Philosophy & Technology (2024) 37:66
https://doi.org/10.1007/s13347-024-00755-7

RESEARCH ARTICLE

AI-Related Risk: An Epistemological Approach

Giacomo Zanotti[1] · **Daniele Chiffi**[2] · **Viola Schiaffonati**[1]

Received: 22 January 2024 / Accepted: 6 May 2024

Abstract
Risks connected with AI systems have become a recurrent topic in public and academic debates, and the European proposal for the AI Act explicitly adopts a risk-based tiered approach that associates different levels of regulation with different levels of risk. However, a comprehensive and general framework to think about AI-related risk is still lacking. In this work, we aim to provide an epistemological analysis of such risk building upon the existing literature on disaster risk analysis and reduction. We show how a multi-component analysis of risk, that distinguishes between the dimensions of hazard, exposure, and vulnerability, allows us to better understand the sources of AI-related risks and effectively intervene to mitigate them. This multi-component analysis also turns out to be particularly useful in the case of general-purpose and experimental AI systems, for which it is often hard to perform both ex-ante and ex-post risk analyses.

Keywords AI · Risk · Components of risk · General-purpose AI systems · Experimental technologies

✉ Giacomo Zanotti
giacomo.zanotti@polimi.it

Daniele Chiffi
daniele.chiffi@polimi.it

Viola Schiaffonati
viola.schiaffonati@polimi.it

[1] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy

[2] Department of Architecture and Urban Studies, Politecnico di Milano, Milan, Italy

Published online: 25 May 2024

1 Introduction

Progress in the field of Artificial Intelligence (AI) has been increasingly rapid, and AI systems are now widespread in societies. In parallel, there has been a growing interest in the ethical and socially relevant aspects of the design and use of AI systems. Among other things, an increasing focus has been directed towards the *risks* associated with the widespread adoption of AI systems.

To begin, a lot of attention (and media coverage) has been devoted to the so-called existential risks, namely the risk of human extinction and global catastrophes due to the development of misaligned AI. In particular, the evoked scenarios focus on the development and deployment of Artificial General Intelligence (AGI), namely human-level or even beyond-human AI. Interestingly, concerns of this kind have motivated initiatives such as the Future of Life Institute's open letter *Pause Giant AI Experiments*,[1] or the Center for AI Safety's public *Statement on AI Risk* according to which "mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war".[2]

That said, many have criticized the insistence on such futuristic scenarios and narratives about AI takeover, arguing that the use of AI systems already involves way more mundane forms of risk.[3] Examples will be made in the article, but we can mention at least problems related to algorithmic discrimination (Buolamwini & Gebru, 2018), privacy violation (Curzon et al., 2021), environmental impacts and exploitation of human labour (Crawford, 2021). Along these lines, notable attempts have been made to regulate the design and use of *current* AI systems and reduce their actual risks, especially within the normative framework of Trustworthy AI (see Zanotti et al., 2023). Most notably, a lot of attention has been devoted to the recently approved regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (AI Act), a unified legal framework for AI.[4] Interestingly, the AI Act explicitly adopts a risk-based approach, that groups together AI systems into different levels of risk. In particular, it explicitly distinguishes systems involving *unacceptable risks* (e.g., those used for social scoring) and *high-risk systems* (e.g., those used for predictive justice). Two other levels of risk can be identified, even if no precise label is employed in the AI Act: *limited-risk* systems (e.g., chatbots) and *minimal-risk* systems (e.g., spam filters). Each level of risk is then associated with a specific level of regulation: unacceptably risky systems are prohibited (Art. 5); high-risk systems need to comply with strict requirements concerning, among other things, traceability, human oversight, accuracy, security, and robustness (Chapter III); limited-risk systems must respect transparency requirements (Art. 50);

[1] https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

[2] https://www.safe.ai/statement-on-ai-risk.

[3] https://www.nature.com/articles/d41586-023-02094-7.

[4] More precisely, the *Regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts* (https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.html).

finally, the development and use of minimal-risk systems should only be subject to codes of conduct.[5]

Now, our work does not directly address the AI Act, but rather aims to lay some philosophical and conceptual bases to better understand AI-related risk. And while this kind of work should ideally inform actual interventions, also in terms of regulation, it is not meant to be interpreted as providing readily applicable instructions and suggestions to policymakers and stakeholders.[6] Our perspective will be distinctively epistemological, and more general in scope. That said, the AI Act provides us with a useful starting point for our discussion. While being a crucial step towards a responsible and trustworthy development of AI, the Act is not free from limitations (e.g., Mahler, 2022; Edwards, 2022; Floridi, 2021; Mökander, 2022).[7] In our view, a potential problem of the Act is that, while explicitly adopting a risk-based approach, it lacks a proper conceptualization of the notion of risk. Improvements were made with the amendments approved in June 2023, for Art. 3 (2) now explicitly defines risk as "the combination of the probability of an occurrence of harm and the severity of that harm". Still, we will argue, this understanding of risk is not enough when it comes to assessing and possibly mitigating AI-related risk.

However, the scope of our analysis extends beyond the AI Act. While other kinds of risks, such as natural risks, have already been investigated from the perspective of the philosophy of science, little work has been done on the epistemology of risk in the context of AI. True, *specific* AI-related risks have been investigated: again, discrimination, privacy violation, environmental impacts, and so on – see Wirtz et al. (2022) for a useful panoramic overview of AI-related risk.[8] However, a comprehensive framework is still lacking.[9] Our aim in this article is to provide an epistemological analysis of AI-related risk that distinguishes its different components by building upon the existing literature on disaster risk analysis and reduction, which is usually (but not exclusively) adopted for natural risks. As we will see, such an approach turns out to be particularly fruitful when it comes to designing risk-mitigation policies, for distinguishing the different components of AI-related risk also opens the way for different kinds of intervention for mitigation.

[5] Although we will not explicitly address the question in this paper, it is noteworthy that, in the latest versions of the AI Act, the category of systemic risk was added specifically in relation to the risks associated with general-purpose AI models (Chapter V).

[6] See Novelli et al. (2023, 2024) for a different framework, meant to be directly applied to the AI Act.

