

Bernhard Steffen (Ed.)

LNCS 14129

Bridging the Gap Between AI and Reality

First International Conference, AISoLA 2023
Crete, Greece, October 23–28, 2023
Selected Papers

 Springer

OPEN ACCESS

Lecture Notes in Computer Science

14129


Founding Editors

Gerhard Goos
Juris Hartmanis

Editorial Board Members

Elisa Bertino, *Purdue University, West Lafayette, IN, USA*

Wen Gao, *Peking University, Beijing, China*

Bernhard Steffen , *TU Dortmund University, Dortmund, Germany*

Moti Yung , *Columbia University, New York, NY, USA*

The series Lecture Notes in Computer Science (LNCS), including its subseries Lecture Notes in Artificial Intelligence (LNAI) and Lecture Notes in Bioinformatics (LNBI), has established itself as a medium for the publication of new developments in computer science and information technology research, teaching, and education.


LNCS enjoys close cooperation with the computer science R & D community, the series counts many renowned academics among its volume editors and paper authors, and collaborates with prestigious societies. Its mission is to serve this international community by providing an invaluable service, mainly focused on the publication of conference and workshop proceedings and postproceedings. LNCS commenced publication in 1973.

Bernhard Steffen
Editor

Bridging the Gap Between AI and Reality

First International Conference, AISoLA 2023
Crete, Greece, October 23–28, 2023
Selected Papers

Editor

Bernhard Steffen 
TU Dortmund University
Dortmund, Germany



ISSN 0302-9743

ISSN 1611-3349 (electronic)

Lecture Notes in Computer Science

ISBN 978-3-031-73740-4

ISBN 978-3-031-73741-1 (eBook)

<https://doi.org/10.1007/978-3-031-73741-1>

© The Editor(s) (if applicable) and The Author(s) 2025. This book is an open access publication.

Open Access This book is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this book are included in the book's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

Preface

This volume contains the postproceedings of AISoLA 2023, an AI-themed sibling of ISoLA, the International Symposium on Leveraging Applications of Formal Methods. AISoLA took place in Crete (Greece) on October 23–28, 2023. It was an in-person event that provided an interdisciplinary forum for discussing the impact of the recent AI developments on research, education, and society. Discussions ranged from philosophical issues that arise from technologies as powerful as indicated by today’s large language models to technical issues and solutions for the responsible uses for AI-applications in safety-critical domains. The program of AISoLA 2023 comprised five keynotes:

Technology and Democracy

by Moshe Vardi, Rice University, USA

Deep Neural Networks, Explanations and Rationality

by Edward Lee, UC Berkeley, USA

Human or Machine: Reflections on Turing-Inspired Testing for the Everyday

by David Harel, Weizmann Institute of Science, Israel

Education and AI – Current Status, Opportunities and Challenges

by Nele McElvany, TU Dortmund, Germany

Graph Neural Networks: Everything Is Connected

Matthias Fey, Kumo.ai, USA

Special Sessions

Technology and Democracy

organized by Jim Larus and Edward Lee

Beyond ChatGPT: The Impact of AI on Academic Research

organized by Viola Schiaffonati

Tracks

1. The Nature of AI-Based Systems
organized by Bernhard Steffen
2. Responsible and Trustworthy AI
organized by Kevin Baum, Torsten Helfer, Markus Langer, Eva Schmidt, Andreas Sesing-Wagenpfeil, and Timo Speith
3. Democracy in the Digital Era
organized by George Metakides and Moshe Vardi
4. Digital Humanism
organized by Viola Schiaffonati and Hannes Werthner
5. Safety Verification of DNNs
organized by Taylor Johnson and Daniel Neider

6. Verification Meets Learning and Statistics
organized by Jan Kretinski, Kim Larsen, Nils Jansen, and Bettina Könighofer
7. Health Care
organized by Martin Leucker
8. AI-Assisted Programming
organized by Wolfgang Ahrendt and Klaus Havelund
9. Safe AI in the Automotive Domain
organized by Falk Howar and Hardi Hungar
10. Digital Humanities
organized by Ciara Breathnach and Tiziana Margaria
11. R@ISE: Research at ISE
organized by Tiziana Margaria and Mike Hinchey
12. AI-Supported Publishing
organized by Jonas Spies

The 26 papers of this volume extend the presentation in the AISoLA 2023 on-site proceedings. This means, in particular, that the corresponding track introductions still apply. Only the Health Care and the Digital Humanities track have some dedicated introductions in this volume.

