (GMOs). While scientific expertise was supposed to deal with the former, vaguely informed and more emotional lay publics were prone to focus on the latter, in keeping with institutionalized expectations about the separation of risk and ethics. Around the same time, Jasanoff (1999: 140–45) pointed out how formal risk assessment is a particular type of expert knowledge, with taken-for-granted yet contingent assumptions about causation, agency and uncertainty. According to her, it understands causation of harms as mechanistic, locates the sources of risk in inanimate objects and renders the cultural and political origins of uncertainty invisible by translating it into formal quantitative language. Jasanoff at least detected the emergence of a different conception of risk-based regulation that would conjoin scientific analysis with political deliberation, encourage feedbacks, recursion and revision based on experience, and acknowledge that the regulatory process is ultimately about decision-making and not science (ibid: 149–50).

At first sight, approaches to AI seem to address these concerns about risk management. The proliferation of ethics guidelines that are initiated and developed by computer scientists and supported by tech companies suggests that an emphasis on ethics is no longer disparaged as the emotional reaction of laypeople. The proposal of the AI Ethics Impact Group, for example, appears to be in line with an understanding of technological risks that recognizes the central role of ethics, values and political deliberation when it comes to the regulation of AI applications. Moreover, the German Data Ethics Commission points out that risks do not only originate from the technological design of an application but also from human decisions in

using the technology (Datenethikkommission 2019: 167). It also notes that the effects of some AI applications may be unacceptable, necessitating a ban.

Yet awareness of the political and cultural context in which risk-based regulation inevitably takes place remains superficial. Firstly, the consideration of ethics is reframed as a task for experts and thus transformed into a question of finding the right experts for it. This is, for example, indicated by the call for "the participation of stakeholders with a broad, interdisciplinary perspective" (AI Ethics Impact Group 2020: 37), which suggests that it will be experts after all who bring different views about the operationalization and prioritization of values to the table. The consideration of ethics thus primarily manifests as the consultation of legal scholars, theologians, ethicists and experts in data protection, who collaborate with computer scientists and industry representatives, as, for example, in the German Data Ethics Commission (Bundesministerium des Innern, für Heimat und Bau 2019).

Secondly, the concession that risks are not only located in technological objects but can arise from human decisions in using the technology is of little consequence when the focus of possible regulation remains on "an AI system's ethical characteristics" (AI Ethics Impact Group 2020: 31). The idea of an ethics label in particular attempts to locate values context-independently in technical objects. It implies that it is possible to decontextualize and objectivize values. As well as the appeal to ethical expertise, this suggests that it is possible to get this kind of assessment 'right'. Appropriate ethical assessments are apparently those that include diverse but well-informed stakeholders who deliberate about the operationalization of values until they find one that all parties involved can agree with.

Thirdly, the political and cultural origins of risk and uncertainty are again blurred as a result. The decontextualized nature of expert ethical assessments inevitably ignores the possibility that values, and in particular their ordering in cases of conflicting values, may vary both spatially and temporally. Admittedly, the European Commission's White Paper calls for "AI based on European values and rules" (European Commission 2020: 3). Yet its overall focus is on working towards global championship and leadership in AI applications. This implies that both European values and solutions to value conflicts are (at least potentially) universal. By contrast, sociological analyses of values and their prioritization in situations of conflict stress that orderings of values are plural, temporal and adaptable to political exigencies (Luhmann 1962; Boltanski/Thévenot 2006; Kusche 2021). The trend towards attempting to quantify specific values veils the contextual nature of value

judgements even more and suggests a distinction between ethics experts, whose function is to remind everyone of abstract principles, and technical experts, who propose and implement appropriate performance measures.

The attention paid to ethical concerns in relation to AI may thus be commendable compared to a risk regulation that focuses on possible future harms without acknowledging that values inevitably enter the equation. Yet it sidesteps the full implications that a thorough consideration of values would have, especially in view of the notion of immaterial risks.