[7] Among other things, the AI Act's list of high-risk systems has been criticized. For instance, Prainsack and Forgó (2024) have recently emphasized how systems classified as "medical devices" are considered high-risk ones regardless of their actual use. A system such as a smartwatch, on the other hand, may pose analogous risks and yet be excluded from the list of high-risk systems due to its being classified as a lifestyle gadget. This, the authors argue, "creates competitive advantages for companies with sufficient economic power to legally challenge high-risk assessments".

[8] Most contributions in the literature have from time to time focused on specific risks related to the deployment of specific systems and/or in specific contexts, while systematic and comprehensive reviews seem to be rarer.

[9] A notable exception is represented by the Artificial Intelligence Risk Management Framework developed by the National Institute of Standards and Technology (NIST, 2023). Although the framework is practically oriented, it provides an analysis of how AI-related risks differ from risks of traditional software systems.

The article is structured as follows. Section 2 presents different approaches to the conceptualization of risk, focusing on multi-component analyses that understand risk as resulting from the interplay of three different components: hazard, exposure, and vulnerability. In Sect. 3, we argue in favour of the application of this multi-component analysis to AI-related risks, showing how it allows us to better capture different aspects of such risks and design more effective interventions for mitigation. In Sect. 4, we develop the analysis presented in Sect. 3 by focusing on the difficulties involved in providing ex-ante analyses of AI-related risks, especially when we deal with general-purpose AI systems having the character of experimental technologies. Section 5 provides a brief summary and closes the article.

2 Components of Risk

Our analysis should probably start with a caveat, namely that there is no univocal notion of risk (Boholm et al., 2016; Hansson, 2023). This is due both to differences in the way risk is defined in the literature and to the fact that technical definitions of risk coexist with the ordinary understanding and usage of this notion.

Focusing on technical definitions, today's dominant approaches conceive of risk in terms of expected utility (Hansson, 2009). That is, risk is given by the combination of the probability of an unwanted event occurring and the magnitude of its consequences.[10] As an example, consider volcanic risk. On the one hand, a large and highly explosive eruption might be associated with a low level of risk if its occurrence is estimated as very unlikely with a fair degree of confidence. A moderately explosive but way more likely eruption, on the other hand, would arguably be associated with a higher level of risk. Despite its simplicity, this way of thinking about risks provides us with an easily applicable and intuitive model for decision-making in contexts of risk that nicely fits with theories of rational choice insisting on the maximisation of the expected utility (see Briggs, 2023).

As already anticipated, this definition of risk is the one the AI Act explicitly refers to. However, always maintaining a definition in terms of expected utility, one can decide to provide further analyses of risk. Most notably, risk can be decomposed into its different *components* – usually, hazard, exposure, and vulnerability. As we will see in a moment, this approach, fairly common in risk analysis, allows to open different areas of intervention for mitigation. Since we aim to provide an analysis of AI-related risk that can fruitfully serve also as a ground for policy making, the multi-component analysis is the one adopted in this article.[11]

Let us now go a bit more into the details of the components of risk. Starting with the first one, the notion of *hazard* refers to the source of potential harm – let us recall that, when it comes to risk, the focus is always on *unwanted* consequences. In addi-

[10] The probabilistic component of risk is also the main ingredient of the Royal Society's (1983) definition of risk as "the probability that a particular adverse event occurs during a stated time period, or results from a particular challenge".

[11] Again, the multi-component analysis of risk is not meant to be a definition of risk, let alone an alternative one with respect to the definitions in terms of expected utility. It should rather be understood as an additional analysis aiming at decomposing specific risks to facilitate mitigation interventions.

tion to the specification of the source of the potential harm as well as of its characteristics in terms of magnitude, the analysis of hazards often involves a probabilistic element – that is, the probability of occurrence of the harmful phenomenon (e.g., UNDRO, 1991). Consider, for instance, risks related to volcanic eruptions. In this case, the hazard is primarily the eruption itself, which in its turn brings about a series of potentially harmful events, such as pyroclastic and lava flows. Different elements contribute to making this kind of hazard more or less impactful for risk analysis, such as the frequency of the eruptions, their intensity and their duration.

The domain of natural disasters offers notable examples of other kinds of hazards, such as earthquakes, tidal waves and flooding. However, importantly for our purpose, hazards can also have origins other than natural ones. For example, the escalation of an armed conflict is a distinctive example of a non-natural hazard. Further narrowing the focus, we can consider technological risks related to technological artefacts, namely objects produced by humans in order to fulfill some kind of practical function (Vermaas et al., 2011, p. 5).[12] In the next section, we will focus on risks stemming from AI systems.

As suggested by the label, the component of *exposure* refers to what could be harmed. Importantly, living beings – most notably, humans – can be exposed, but we can also think of risks in which material assets such as buildings and infrastructures are involved. Going back to the example of volcanic risk, the exposure has to do with the number of people, buildings, infrastructures, and other assets that would be affected by the eruption. Importantly, as we will see in the next section in relation to some AI systems, even minimal hazards can be associated with high levels of risk – and eventually bring about disastrous outcomes – when exposure is high.

Finally, the component of *vulnerability* unsurprisingly has to do with *how much* the exposed people or assets are susceptible to the impacts of hazards.[13] Providing a precise characterization of vulnerability is far from easy, for a number of different definitions are available (Thywissen, 2006), and factors affecting vulnerability may significantly vary. In general, however, they include all those circumstances and measures that could make people or assets more or less defenceless against harming events. In the case of volcanic eruptions, this might translate into the existence

[12] The distinction between natural and technological risk is sometimes blurred. In fact, one could also refer to risks having mixed origins, both natural and technological – the so-called "Natech" risks (UNISDR, 2017). That said, it is important to note that the division between natural and human-made (or technological) risks is highly practical in various scenarios, but it is difficult to make a clear distinction between the two categories (Hansson, 2016). Frequently, some aspects of the same risk may be labelled as natural in certain situations and as technological in others. It is therefore advisable to avoid the oversimplified division of accidents into two rigid categories, natural and human-made.