I thank the track organizers, the members of the program committee and their reviewers for their effort in selecting the papers to be presented, the local Organization Chair, Petros Stratis, and the EasyConferences team for their continuous precious support during the entire period preceding the events, and Springer Nature for being, as usual, a very reliable partner for the proceedings production. Finally, I am grateful to Nicolas Stratis and Tim Tegeler for continuous support for the website and the program, and to Steve Bosselmann for his help with the editorial system EquinOCS.

Special thanks are due to the Center for Trustworthy Data Science and Security and the Lamarr Institute for their support in the organization of the event, and to the Technical University of Dortmund, my home institution.

With over 150 international participants and very lively discussions, AISoLA was a very successful event and I am looking forward to seeing many of you in Crete in October for AISoLA 2024.

October 2024

Bernhard Steffen

Organization

Program Committee

Ahrendt, Wolfgang	Chalmers University of Technology, Sweden
Baum, Kevin	Saarland University, Germany
Breathnach, Ciara	University of Limerick, Ireland
Havelund, Klaus	Jet Propulsion Laboratory, USA
Helfer, Thorsten	Saarland University, Germany
Howar, Falk	TU Dortmund, Germany
Hungar, Hardi	DLR Braunschweig, Germany
Jansen, Nils	Radboud Universit�at Nijmegen, Germany
Johnson, Taylor T.	Vanderbilt University, USA
Kretinsky, Jan	Brno University, Czech Republic
K�onighofer, Bettina	Graz University of Technology, Austria
Langer, Markus	University of Marburg, Germany
Larsen, Kim	Aalborg University, Denmark
Leucker, Martin	University of L�ubeck, Germany
Margaria, Tiziana	University of Limerick, and Lero, Ireland
Neider, Daniel	TU Dortmund, Germany
O'Shea, Enda	University of Limerick, Ireland
Schmidt, Eva	TU Dortmund, Germany
Sesing-Wagenpfeil, Andreas	Saarland University, Germany
Speith, Timo	University of Bayreuth, Germany
Steffen, Bernhard (Chair)	TU Dortmund, Germany

Reviewers

Ahrendt, Wolfgang	Chalmers University of Technology, Sweden
Aichernig, Bernhard	TU Graz, Austria
Bao, Tianshu	Guizhou University, China
Baum, Kevin	Saarland University, Germany
Baumann, Jonas	Saarland University, Germany
Bollig, Benedikt	ENS Cachan, France
Bongo, Lars Ailo	UiT, Arctic University of Norway, Troms�o, Norway
Breathnach, Ciara	University of Limerick, Ireland
Chaudhary, Hafiz Ahmad Awais	University of Limerick, Ireland

Ferilli, Stefano	University of Bari, Italy
Floris, Francesco	Università di Torino, Italy
Fradiente, Valeria	Università di Torino, Italy
Havelund, Klaus	Jet Propulsion Laboratory, USA
Helfer, Thorsten	CISPA Helmholtz Center for Information Security, Saarbrücken, Germany
Howar, Falk	TU Dortmund, Germany
Hungar, Hardi	DLR Braunschweig, Germany
Hösterey, Steffen	Humboldt-Universität zu Berlin, Germany
Langer, Markus	University of Freiburg, Germany
Le Roux, Maelle	University of Limerick, Ireland
Leucker, Martin	University of Lübeck, Germany
Margarita, Tiziana	University of Limerick, CSIS and Lero, Ireland
Mitwalli, Daniel Sami	University of Limerick, Ireland
Mourits, Rick	International Institute of Social History, Amsterdam, the Netherlands
Pedersen, Bjørn-Richard	UiT, Arctic University of Norway, Tromsø, Norway
Peled, Doron	Bar Ilan University, Israel
Sachenbacher, Martin	University of Lübeck, Germany
Schmidt, Eva	TU Dortmund, Germany
Schrills, Tim	University of Lübeck, Germany
Sesing-Wagenpfeil, Andreas	Saarland University, Germany
Singh, Amandeep	University of Limerick, Ireland
Speith, Timo	University of Bayreuth, Germany
Teumert, Sebastian	TU Dortmund, Germany
Thoma, Daniel	University of Lübeck, Germany
Walsh, Oonagh	Glasgow Caledonian University, UK
Xiang, Weiming	Augusta University, USA
Yalciner, Mustafa	Fraunhofer IML, Germany
Zafeiridi, Evi	University College Cork, Ireland