Risk, Decision-making and Non-knowledge

Although AI ethics guidelines fail to regulate design and business decisions regarding AI applications, their proliferation indicates the recognition that risks and decisions are closely connected. Value conflicts do not just disappear; they demand decisions. Moreover, the concretization of values is a matter of making decisions in the first place, and talking about deliberation instead indicates primarily a preference for involving many decision-makers instead of only a few. In the absence of deducible criteria or indicators for specific values, any decision about such criteria is itself uncertain. Adhering to such criteria could ultimately lead to the violation of values and to corresponding negative consequences – it is, inevitably, itself a risk. If such criteria are quantified, the role that decisions played in arriving at the respective measures is rendered invisible; by contrast, an emphasis on deliberation is at least a reminder that decision-making is unavoidable. Yet the implications of the necessity of risky decisions when dealing with risks of AI only become clear once they are considered as only one instance of the constitutive character that risk has for modern society.

Risk does not denote an objectively measurable hazard, but is a way to deal with contingency, attributing uncertain negative events in the future to decisions, as opposed to misfortune, God's will, laws of nature or any other external cause. The probabilistic approach to risk, common to classical statistics and advanced machine-learning algorithms, is a symptom of how ubiquitous the attribution of future events to contingent decisions is. Yet for the same reason, as Luhmann (1991: 28–31) argues, the opposite of risk is not safety but danger, that is a possible future negative event attributed externally and not to one's own decision-making. By contrast, safety is something to strive for, be it in the face of risk or danger, but not in the

sense of a specific goal that can be reached. It is a value (ibid: 28) that may orient decisions in the face of an unknown future, which always entails the possibility of events that one would prefer to avoid.

The distinction between risk and danger highlights the difference that it makes whether negative future events are attributed internally, that is to one's own decisions, or externally. Since the attribution to external causes includes the attribution to decisions others have made, the distinction risk/danger is closely connected to another distinction, namely that between decision-makers and those affected by decisions (ibid: 111–19). Risks run by decision-makers can turn into dangers for those who experience consequences without having been involved in the respective decisions. When dangers are deemed considerable and can be attributed to risks taken by others, a conflict between decision-makers and those affected becomes likely (Japp/Kusche 2008: 90–92). The latter may refuse to accept what they observe as danger and turn against those seen as responsible for it. Excluded from the decision-making, they can take recourse to protesting against the danger and against the decision-makers to whom it is attributed.

The introduction and spread of new technologies are, although not the only case, a very prominent case in which attributions to risk and danger have often fueled conflicts between decision-makers and those affected. Policy-makers, companies and business associations have become increasingly aware that broad popular resistance against technologies can pose both political and economic problems. Due to its capacity to make collectively binding decisions, the political system attracts a plethora of expectations. Resistance against technologies will almost inevitably turn into demands for regulation or even bans. Based on past experience, with the prolonged protests against nuclear power being probably the most impactful in Germany, political actors can anticipate the necessity to get involved and address the question of potential harms. Moreover, in the case of AI national governments and the European Commission even take explicit responsibility for the various effects that these applications may have when they actively promote their adoption and further development in the interest of competitive economies (European Commission 2020). This is politically risky in the sense that it would seem to create clear targets for blame in case AI applications turn out to have consequences deemed negative by significant numbers of voters. Similarly, businesses wishing to develop and sell or use a new technology can anticipate not only legal problems with liability in case of possible harms but also

threats to revenues when the respective technology meets broad resistance from clients and consumers. As political consumers, the latter may choose to prioritize ethical concerns even when they are not directly affected by negative effects of a particular product or service (Brenton 2013).

Against this backdrop, both political actors and businesses can anticipate that they will be seen as decision-makers with regard to AI. A common way to deal with the political risks implied is to defer to specialized expertise and science-based decision-making (Jasanoff 1990). Scientific research routinely deals with non-knowledge, but typically in a way that specifies what is not yet known, thereby laying the foundation for new knowledge (Merton 1987: 7). Specified non-knowledge entails the expectation that it will be transformed into knowledge, given enough time, as a result of further research. This does not mean that scientific activity gradually decreases the amount of specific non-knowledge and increases the amount of knowledge. Rather, the specification of non-knowledge defines a soluble scientific problem, whose solution inevitably points to new non-knowledge to be specified by further scientific research (Merton 1987; Japp 2000). Yet when a soluble scientific problem is defined in congruence with a political decision-making problem, specified non-knowledge also suggests the possibility of informed decision-making in combination with certified experts that can be invoked to justify the decisions made.