[13] Note that risk is given only in those cases in which all three components are present. There is clearly no risk if there is no hazard, but there is no risk also if no-one is exposed to harm or vulnerable. A reviewer interestingly points out that we might imagine situations in which interventions are successfully performed that significantly reduce the vulnerability of some components of the population with respect to a certain hazard, to the point where the risk related to the hazard in question becomes negligible for them. In such a case, from a population perspective, the hazard remains but the overall risk is mitigated, for there are fewer exposed people who are also vulnerable.

and feasibility of plans for evacuation, shelters as well as food and water emergency supplies.[14]

Once these three components of risk are clearly identified, different interventions for mitigating risk can be designed. First of all, one may take measures to reduce hazards. Here, some distinctions between different kinds of risks shall be made, for hazard reduction is not always possible. In particular, hazard mitigation is not so easy when it comes to natural risk. True, there are some cases in which interventions for reducing hazards are possible – for instance, flooding is influenced by anthropogenic climate change, and hazards like landslides can be due to logging and land abuse. However, in many other cases, including volcanic risk, hazard mitigation is simply not possible, for the occurrence of the unwanted event is independent of human action.

On the contrary, if the risks in question are related to the use of a certain technological artefact, the hazard can sometimes be reduced or even eliminated. Most notably, measures could be taken by prohibiting the use of the artefact and withdrawing it from the market, as it happened in several countries with the ban on asbestos. Sure, things are not always easy. In many cases, hazard mitigation in contexts of technological risk presents significant challenges – think about interventions to cap carbon dioxide emissions.[15] Still, as it happened with asbestos, there seem to be cases in which mitigating technological hazard is feasible.

That said, hazard reduction is not the only way to mitigate risk. Among other things, risk mitigation strategies might attempt to reduce the exposure. Thinking about natural risks stemming from the occurrence of geographically circumscribed events, building and access permits could be denied in the potentially affected areas – such as the vicinities of a volcano. However, exposure can be reduced also when it comes to technological risks. If some signs of structural failure are detected in a bridge, for example, then exposure can be significantly reduced by forbidding access to the bridge. In the field of information and communication technologies (ICT), instead, age restrictions on the use of services and products – e.g., social networks – can be seen, at least in principle, as measures for exposure reduction.

Finally, interventions could aim at making the population and assets less vulnerable. These interventions can significantly vary as a result of the fact that vulnerability is a very broad notion and involves different factors. We have seen how, in the case of volcanic risk, they involve actions such as building shelters, designing evacuation plans and planning basic necessities and supplies. In the case of ICT, antivirus software and spam filters play an analogous function, protecting the user from malwares and potentially dangerous content.

[14] Sometimes a fourth component of risk is acknowledged, i.e. capacity, even if this concept is usually assumed as something pertaining to vulnerability. In particular, capacity is defined as "the combination of all the strengths, attributes and resources available within an organisation, community or society to manage and reduce disaster risks and strengthen resilience. […] Capacity may include infrastructure, institutions, human knowledge and skills, and collective attributes such as social relationships, leadership and management" (Sendai Framework Terminology on Disaster Risk Reduction, https://www.undrr.org/terminology/capacity).

[15] We thank one of the reviewers for suggesting this clarification.

As anticipated, there is no univocal notion of risk. On the contrary, different ways to conceptualise and analyse risk are possible and could be more or less useful depending on the context. Here, we have focused on a multi-component analysis of risk that distinguishes between the components of hazard, vulnerability, and exposure.

3 AI and Multi-Component Analysis of Risk

We have seen how a multi-component analysis of risk can be employed to understand both natural and technological risks and at the same time pave the way for different kinds of interventions aiming at risk-mitigation. We now wish to narrow the focus to risks stemming from AI systems. However, a problem immediately emerges concerning what we mean by "AI system", for the definition of AI is a long-standing problem at least since the foundation of the discipline, and a number of different definitions are available (Russell & Norvig, 2021).[16]

For the purpose of this article, the OECD's (2023) definition can be kept in mind, according to which an AI system is.

> [...] according to which an AI system is "a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment".[17]

With this definition of AI systems in mind, we will show through some relevant examples how the multi-component analysis of risk we have considered fruitfully applies to AI-related risk.

Let us start by considering the hazard component involved in the use of AI systems, which is arguably the most discussed aspect of AI-related risk. As already noted, AI systems are now increasingly employed in a number of contexts that we intuitively perceive as highly risky. Among the most discussed cases, one can think about systems used in medicine (Panayides et al., 2020), in courts (Queudot & Meurs, 2018), and in war scenarios (Amoroso & Tamburrini, 2020). In these cases, it is pretty straightforward why the use of AI systems involves considerable risks. Though increasingly accurate in their predictions and classifications, many state-of-the-art AI systems are still subject to errors and malfunctions. Consider, as an example, a system used for the detection of skin cancers. Such a system might be remarkably accurate in distinguishing cancerous tissues from benign lesions, maybe even more than a human doctor (Soenksen et al., 2021). Still, the possibility of a misdiagnosis

[16] See Floridi (2023) for an up-to-date overview of different (legal) definitions of AI.

[17] https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. Note that this is not the only possible definition of AI, not even if we narrow it down to legal definitions. The AI Act, for example, defines an AI system as "a machine-based system designed to operate with varying levels of autonomy, that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" (Art. 3, 1).

is open, with potential life-threatening consequences for the patients. Analogously, AI systems employed in war scenarios can make errors in target identifications, and biased systems employed in courts can result in unjust incarceration (Angwin et al., 2016). And when the stakes are high, such errors and malfunctions result in high levels of hazards.[18]

Hazard, however, is not the only component that we should take into account. Consider AI-based recommender systems. These systems are nowadays widespread and integrated into a number of online services and platforms, and they are used to filter content – advertisements, buying suggestions, music, videos, and so forth – based on the user's interests. These interests are typically predicted on the basis of the users' online habits and previous choices. If one focuses exclusively on the hazard component, these systems do not strike as particularly risky, especially when compared with systems whose failure can result in human victims.