Contents

Digital Humanities

Digital Humanities and Cultural Heritage in AI and IT-Enabled Environments	3
<i>Ciara Breathnach and Tiziana Margaria</i>	
Common Language for Accessibility, Interoperability, and Reusability in Historical Demography	10
<i>Rick J. Mourits, Tim Riswick, and Rombert Stapel</i>	
Coding Historical Causes of Death Data with Large Language Models	30
<i>Bjørn-Richard Pedersen, Maisha Islam, Doris Tove Kristoffersen, Lars Ailo Bongo, Eilidh Garrett, Alice Reid, and Hilde Sommersth</i>	
Teaching the Specialized Language of Mathematics with a Data-Driven Approach: What Data Do We Use?	48
<i>Cecilia Fissore, Francesco Floris, Marina Marchisio Conte, and Matteo Sacchet</i>	
Interoperating Civil Registration of Death and Census Data: Old Age and Marriage as Categories of Analysis	65
<i>Ciara Breathnach, Rachel Murphy, Alexander Schieweck, and Tiziana Margaria</i>	
From Data Science to Modular Workflows Changing Perspectives from Data to Platform: DBDIrl 1864-1922 Case Study	84
<i>Enda O'Shea, Marco Krumrey, Daniel Sami Mitwalli, Sebastian Teumert, and Tiziana Margaria</i>	
Mapping Madness: HGIS and the Analysis of Irish Patient Records	104
<i>Oonagh Walsh and Stuart Clancy</i>	
Digitised Historical Sources and Non-digital Humanists: An Interdisciplinary Challenge?	119
<i>Maelle Le Roux and Anna Gasperini</i>	
Using Passive Sensing to Identify Depression	132
<i>Evi Zafeiridi, Malik Muhammad Qirtas, Eleanor Bantry White, and Dirk Pesch</i>	

The GraphBRAIN Framework for Knowledge Graph Management and Its Applications to Cultural Heritage 144
Stefano Ferilli, Eleonora Bernasconi, Davide Di Pierro, and Domenico Redavid

Health Care

Challenges for AI in Healthcare Systems 165
Markus Bertl, Yngve Lamo, Martin Leucker, Tiziana Margaria, Esfandiar Mohammadi, Suresh Kumar Mukhiya, Ludwig Pechmann, Gunnar Piho, and Fazle Rabbi

Towards a Multi-dimensional Health Data Analysis Framework 187
Fazle Rabbi, Bahareh Fatemi, Suresh Kumar Mukhiya, and Yngve Lamo

Future Opportunities for Systematic AI Support in Healthcare 203
Markus Bertl, Gunnar Piho, Dirk Draheim, Peeter Ross, Ludwig Pechmann, Nicholas Bucciarelli, and Rahul Sharma

CRISP-PCCP – A Development Methodology Supporting FDA Approval for Machine Learning Enabled Medical Devices 225
Ludwig Pechmann, Yannik Potdevin, Kai Brehmer, Dirk Nowotka, and Martin Leucker

Model Driven Development for AI-Based Healthcare Systems: A Review 245
Colm Brandon, Amandeep Singh, and Tiziana Margaria

Responsible and Trustworthy AI

Balancing Transparency and Risk: An Overview of the Security and Privacy Risks of Open-Source Machine Learning Models 269
Dominik Hintersdorf, Lukas Struppek, and Kristian Kersting

AI-Related Risk and Uncertainty 284
Giacomo Zanotti, Daniele Chiffi, and Viola Schiaffonati

Leveraging Actionable Explanations to Improve People’s Reactions to AI-Based Decisions 293
Markus Langer and Isabel Valera

From Explanation Correctness to Explanation Goodness: Only Provably Correct Explanations Can Save the World 307
Maike Schwammbeger