That is why risk as a calculative technique appears to be attractive to decision-makers faced with uncertainty. It transforms specified non-knowledge into knowledge about likelihoods. This sort of knowledge is enticing in many policy fields, for example policing and crime. Whenever there is extensive data about past events, risk calculations are feasible. The availability of big data extends the reach of such calculations to new fields of application. A reliance on risk calculations, whether based on classical statistics or sophisticated deep-learning algorithms, transforms the political problem of crime into various problems of specified non-knowledge. A resulting prediction, for example about the neighborhoods in a particular city where break-ins will most likely occur within the next month, presents an actionable knowledge that can guide decisions about the deployment of limited police forces. Such knowledge is not expected to prevent all break-ins, but only more break-ins than if decisions were taken without such knowledge.

However, the problem of risky AI cannot be addressed in the same way. It is a problem that is potentially created by all the unspecified non-knowledge that is excluded in the course of specifying non-knowledge, to which

the selection of training data, of a particular AI model and all the other steps involved in the creation and implementation of an AI application contribute. Accordingly, there is no empirical data on which to base a calculation that could deal with this non-knowledge; decisions can only rely on judgements of those deemed to be in possession of relevant experience. Yet whenever there is few or no empirical data, the idea that risks could be estimated benefits from a spillover effect (Tierney 1999: 219). Although the method of specifying non-knowledge is utterly different when there is no data to calculate likelihoods, the notion of risk analysis invokes a scientific specification of non-knowledge to legitimate its results. By contrast, if the term risk were dropped or clearly delineated as an everyday expression marking the possibility of negative consequences that decisions about using AI may have, the deeply political dimension of such decisions would become obvious.

Conclusion: Risk and Depoliticization

The plausibility of risk estimates for many technologies benefits from spillover effects. Yet this did not prevent public resistance, for example against nuclear energy or genetically modified organisms, in the past. Firstly, those opposed to a technology aligned themselves with experts who arrived at other conclusions (van den Daele 1996). Secondly, they rejected the notion of specified non-knowledge and observed unspecified non-knowledge instead (Japp 2000), interpreting the existence of different expert opinions as proof that the non-knowledge could not be specified. Unspecified non-knowledge implies non-quantifiable, catastrophic risk (ibid: 231), with people rejecting possible future harms completely, deeming them unacceptable on principle.

As of now, such a politicization of technology based on the distinction between decision-makers and those affected is not in sight in the case of AI. Although this may change in the future, which is, of course, unknown in this respect as in any other, there is reason to believe that it will not change as long as the notion of risk continues to frame the debate. One of its peculiarities is that the underlying technology of AI applications is itself based on probabilistic calculation and aimed at decision problems. Hence the depoliticization of problems that AI applications are supposed to tackle and the depoliticization of problems that AI applications potentially create are intertwined. Automating decisions about eligibility for welfare benefits,

the allocation of police forces, or the deletion of social media posts means by definition that the attribution of responsibility and associated risks shifts. What used to be decisions of policy makers and administrators in relation to particular issues or cases turn into decisions about whether and how to deploy a corresponding AI system. They thus turn into a matter of the risks related to that system. To the extent to which these risks are framed as a matter of research specifying non-knowledge, the depoliticization by AI and the depoliticization of AI are likely to reinforce each other.

Concurrently, the incorporation of ethical principles into the discourse about risks of AI sidesteps the distinction between specifiable non-knowledge and unspecified non-knowledge that fueled resistance to technologies in the past. When the principles on which one might base a rejection of possible future harms categorically are drawn into the framework of risk, they are presented as negotiable, if not quantifiable, and as unpolitical at the same time. This is good news for those who prioritize the further development and spread of AI applications. It is rather bad news for those who fear irreversible societal consequences of some AI applications and believe that trade-offs between values or ethical principles are common and inevitable, but inherently political.