However, things change when the component of exposure is considered. Due to their being pervasive in online environments, including extremely popular platforms, recommender systems virtually monitor and influence the behaviour of all users. As a result, all the possible concerns about privacy, addiction, and manipulation apply on a massive scale. The unwanted consequences might not be so detrimental for the individual, but the potential risks involved in their use are characterised by extremely high levels of exposure.

What is more, the use of AI systems involves risks whose component of exposure goes way beyond the system's users. In particular, we have in mind risks related to AI systems' environmental impact. Among other things, there is a growing awareness that the training and use of ML models require significant amounts of energy, which results in an increasing carbon footprint (OECD, 2022; Verdecchia & Cruz, 2023). In this sense, the emission of greenhouse gases related to the use of AI systems involves risks with a potentially global impact (Tamburrini, 2022).

Finally, some AI systems strike for the vulnerability of their users. In this regard, interesting examples come from those contexts in which AI systems interact in social environments with specific kinds of population. These systems, explicitly designed to interact with humans by following social rules, are often equipped with modules and software that allow them to recognize users' affective states and suitably simulate emotion-driven behaviour. A possible example is represented by AI-powered social robots. These systems come in many forms and shapes and are increasingly used in the context of older adults' care (Miyagawa et al., 2019) and in different educational environments (Tanaka et al., 2015) – e.g., with children with autism spectrum disorders (Rakhymbayeva et al., 2021). When it comes to these settings, the weight distribution among the components of risk is different again, for the real cause of concern has to do with vulnerability. As a matter of fact, older adults and children

[18] Analogous considerations are explicitly made in the AI Act concerning AI-based safety components in digital infrastructures, road traffic and the supply of water, gas, heating and electricity, whose failure or malfunctioning "may put at risk the life and health of persons at large scale and lead to appreciable disruptions in the ordinary conduct of social and economic activities" (recital 55). Consistently with the literature on the non-epistemic aspects of inductive risk (see Douglas, 2000), this kind of error also has an impact at the level of values (Karaca, 2021).

are the prototypical vulnerable populations.[19] On the contrary, hazard levels are reasonably low. True, it is an open possibility that a malfunction of the robot results in someone getting physically hurt. More frequently, however, errors are "social errors" (Tian & Oviatt, 2021), episodes involving the breaking of social rules and failures in the recognition and display of emotions. At the same time, the deployment of social robots in older adults' care and education does not necessarily involve high levels of exposure, for it typically takes place in small and controlled environments.

Summing up, distinguishing between the components of hazard, exposure and vulnerability allows us to better identify the different sources of the risks involved in the use of different AI-based technologies. It is worth noting that, just like all other risks, AI-related risk always results from the interplay of all three components. Consider – again – the case of a recommender system implemented in a social network suggesting links to products on an e-commerce website. We have seen how, being integrated into widely used online platforms, these systems involve significant risks due to their high level of exposure. However, one could also take into account how they may end up exploiting users' weaknesses to maximise sales profits. In this case, when assessing risk, the focus has to be on the interplay between exposure and users' vulnerability. Besides the schematism of the discussion presented here, the point is that adopting a multi-component analysis of risk enables a better assessment of AI-related risk and could pave the way for better mitigation strategies.

4 AI-Related Risk: Flexibility and Experimentality

While the adoption of a multi-component approach to risk analysis puts us in a better position to deal with AI-related risks, some difficulties remain. In this section, we focus on the risks stemming from the deployment of AI systems having the character of *experimental technologies* and qualifying as *general-purpose* ones – more on this in a moment. These two features, increasingly common in many AI systems, give rise to difficulties when it comes to providing an ex-ante analysis of the involved risks, and may reduce the usefulness of ex-post ones. We will show how, even in these cases, a multi-component analysis of risk allows us to better understand sources of AI-related risk and thereby plan mitigation interventions.

4.1 Ex-Ante and Ex-Post Risk Analyses

Let us start by introducing the general notions of ex-ante and ex-post risk analysis. In risk analysis and methods for economic evaluations like cost-benefit analysis, it is quite standard to distinguish ex-ante evaluations of risk, namely risk assessments conducted *before* the realization of a project or policy, from ex-post evaluations of risk, which occur after a specific project or policy – in our case, for example, the

[19] Note that vulnerability would arguably deserve a separate and detailed treatment, for the question of vulnerability and AI is often raised but seldom investigated. In particular, it would be interesting to analyse to what extent we can exclusively rely on "classic" vulnerable groups (such as older adults and children) when assessing AI risk or whether we should rethink the very concept of vulnerability and its categories in light of technical and social changes in the AI landscape.

introduction of a new AI-based technology – has been completed (de Rus, 2021). Note that ex-ante evaluations may face severe forms of empirical uncertainty, particularly when evaluating risks that may occur in a distant future and are not limited to specific geographic areas, which can be difficult to identify and quantify (Hansson, 1996). For example, new emerging risks may manifest during the implementation of a project and could be in some circumstances quite unimaginable in the ex-ante phase. As we will see in a moment, this is particularly relevant when it comes to the so-called experimental technologies.

A possible way to address these risks is to provide ex-post risk evaluations of projects or policies. Through this kind of retrospective analysis, it is possible to identify and address emergent risks that would have been difficult to consider in the ex-ante phase. In an ex-post evaluation, factual uncertainty can be dramatically reduced, even though there may still be uncertainty regarding the counterfactual scenario in which a specific intervention was not executed.[20] Moreover, through ex-post analysis, we can gain a better understanding of the exact magnitude of the risks, the extent of their exposure, and the key factors influencing vulnerability. In light of this, ex-post evaluations can be used to inform and partially shape future ex-ante evaluations of similar new projects and risks.

Having introduced the notions of ex-ante and ex-post risk analysis, we can now consider the difficulties involved in performing such analyses within the context of AI.[21] To this aim, let us go back for a moment to the OECD's definition presented in Sect. 3, stressing the artefactual nature of AI systems, their interaction capabilities and ability to learn, the role played by inferences and their impact on decisions, and the different levels of autonomy. This definition captures pretty well the features of many current AI systems, that are capable of performing complex tasks in unknown environments by constantly using new data. Now, several of the current AI techniques adopted to achieve these capabilities produce results that are opaque and very often difficult to explain. However, complexity does not only concern the very nature of these systems. On the contrary, it has a lot to do with their interaction with environments (including humans) that in many cases are not known in advance.