Thinking Outside the Box?: Regulatory Sandboxes as a Tool for AI Regulation 318
Hannah Ruschemeier

AI and Democratic Equality: How Surveillance Capitalism and Computational Propaganda Threaten Democracy 333
Ashton Black

AI Assisted Programming

Safeguarding AI-Based Software Development and Verification Using Witnesses (Position Paper) 351
Dirk Beyer

End-to-End AI Generated Runtime Verification from Natural Language Specification 362
Itay Cohen and Doron Peled

AI-Assisted Programming with Test-Based Refinement 385
Bernhard K. Aichernig and Klaus Havelund

Automotive Driving

Safer Than Perception: Increasing Resilience of Automated Vehicles Against Misperception 415
Martin Fränzle and Andreas Hein

Towards ML-Integration and Training Patterns for AI-Enabled Systems 434
Sven Peldszus, Henriette Knopp, Yorick Sens, and Thorsten Berger

Safety Verification of DNNs

The Reachability Problem for Neural-Network Control Systems 455
Christian Schilling and Martin Zimmermann

Author Index 471



AI-Related Risk and Uncertainty

Giacomo Zanotti¹(✉), Daniele Chiffi², and Viola Schiaffonati¹

¹ Department of Electronics, Information and Bioengineering (DEIB), Politecnico di Milano, Milan, Italy

giacomo.zanotti@polimi.it

² Department of Architecture and Urban Studies (DAStU), Politecnico di Milano, Milan, Italy

Abstract. Discussions on the risks involved in the deployment of AI systems are increasingly prominent in both public discourse and scientific debates. While talk of risk plays a crucial role in framing ethical and societal problems related to AI, we argue that it could profitably be associated with a clear analysis of uncertainty. Starting from a multi-component approach to AI-related risk assessment and mitigation, this chapter discusses the way the deployment of AI systems often takes place in contexts in which uncertainty is not meaningfully quantifiable.

Keywords: AI · Risk · Uncertainty · Philosophy of risk

1 Introduction

Recent advances in the field of Artificial Intelligence (AI) have resulted in a widespread diffusion of AI systems to be applied for significantly heterogeneous purposes in a wide range of situations. In many cases, these systems are delegated with complex tasks that would typically require human intervention. What is more, they are increasingly employed in delicate contexts in which their decisions, predictions and classifications can have a significant impact on people's life. Most notably, we can think about the fields of medical AI and predictive justice, or systems employed for loan processing and autonomous driving.

With things being this way, the growing prominence of the notion of risk in discussions on the ethical and social implications of AI does not come as a surprise. On the one hand, a fair deal of literature and public discourse has been focusing on the so-called *existential* risks related to the deployment of AI systems, often involving human extinction or global catastrophes. On the other hand, usually in open contrast with the talk on existential risk, increasing attention has been devoted to more mundane forms of AI-related risk.¹ This latter approach – which is also the one behind this contribution – has led, among other things, to the recently approved European proposal for the first comprehensive regulation on AI – the so-called AI Act² – where systems are classified

¹ <https://www.nature.com/articles/d41586-023-02094-7>, last accessed 2024/04/04.

² 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, last accessed 2024/04/04).

in accordance to their level of risk (unacceptable, high, limited, minimal) and regulated accordingly.

Notwithstanding its importance, the assessment of risk may have some limitations when it comes to the outcomes and implications of AI systems. In fact, risk is understood as a normative notion associated with potential negative consequences, and it is often characterized by a distinct probabilistic component. More specifically, when referring to the risk of an event x , we typically imply that we can meaningfully assign a probabilistic value to the occurrence of x . This possibility is not always feasible when it comes to AI systems and the potential effects of their use. This contribution contends that talk of risk in AI should make room for the notion of uncertainty, both in quantifiable and unquantifiable forms. While our analysis is distinctively philosophical in scope and methodology, we believe it may be used as a theoretical ground for devising risk assessment practices in AI.