References

AI Ethics Impact Group (2020): "From Principles to Practice. An interdisciplinary framework to operationalise AI ethics." In: *Bertelsmann Stiftung*. URL: https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WKIO_2020_final.pdf.

Aradau, Claudia, and Tobias Blanke (2017): "Politics of Prediction: Security and the Time/Space of Governmentality in the Age of Big Data." In: *European Journal of Social Theory* 20.3, pp. 373–391.

Barry, Laurence, and Arthur Charpentier (2020): "Personalization as a Promise: Can Big Data Change the Practice of Insurance?" In: *Big Data & Society* 7.1. URL: https://doi.org/10.1177/2053951720935143.

Begoli, Edmon, Tanmoy Bhattacharya, and Dimitri Kusnezov (2019): "The Need for Uncertainty Quantification in Machine-Assisted Medical Decision Making." In: *Nature Machine Intelligence* 1.1, pp. 20–23.

Boltanski, Luc, and Laurent Thévenot (2006): *On Justification: Economies of Worth. Princeton Studies in Cultural Sociology*. Princeton: Princeton University Press.

Brenton, Scott (2013): "The political motivations of ethical consumers." In: *International Journal of Consumer Studies* 37.5, pp. 490–497.

Buiten, Miriam C. (2019): "Towards Intelligent Regulation of Artificial Intelligence." In: *European Journal of Risk Regulation* 10.1, pp. 41–59.

Bundesministerium des Innern, für Heimat und Bau (2019): "Mitglieder der Datenethikkommission der Bundesregierung". URL: https://www.bmi.bund.de/DE/themen/it-und-digitalpolitik/datenethikkommission/mitglieder-der-dek/mitglieder-der-dek-node.html.

Campolo, Alexander, and Kate Crawford (2020): "Enchanted Determinism: Power without Responsibility in Artificial Intelligence." In: *Engaging Science, Technology, and Society* 6, pp. 1–19. URL: https://doi.org/10.17351/ests2020.277.

Castel, Robert (1991): "From Dangerousness to Risk." In: *The Foucault Effect. Studies in Governmentality*. Ed. by Graham Burchell, Colin Gordon, and Peter Miller, Chicago: Chicago University Press, pp. 281–298.

Cath, Corinne, Sandra Wachter, Brent Mittelstadt, Mariarosaria Taddeo, and Luciano Floridi (2018): "Artificial Intelligence and the 'Good Society': The US, EU, and UK Approach." In: *Science and Engineering Ethics* 24.2, pp. 505–528.

Collier, Stephen J. (2008): "Enacting catastrophe: preparedness, insurance, budgetary rationalization." In: *Economy and Society* 37.2, pp. 224–250.

Daele, Wolfgang van den (1996): "Objektives Wissen als politische Ressource: Experten und Gegenexperten im Diskurs." In: *Kommunikation und Entscheidung: Politische Funktion öffentlicher Meinungsbildung und diskursiver Verfahren, WZB-Jahrbuch 1996*. Ed. by Wolfgang van den Daele and Friedhelm Neidhardt, Berlin: Edition Sigma, pp. 297–326.

Datenethikkommission (2019): *Gutachten der Datenethikkommission*. URL: https://datenethikkommission.de/gutachten/.

European Commission (2020): *White Paper on Artificial Intelligence: A European Approach to Excellence and Trust*. URL: https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

Garfinkel, Simson (2017): "How angry truckers might sabotage self-driving cars." In: *MIT Technology Review* 120.6, p. 14.

Greene, Daniel, Anna Lauren Hoffmann, and Luke Stark (2019): "Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning." Hawaii International Conference on System Sciences. URL: http://hdl.handle.net/10125/5965.