To make things worse, many of these technologies are radically innovative, and their introduction into society is *de facto* unprecedented. Indeed, they could be described as *experimental technologies* according to the characterization provided by van de Poel (2016). By definition, experimental technologies are those technologies whose risks and benefits are hard to estimate before they are properly inserted in their context of use, for "there is only limited operational experience with them, so

[20] Note that we are significantly simplifying the matter for the sake of exposition. As we will see in a moment, ex-post analyses often present significant challenges.

[21] The distinction between ex-ante and ex-post risk analysis is not reducible to the difference between 'inherent risk' and 'residual risks', although there are some commonalities. Inherent risk represents the amount of risk that exists in the absence of risk mitigation measures, while residual risk refers to the risk that remains after an organization has implemented measures to mitigate the inherent risks (Gorecki, 2020). From an ex-ante perspective, we need to consider in advance the risks occurring either in the presence or absence of *possible* mitigation strategies, while from an ex-post perspective, we need to evaluate the risks resulting from *concretely adopted* mitigation strategies and imagine counterfactually what risks could have materialized without such strategies.

that social benefits and risks cannot, or at least not straightforwardly, be assessed on basis of experience" (van de Poel, 2016, p. 669). Several technologies, such as nanotechnologies or human enhancement drugs, may qualify as experimental according to this definition. And indeed, many of the current AI systems seem to perfectly fit in the category, given their radically innovative character and their being designed to interact with unknown environments.

The assessment of AI-related risks is also complicated by the fact that many current AI applications are based on so-called general-purpose AI systems (GPAIS). In a nutshell, GPAIS are pre-trained models that can constitute the basis for very different AI systems that can in their turn be fine-tuned to better perform in specific contexts of application (Gutierrez et al., 2023).[22] As a result, they can be used for a variety of purposes, that need not be anticipated in the training phase.

The combination of the two features we have just seen, namely being experimental technologies and being general-purpose systems, makes it particularly difficult to assess *ex-ante* – that is, before the deployment of the system – all the risks involved in the use of certain AI systems. Consider, for instance, Large Language Models (LLMs). LLMs are relatively new in the AI landscape, at least if we focus on transformer-based architectures (Vaswani et al., 2017). Among other things, these systems, trained on huge datasets, can produce impressively convincing texts on the basis of prompts given by the users. While transformer-based LLMs have been available for some years now, they have become extremely popular among the general public after OpenAI's launch of ChatGPT in November 2022. From a technical point of view, ChatGPT was not dramatically revolutionary: it was based on a pre-existing model that was fine-tuned to make it suitable for conversation. From a more societal perspective, however, it was groundbreaking: since November 2022, everyone has the possibility to interact free of charge and through a user-friendly interface with a state-of-the-art language model that impressively performs in a number of tasks, generating textual outputs that are often hardly distinguishable from human-produced ones. And in fact, ChatGPT reached one million users in five days.[23]

Somewhat unsurprisingly, in a relatively short period of time, LLM-based applications have multiplied.[24] Among the interesting features of these models, their flexibility stands out. Leaving aside specific limitations imposed by programmers – e.g., they should typically be unable to produce discriminatory and pornographic outputs – they can generate virtually any type of text, can be combined with other models for multimodal processing and generation, and can easily be fine-tuned to be adapted to specific domains. This flexibility allows for a multiplicity of uses. Among other things, LLMs show promising applications in medicine (Thirunavukarasu et al., 2023), finance (Wu et al., 2023), coding (Xu et al., 2022), and education (Kasneci et al., 2023). And even within these contexts, LLM-based systems can be used for a variety of applications, as a testimony to their general-purpose character. However,

[22] A closely related – and sometimes overlapping – notion is the one of foundation model (Bommasani et al., 2022).

[23] https://www.statista.com/chart/29174/time-to-one-million-users/.

[24] As it often happens with other AI applications, the pace of technological innovation also makes it hard to keep regulations up to date.

this flexibility comes with a cost, namely a greater potential for hazard generation: different kinds of errors and failures can occur in the different applications of LLMs, from misdiagnoses to wrong predictions of the stock market resulting in monetary loss. On top of all that, LLMs' flexibility opens the way to a great deal of misuses. For instance, one could use a LLM to write hardly detectable malwares,[25] or generate disinformation (Bagdasaryan & Shmatikov, 2022), also through the generation of fooling images. Provided that other technologies (even non-AI ones) are potentially related to many of the hazards involved in the use of LLMs, LLMs stand out in that they qualify as *multi-hazard* systems.

So, not only the large-scale use of LLM-based applications is unprecedented, which makes these technologies experimental in van de Poel's (2016) sense. It is also characterised by high degrees of flexibility, for LLMs can easily be fine-tuned to specific tasks that need not be anticipated by the initial designers. These two features of LLMs make it extremely difficult to predict *ex-ante* the risks involved in the use of such technologies, for the operational experience with them is limited and it is hard to anticipate all their possible applications, and therefore all the associated risks.

Sure, *ex-post* analyses can be performed after the occurrence of unwanted events related to the use of LLMs. However, conducting ex-post risk evaluations in the presence of a recently developed AI technology can be extremely difficult, as we may not yet have all the required information needed to provide a complete retrospective risk analysis. More specifically, these analyses show severe limitations when it comes to considering all the unexplored risks of novel uses of LLMs. Given the experimental character of these technologies and their flexibility, it is reasonable to assume that uses and misuses other than those targeted by the *ex-post* analysis in question will be a reason for concern.[26]

4.2 Risk Mitigation Strategies

Given these premises, how can we intervene to reduce the risks related to the use of these systems? Here, the analysis of risk we have presented in Sect. 3 turns out to be particularly helpful.