In Sect. 2, the notion of risk is presented, paying particular attention to multi-component approaches to risk and their quantifiable uncertainties, with a specific focus on AI-related risk. Then, in Sect. 3, the notion of severe uncertainty is introduced as a way to better understand and assess those cases where uncertainty cannot be meaningfully quantified. We will focus on the use of general-purpose AI systems as paradigmatic examples of a context in which this dimension of severe uncertainty is particularly relevant. Section 4 concludes the chapter by discussing possible future lines of research.

2 Risk and Its Components

Providing a univocal characterization of risk is not an easy task, for non-technical understandings of this notion come together with a number of technical definitions. Among these, the one provided by the Royal Society in 1983 is often referred to as the “classic” one, equating risk with “the probability that a particular adverse event occurs during a stated period of time, or results from a particular challenge” (Royal Society, 1983). Needless to say, it does not all come down to probability. As a matter of fact, assessing risk typically involves some form of expectation in which the probability of the unwanted event becomes the weight for the magnitude of its consequences: a higher magnitude might counterbalance a lower probability of occurrence, and *vice versa*. Still, probability is usually required in many definitions of risk.

Among other things, this way of understanding risk seems to be at the basis of the AI Act, that explicitly defines risk as “the combination of the probability of an occurrence of harm and the severity of that harm” (Art. 3, 2). However, it is not the only way to approach risk, in particular when it comes to designing risk-mitigation policies and interventions. Most notably, approaches adopted in the domain of disaster risk mitigation understand risk as the result of the interaction between three different components: hazard, exposure, and vulnerability (UNISDR, 2015). *Hazard* refers to the source of potential harm, *exposure* to the people and resources that could be harmed, and *vulnerability* has to do with how much what is exposed is susceptible to the impacts of the hazard. As an example, consider seismic risk. In this case, the hazard component refers to the earthquake itself, and its assessment involves estimates concerning both the probability and the magnitude of the earthquake. When it comes to exposure, instead, we

focus on what could be harmed by the earthquake, considering both people and material assets (e.g., buildings and infrastructures) that are found in the seismic hazard zone. Finally, one should take into account the circumstances and measures that could render these individuals and assets more or less susceptible to the potential harm: in the case of an earthquake, relevant elements could be the seismic safety standards of the potentially affected buildings, the existence of response plans, and the availability of temporary shelters.

Distinguishing between the different components we have just seen allows us to intervene on several fronts to reduce risk. Now, reducing the hazard is not always possible, especially in the case of some *natural* risks: we simply cannot prevent an earthquake from occurring. However, there are many cases, especially those in which the hazard is related to human action, in which there is much we can do (e.g., we might relocate polluting factories away from population centers, or withdraw from the market a potentially dangerous technology). At the same time, we can intervene on the exposure. In the case of seismic risk, the most straightforward way to do this involves limiting the number of people and assets in the areas that are more likely to be affected by earthquakes. Finally, efforts can be made to reduce the vulnerability of such people and assets by intervening on buildings to improve their safety, designing evacuation plans, and so on.

While the domain of natural risk offers intuitive examples of how different risks can be better analyzed and managed by distinguishing their components, nothing prevents us from applying the same kind of analysis to technological risks – that is, risks stemming from the use of technological artifacts.³ AI systems make no exception. On the contrary, thinking of AI-related risk through the conceptual and methodological lens of multi-component analyses of risk allows us to understand how and why significantly different kinds of AI systems involve non-negligible levels of risk.⁴

In some cases, AI systems strike us as involving considerable levels of risk as a result of the hazard's magnitude. Let us take a look at the AI Act's Annex III, listing (some of) the systems that are considered as "high-risk" within the scope of the Act, and are therefore subject to stricter regulation. Among these, we can find AI systems that serve as "safety components in the management and operation of critical digital infrastructure, road traffic and the supply of water, gas, heating and electricity", or systems used by law enforcement. It is fairly straightforward that malfunctions in such systems directly result in potentially harming events. The failure of an AI system used to manage road traffic can result in life-threatening accidents, and a system used to predict recidivism in courts can be affected by biases that may ultimately result in unfair judgments and unjustified detention (Angwin *et al.*, 2016). In these cases, regardless of the levels of exposure and vulnerability, the fact that these systems involve high levels of hazard seems to be enough for labeling them as "high-risk".