Guikema, Seth (2020): "Artificial Intelligence for Natural Hazards Risk Analysis: Potential, Challenges, and Research Needs." In: *Risk Analysis* 40.6, pp. 1117–1123.

Hagendorff, Thilo (2020): "The Ethics of AI Ethics: An Evaluation of Guidelines." In: *Minds and Machines* 30.1, pp. 99–120.

Japp, Klaus P. (2000): "Distinguishing Non-Knowledge." In: *Canadian Journal of Sociology* 25.2, pp. 225–238.

Japp, Klaus P., and Isabel Kusche (2008): "Systems Theory and Risk." In: *Social Theories of Risk and Uncertainty: An Introduction*. Ed. by Jens O. Zinn, Malden, MA: Blackwell, pp. 76–105.

Jasanoff, Sheila (1990): *The fifth branch: science advisers as policymakers*. Cambridge, Mass: Harvard University Press.

Jasanoff, Sheila (1999): "The Songlines of Risk." In: *Environmental Values* 8, pp. 135–152.

Jobin, Anna, Marcello Ienca, and Effy Vayena (2019): "The Global Landscape of AI Ethics Guidelines." In: *Nature Machine Intelligence* 1.9, pp. 389–399.

Kusche, Isabel (2021): "Systemtheorie und Ideologie. Eine Spurensuche." In: *Die Rückkehr der Ideologie*. Ed. by Heiko Beyer and Alexandra Schauer, Frankfurt: Campus, pp. 111–139.

Laat, Paul B. de (2019): "The Disciplinary Power of Predictive Algorithms: A Foucauldian Perspective." In: *Ethics and Information Technology* 21.4, pp. 319–329.

Luhmann, Niklas (1962): "Wahrheit und Ideologie: Vorschläge zur Wiederaufnahme der Diskussion." In: *Der Staat* 1.4, pp. 431–448.

Luhmann, Niklas (1991): *Soziologie des Risikos*. Berlin; New York: W. de Gruyter.

Makropoulos, Michael (1990): "Möglichkeitsbändigungen: Disziplin und Versicherung als Konzepte zur sozialen Steuerung von Kontingenz." In: *Soziale Welt* 41.4, pp. 407–423.

Merton, Robert K. (1987): "Three Fragments from a Sociologist's Notebooks: Establishing the Phenomenon, Specified Ignorance, and Strategic Research Materials." In: *Annual Review of Sociology* 13.1, pp. 1–29.

O'Malley, Pat (2009): "'Uncertainty makes us free'. Liberalism, risk and individual security." In: *Behemoth – A Journal on Civilisation* 2.3, pp. 24–38.

Perrow, Charles (1984): Normal accidents: living with high-risk technologies. New York: Basic Books.

Rességuier, Anaïs, and Rowena Rodrigues (2020): "AI Ethics Should Not Remain Toothless! A Call to Bring Back the Teeth of Ethics." In: *Big Data & Society* 7.2. URL: https://doi.org/10.1177/2053951720942541.

Rouvroy, Antoinette, and Thomas Berns (2013): "Gouvernementalité algorithmique et perspectives d'émancipation: Le disparate comme condition d'individuation par la relation?" In: *Réseaux* 177.1, pp. 163–196. URL: https://doi.org/10.3917/res.177.0163.

Scherer, Matthew U. (2015): "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies." In: *Harvard Journal of Law and Technology* 29.2, pp. 354–400.

Speicher, Till, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar (2018): "A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual &Group Unfairness via Inequality Indices." In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, London/United Kingdom: ACM, pp. 2239–2248. URL: https://doi.org/10.1145/3219819.3220046.

Tierney, Kathleen J. (1999): "Toward a Critical Sociology of Risk." In: *Sociological Forum* 14.2, pp. 215–242.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell (2021): "Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI." In: *Computer Law & Security Review* 41, July 2021, 105567. URL: https://doi.org/10.1016/j.clsr.2021.105567.

Wynne, Brian (2001) "Creating Public Alienation: Expert Cultures of Risk and Ethics on GMOs." In: *Science as Culture* 10.4, pp. 445–481.