It is quite straightforward that we cannot intervene to reduce unspecified hazards – think about a hazard related to an unanticipated misuse of an LLM-based program. However, we can identify some *areas of concern* and intervene to limit the use of LLMs in that context. For instance, we might be concerned about and therefore limit the use of LLMs for medical diagnosis, if only for the fact that LLMs are still prone to hallucinations and the consequences of a misdiagnosis can be fatal. More generally, the idea is that, provided that we cannot anticipate with precision the unwanted

[25] https://www.cyberark.com/resources/threat-research-blog/chatting-our-way-into-creating-a-polymorphic-malware.

[26] When it comes to LLMs, the issue of emergent capabilities further complicates the matter. The point is that, once they are actually used, these models exhibit abilities that could not be foreseen during the training phase. Such capabilities can also emerge when an LLM is scaled up in terms of hyperparameters and trained on a broader dataset. Needless to say, this makes it extremely hard to predict all possible uses and thus the potential risks of these systems.

consequences of the employment of LLMs, we can significantly mitigate the hazards by limiting the contexts of their use.

To do so, our first move could involve intervening at the design level – i.e., by placing constraints on the kind of answers an LLM-based tool could provide. Otherwise, we could intervene through regulation, by limiting the use of LLMs in critical contexts. Furthermore, approaches that integrate normative and technical solutions can be pursued. Consider the growing problem of distinguishing artificially generated contents (text, images, and so on) from those produced by humans, whose implications can be dramatic – just think about the use of deepfake in warfare and political contexts (Twomey, 2023) as well as for the so-called revenge porn (Kirchengast, 2020). Among others, Knott et al. (2023) argue for a legislative mechanism according to which the public release of general-purpose generative AI models should be possible only if a reliable tool for the detection of contents generated by the model is also made available. This way, we would intervene at the regulatory level to impose technical interventions.

That being said, hazard is not the only component of risk we should focus on when it comes to GPAIS, whose applications are not always significantly problematic from the point of view of the hazard. It is also important to consider that exposure levels play a crucial role in determining risk. We have seen how GPAIS have suddenly gained popularity, how their flexibility makes them suitable for a number of different tasks and how they are at the basis of easily accessible and user-friendly applications. When assessing the risks related to their use, all of this translates into significantly high levels of exposure. At the same time, as we have seen, we have very little experience with their being used at a large scale, and we can hardly predict the involved risks. A promising strategy, at least in principle, would therefore be to introduce these technologies into society by initially limiting their users, thereby intervening on exposure. This would allow us to collect data and monitor the impact of such technologies, which would in turn allow us to take the appropriate measures to ensure their safe and beneficial large-scale deployment (van de Poel, 2016).

Now, things are easier said than done, for a certain tension underlies this strategy. On the one hand, we can hardly anticipate the consequences of the large-scale use of GPAIS as well as their societal impact. On the other hand, a possible solution to deal with this uncertainty consists in preliminarily testing, so to say, these technologies in restricted and monitored environments. However, it is not clear to what extent this strategy could work, for there may be significant risks that emerge only when a system is extensively used. Note that this is not to say that small-scale testing is not helpful. However, it is important to keep in mind that some risks may emerge only at a societal scale, remaining therefore unanticipated. Moreover, a trade-off need to be made between the significance of the smaller-scale testing and its risks: testing the impact of a new technology in a small and controlled environment does not give us much grasp on what could happen at a societal scale but the more controlled environment allows us to easily detect risks and intervene, while larger-scale testing is a better approximation of the introduction of a technology in society but involves greater risks. While there seems to be no easy solution to the problem, this discussion brings forth a crucial point: monitoring processes must play a central role. Given the difficulties involved in anticipating the risks in the use of a certain AI system, especially

when it qualifies as an experimental and a general-purpose one, continuous assessment is required to promptly detect and intervene on emerging risks.[27]

Finally, provided that an eye should be kept on exposure, some criteria are needed to prioritise the protection of certain populations when it comes to the deployment of new AI systems. Here, vulnerability is the key notion. Whether testing a new AI system or considering the impacts of its introduction, prioritising vulnerable populations is crucial. On one hand, there are traditionally vulnerable groups such as older people and children, who, having little to no familiarity with AI systems, are more susceptible to dangers like deception or manipulation associated with the use of AI, and especially of generative AI. On the other hand, it is equally important to pay attention to users employing AI systems in *contexts* of vulnerability, such as in education. While the use of GPAIS in educational settings holds promising applications, it also imposes changes that do not always benefit the involved users. For instance, it might compel us to modify teaching and testing methods, possibly abandoning well-established and effective practices.

Taking stock, we have seen how keeping in mind a multi-component analysis of risk puts us in a better position to cope with AI-related risks emerging from the use of AI experimental technologies and general-purpose systems. Besides being increasingly popular, LLMs are a perfect example of such technologies. However, flexibility and experimentality are distinctive features of many kinds of models, especially if it comes to GPAIS. And while LLMs are among the most studied and used GPAIS, the class is wider and encompasses increasingly powerful models for computer vision – such as Meta AI's Segment Anything Model (SAM),[28] a zero-shot learning system for image segmentation – as well as multimodal processing – such as Google DeepMind's Gemini,[29] a family of models designed to handle image, audio, video, and text. Our general analysis should therefore be applicable to systems implementing all these models, allowing us to better understand and mitigate risk in all those cases in which the generality and experimentality of the model make it difficult to perform and rely on ex-ante and ex-post risk analyses.

5 Conclusions

Increasing attention has recently been devoted to the risks associated with the deployment of AI systems. However, a general epistemological framework for understanding such risks is still lacking. This article attempted to fill this gap by starting from multi-component analyses of risk, typically used in natural risk assessment and reduction, and trying to apply them to the context of AI. We argued that distinguishing between the components of hazard, exposure, and vulnerability allows us to better understand and deal with AI-related risk. This holds also for those AI systems that qualify as

[27] Although we cannot afford to get into the details, incremental approaches based on trial-and-error and small steps testing represent a promising way to deal with these issues (Woodhouse & Collingridge, 1993; van de Poel, 2016).

[28] https://segment-anything.com/.