The reasoning is diametrically opposite when it comes to those systems that qualify as highly risky due to their affecting and/or being used by a considerable number of people. In this regard, recommender systems are the most prominent example, especially those

³ Note that the dichotomy between natural and technological risks is not meant to be always completely exhaustive (Hansson, 2016).

⁴ For a detailed analysis of a multi-component approach to AI-related risk, see (Zanotti, Chiffi, Schiaffonati, 2024).

implemented in the so-called *very large online platforms* (VLOPs).⁵ In these cases, the potentially low levels of hazard and vulnerability are negatively counterbalanced by significantly high levels of exposure.

Finally, some systems might qualify as high-risk as a result of the vulnerability of their users. Examples abound. For instance, AI-systems – including intelligent robotic systems – are increasingly used in the context of education and elderly care (Miyagawa *et al.*, 2019; Tanaka *et al.*, 2015). In these cases, even assuming low levels of hazard and exposure, the vulnerability of the people using or being affected by AI systems compels us to guard against potential unwanted outcomes and accordingly treat the involved technologies as high-risk ones.

We have now seen how adopting a specific approach to risk, namely a multi-component analysis, can help us better understand AI-related risk. The two general features of risk that have been partially anticipated, however, remain valid. First of all, risk refers to the possible occurrence of an *unwanted event*, of something that is negatively valued. This is immediately evident in the classic definition of risk, that explicitly refers to *adverse* events, and it is clear in multi-component analyses of risk, that understand hazard as the source of *harm*. Accordingly, when referring to AI-related risk, we always focus on the *negative* potential consequences of AI systems' deployment. The second feature that typically characterizes conceptions of risk is that they involve the possibility of a meaningful probabilistic evaluation of the unwanted events in question.⁶

Sometimes, probabilistic risk assessment is assumed and conducted by using point-like probabilistic values, since we trust such probabilities. This can be a good choice when the uncertainty and complexity of the risk are not particularly noteworthy. This is what typically happens in textbook cases and idealized scenarios: if we bet on dice games and the dice is a fair one, we know exactly which our risk of losing is. However, more commonly some quantifiable forms of uncertainty are acknowledged within risk, and this is why risks may be quantified and evaluated by means of probabilistic intervals, second-order probabilities, imprecise probabilities, belief-functions, possibility theories, and fuzzy logic, just to mention some of these methods (Hansson, 2018; Dencœur *et al.*, 2020a, 2020b). AI makes no exception. On the contrary, providing point-like probabilities may be hard in the case of AI systems' deployment, for such systems are often used in complex contexts in which unanticipated circumstances might influence the course of events, and their being often relatively new technologies may result in a paucity of data concerning their use and its possible negative outcomes.

3 AI-Related Risks and Severe Uncertainty

Taking stock, we have seen how the notion of risk is associated with the possibility of making probabilistic estimates about unwanted events and their outcomes. True, some components of uncertainty are typically involved in real-world scenarios, for it is often hard to assess risk by means of point-like probabilistic values, and AI makes no exception. Still, the uncertainty in question can be quantified. However, this is not always the case:

⁵ According to the European *Digital Services Act*, a platform qualifies as a VLOP if it has more than 45 million users per month in the EU (DSA, 2022).

⁶ In the literature, these situations are understood as “known unknowns” (Hansson, 2009).

several types of uncertainty exist, and not all of them can be meaningfully quantified (Hansson, 2022). In this section, we analyze how non-quantifiable uncertainty may play a role in the assessment of AI-related risk. Note that, while we focus on the way risk (and quantifiable forms of uncertainty) differ from non-quantifiable forms of uncertainty with respect to our probabilistic knowledge of possible scenarios, other differences exist. Most notably, we have seen how risk is a normative and evaluative concept with a negative connotation. This does not always happen with uncertainty. On the contrary, some forms of uncertainty are usually assumed to be possible triggers for technological innovation (Chiffi, Moroni, Zanetti, 2022).