[29] https://deepmind.google/technologies/gemini/#introduction.

general-purpose and experimental technologies, for which it is often hard to perform ex-ante and ex-post risk analyses and for which continuous assessment and monitoring should be performed. This aligns, for instance, with the risk-based indications of the Food & Drug Administration (FDA) concerning the regulation of AI-based medical products. The FDA clarifies that the deeply iterative, autonomous, and often flexible nature of medical products necessitates a novel regulatory framework for the total product lifecycle. This framework fosters a rapid cycle of product enhancement, empowering these devices to continually improve their functionalities while maintaining robust protective measures even during the pre-market phase (FDA, 2024).

Needless to say, open questions abound. For instance, provided that the multi-component approach to risk borrowed from disaster risk analysis can fruitfully be applied to the domain of AI, it is still unclear whether there are some specificities of AI-related risk that call for a distinct and additional treatment. Or again, it would be interesting to investigate whether and how the development, use and societal impact of AI systems could be understood in terms of (deep forms of) *uncertainty*, further drifting apart from the dimension of probabilistic assessment that is often involved in talk of risk. With our work here, we hope to have provided the epistemological ground for addressing such questions and paved the way to the articulation of a more comprehensive methodology to deal with these risks.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13347-024-00755-7.

Author Contributions All authors have contributed substantially to the manuscript.

Funding The research is supported by (1) Italian Ministry of University and Research, PRIN Scheme (Project BRIO, no. 2020SSKZ7R); (2) Italian Ministry of University and Research, PRIN Scheme (Project NAND no. 2022JCMHFS); (3) PNRR-RETURN-NextGeneration EU program, PE0000005; (4) PNRR-PE-AI FAIR-NextGeneration EU program.
Open access funding provided by Politecnico di Milano within the CRUI-CARE Agreement.

Data Availability Not applicable.

Code Availability Not applicable.

Declarations

Ethical Approval Ethics approval was not required for this study.

Consent to Publish Not applicable.

Consent to Participate Not applicable.

Competing Interests The authors declare they have no competing interests.

References

Amoroso, D., & Tamburrini, G. (2020). Autonomous weapons systems and meaningful human control: Ethical and legal issues. *Current Robotics Reports*, *1*, 187–194. https://doi.org/10.1007/s43154-020-00024-3.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *Pro Publica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Bagdasaryan, E., & Shmatikov, V. (2022). Spinning language models: Risks of propaganda-as-a-service and countermeasures. *2022 IEEE Symposium on Security and Privacy (SP)*, San Francisco (CA), 769–786, https://doi.org/10.1109/SP46214.2022.9833572.

Boholm, M., Möller, N., & Hansson, S. O. (2016). The concepts of risk, safety, and security application in everyday language. *Risk Analysis*, *36*(2), 320–338. https://doi.org/10.1111/risa.12464.

Bommasani, R., Hudson, D. A., Adeli, E., et al. (2022). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Briggs, R. A. (2023). Normative theories of rational choice: Expected utility. In Edward N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy.* https://plato.stanford.edu/archives/fall2023/entries/rationality-normative-utility/.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on Fairness, Accountability and Transparency*, New York: PMLR, 77–91.

Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Curzon, J., Kosa, T. A., Akalu, R., & El-Khatib, K. (2021). Privacy and artificial intelligence. *IEEE Transactions on Artificial Intelligence*, *2*(2), 96–108. https://doi.org/10.1109/TAI.2021.3088084.

de Rus, G. (2021). *Introduction to cost benefit analysis: Looking for reasonable shortcuts*. Edward Elgar Publishing.

Douglas, H. (2000). Inductive risk and values in science. *Philosophy of Science*, *67*(4), 559–579. https://doi.org/10.1086/392855.

Edwards, L. (2022). *Regulating AI in Europe: Four problems and four solutions*. Ada Lovelace Institute.

FDA. (2024). *Artificial Intelligence and medical products: How CBER, CDER, CDRH, and OCP are working together.*https://www.fda.gov/media/177030/download?attachment.

Floridi, L. (2021). The European legislation on AI: A brief analysis of its philosophical approach. *Philosophy and Technology*, *34*, 215–222. https://doi.org/10.1007/s13347-021-00460-9.

Floridi, L. (2023). On the Brussels-Washington consensus about the legal definition of Artificial Intelligence. *Philosophy and Technology*, *36*, 87. https://doi.org/10.1007/s13347-023-00690-z.

Gorecki, A. (2020). *Cyber breach response that actually works: Organizational approach to managing residual risk*. Wiley.

Gutierrez, C. I., Aguirre, A., Uuk, R., Boine, C. C., & Franklin, M. (2023). A proposal for a definition of general purpose Artificial Intelligence systems. *Digital Society*, *2*, 36. https://doi.org/10.1007/s44206-023-00068-w.

Hansson, S. O. (1996). Decision making under great uncertainty. *Philosophy of the Social Sciences*, *26*(3), 369–386. https://doi.org/10.1177/004839319602600304.

Hansson, S. O. (2009). From the casino to the jungle: Dealing with uncertainty in technological risk management. *Synthese*, *168*(3), 423–432. https://doi.org/10.1007/s11229-008-9444-1.

Hansson, S. O. (2016). Managing risks of the unknown. In P. Gardoni, C. Murphy, & A. Rowell (Eds.), *Risk analysis of natural hazards* (pp. 155–172). Springer.

Hansson, S. O. (2023). Risk. In E. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy.* https://plato.stanford.edu/archives/sum2023/entries/risk.

Karaca, K. (2021). Values and inductive risk in machine learning modelling: The case of binary classification models. *European Journal of Philosophy of Science*, *11*, 102. https://doi.org/10.1007/s13194-021-00405-1.

Kasneci, E., Seßler, K., Küchemann, S., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274.

Kirchengast, T. (2020). Deepfakes and image manipulation: Criminalisation and control. *Information & Communications Technology Law*, *29*(3), 308–323. https://doi.org/10.1080/13600834.2020.1794615.

Knott, A., Pedreschi, D., Chatila, R., et al. (2023). Generative AI models should include detection mechanisms as a condition for public release. *Ethics and Information Technology*, *25*, 55. https://doi.org/10.1007/s10676-023-09728-4.

Mahler, T. (2022). Between risk management and proportionality: The risk-based approach in the EU's Artificial Intelligence Act proposal. *Nordic Yearbook of Law and Informatics 2020–2021*, 247–270. https://doi.org/10.53292/208f5901.38a67238.

Miyagawa, M., Kai, Y., Yasuhara, Y., Ito, H., Betriana, F., Tanioka, T., & Locsin, R. (2019). Consideration of safety management when using Pepper, a humanoid robot for care of older adults. *Intelligent Control and Automation*, *11*, 15–24. https://doi.org/10.4236/ica.2020.111002.

Mökander, J., Juneja, P., Watson, D. S., et al. (2022). The US algorithmic accountability act of 2022 *vs* the EU Artificial Intelligence Act: What can they learn from each other? *Minds & Machines*, *32*, 751–758. https://doi.org/10.1007/s11023-022-09612-y.

National Institute of Standards and Technology (NIST) (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

NIST. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. https://doi.org/10.6028/NIST.AI.100-1.

Novelli, C., Casolari, F., Rotolo, A., Taddeo, M., & Floridi, L. (2023). Taking AI risks seriously: A new assessment model for the AI act. *AI & SOCIETY*, 1–5. https://doi.org/10.1007/s00146-023-01723-z.

Novelli, C., Casolari, F., Rotolo, A., Taddeo, M., & Floridi, L. (2024). AI risk assessment: A scenario-based, proportional methodology for the AI act. *Digital Society*, *3*(1), 1–29. https://doi.org/10.1007/s44206-024-00095-1.

OECD (2022). Measuring the environmental impacts of artificial intelligence compute and applications: The AI footprint. *OECD Digital Economy Papers*, 341. Paris: OECD Publishing. https://doi.org/10.1787/7babf571-en.

OECD (2023). *Recommendation of the Council on Artificial Intelligence*. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

Panayides, et al. (2020). AI in medical imaging informatics: Current challenges and future directions. *IEEE Journal of Biomedical and Health Informatics*, *24*(7), 1837–1857. https://doi.org/10.1109/JBHI.2020.2991043.

Prainsack, B., & Forgó, N. (2024). New AI regulation in the EU seeks to reduce risk without assessing public benefit. *Nature Medicine*. https://doi.org/10.1038/s41591-024-02874-2.

Queudot, M., & Meurs, M. J. (2018). Artificial Intelligence and predictive justice: Limitations and perspectives. In M. Mouhoub, S. Sadaoui, & O. Ait Mohamed (Eds.), *Recent trends and future technology in applied intelligence*. Springer. https://doi.org/10.1007/978-3-319-92058-0_85.

Rakhymbayeva, N., Amirova, A., & Sandygulova, A. (2021). A long-term engagement with a social robot for autism therapy. *Frontiers in Robotics and AI*, *8*, 669972. https://doi.org/10.3389/frobt.2021.669972.

Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Soenksen, L. R., Kassis, T., Conover, S. T., Marti-Fuster, B., et al. (2021). Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images. *Science Translational Medicine*, *13*(581), eabb3652. https://doi.org/10.1126/scitranslmed.abb3652.

Tamburrini, G. (2022). The AI carbon footprint and responsibilities of AI scientists. *Philosophies*, *7*(1), 4. https://doi.org/10.3390/philosophies7010004.

Tanaka, F., Isshiki, K., Takahashi, F., Uekusa, M., Sei, R., & Hayashi, K. (2015). Pepper learns together with children: Development of an educational application. *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, 270–275. https://doi.org/10.1109/HUMANOIDS.2015.7363546.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., et al. (2023). Large language models in medicine. *Nature Medicine*, *29*, 1930–1940. https://doi.org/10.1038/s41591-023-02448-8.

Thywissen, K. (2006). Components of risk: a comparative glossary. *Source*, *2*. Bonn: UNU-EHS.

Tian, L., & Oviatt, S. (2021). A taxonomy of social errors in human-robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)*, *10*(2), 1–32. https://doi.org/10.1145/3439720.

Twomey, J., Ching, D., Aylett, M. P., Quayle, M., Linehan, C., & Murphy, G. (2023). Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine. *Plos One*, *18*(10), e0291668. https://doi.org/10.1371/journal.pone.0291668.

UNDRO. (1991). *Mitigating natural disasters. Phenomena, effects and options. A manual for policy makers and planners*. United Nations.

UNISDR (2017). *Natech Hazard and Risk Assessment.* https://www.undrr.org/quick/11674.

Van de Poel, I. (2016). An ethical framework for evaluating experimental technology. *Science and Engineering Ethics*, *22*(3), 667–686. https://doi.org/10.1007/s11948-015-9724-3.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Aidan, Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Verdecchia, R., Sallou, J., & Cruz, L. (2023). A systematic review of Green AI. *WIREs Data Mining and Knowledge Discovery*, *13*(4), e1507. https://doi.org/10.1002/widm.1507.

Vermaas, P., Kroes, P., Van de Poel, I., Franssen, M., & Houkes, W. (2011). *A philosophy of technology: From technical artefacts to sociotechnical systems.* Springer.

Wirtz, B. W., Weyerer, J. C., & Kehl, I. (2022). Governance of artificial intelligence: a risk and guideline-based integrative framework. *Government Information Quarterly*, *39*(4), 101685.

Woodhouse, E. J., & Collingridge, D. (1993). Incrementalism, intelligent trial-and-error, and political decision theory. In H. Redner (Ed.), *An heretical heir of the enlightenment: science, politics and policy in the work of Charles E. Lindblom* (pp. 131–154). Westview.

Wu, S., Irsoy, O., Lu, S. (2023). Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*

Xu, F. F., Alon, U., Neubig, G., & Hellendoorn, V. J. (2022). A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming (MAPS 2022)*. New York: Association for Computing Machinery, 1–10. https://doi.org/10.1145/3520312.3534862.

Zanotti, G., Petrolo, M., Chiffi, D., & Schiaffonati, V. (2023). Keep trusting! A plea for the notion of trustworthy AI. *AI & Society*. https://doi.org/10.1007/s00146-023-01789-9.