Based on what we have seen in the previous section, risk assessment seems to depend on our evaluation of the potential unwanted events in question, their consequences and contexts of occurrence. For instance, in the case of seismic risk, assessing exposure requires to possess reliable knowledge about the location and extension of the potentially affected area as well as the number of people, buildings and infrastructures therein. In addition, up-to-date information concerning (among other things) the existence of evacuation plans and buildings' safety standards is needed to evaluate the vulnerability of exposed people and assets. All of this applies to the case of AI. Suppose you want to estimate the risk associated with the deployment of a certain AI system. First, you need to identify possible inaccuracies, malfunctions, misuses, and more generally all unintended and unwanted consequences resulting from the deployment of the system, and possibly associate them with a probability. Then, you must have a sufficiently precise idea of the people and assets exposed to such consequences. Finally, you should be able to assess their vulnerability by considering all those factors and circumstances that make them more or less prone to be harmed by the potential events in question.

While this might be doable for some AI systems and in some contexts (e.g., AI systems based on symbolic techniques to be used in controlled environments), it is not always possible. In some cases, it might be hard to make predictions on the possible *inaccuracies* and *malfunctions* of AI systems, often due to their complexity and working opacity. In addition to this, we might not be able to anticipate their possible uses, and therefore their *misuses*, and identify who could be affected by their negative outcomes. As we will see in a moment, such difficulties might be due to the fact that some kinds of AI systems can be adapted to a wide variety of uses and applications. On top of that, we should keep in mind that, in many cases, the technologies we are referring to are relatively recent, and we largely lack data on their real-world use that could inform our predictions.

In the literature, analogous situations are captured through the notion of *severe uncertainty*. Severe uncertainty is typically conceived in open contraposition to probabilistic conceptualizations of risk such as the Royal Society's one we have seen in Sect. 2. Consider the (fair) dice game example. In this case, we have exhaustive and reliable knowledge of both (i) the possible outcomes of the roll of the dice and (ii) the probability associated with each outcome.

In situations of severe uncertainty, things are less clear. For a specific set of events, we might be able to anticipate the possible outcomes while ignoring their probability distribution. Many of the recent and most impactful AI technologies seem to be used in and give rise to contexts of severe uncertainty. The example we propose to consider is that

of so-called *general-purpose AI systems* (GPAIs). This expression, that for the purpose of this chapter we take to be largely overlapping with the one of *foundation models* (Bommasani *et al.*, 2022), refers to any AI system that can “accomplish or be adapted to accomplish a range of distinct tasks, including some for which it was not intentionally and specifically trained” (Gutierrez *et al.*, 2023). The class of GPAIs includes different models and systems, from those designed for computer vision to those for multimodal processing. Among these, however, Large Language Models (LLMs) are increasingly widespread, especially after OpenAI’s decision in November 2022 to implement their model GPT3.5 in a freely available chatbot with a user-friendly interface. From that moment on, different companies and developers rushed to offer easily accessible LLM-based platforms at users’ fingertips.

When it comes to these systems, assessing risk is particularly difficult. First of all, the identification of malfunctions, misuses, and unintended consequences might be quite critical. As a matter of fact, their being general-purpose models, so capable of tasks for which they have not been specifically designed and trained, makes it very difficult to anticipate all the potential consequences of their use. Moreover, the fact that these systems are most of the time running proprietary software (not an open source one) further exacerbates the possibility to predict malfunctions. True, many possible scenarios of malfunctions and abuses can be foreseen. For instance, once we know that certain GPAIs can be used for code generation, we can easily anticipate that someone may jailbreak them to write malware. However, it is not clear how we could associate a probability to this scenario before the system’s large-scale deployment.

Analogous considerations can be made when it comes to estimating the exposure component of the risks involved in the deployment of GPAIs. Many GPAIs are now implemented in free and accessible platforms, and the number of people making use of these systems in their daily life is increasing – again, their flexibility makes them potentially applicable to significantly different tasks and in a wide range of situations. Such an evolving scenario also makes it very difficult to have a sufficiently precise idea of the people exposed to their consequences.

Finally, in light of this, it is not hard to see the difficulties involved in the attempt to estimate the component of vulnerability associated with these systems’ risk. To do so, as a matter of fact, we should be able to identify both the potential harmful uses of GPAIs as well as those affected by their possible negative consequences. And again, this is not an easy task.

Summing up, we could say that the extreme flexibility of some AI systems, GPAIs in particular, plays a major role in giving rise to severe forms of uncertainties: as their possible uses are wide and open, it is hard to anticipate and assess all of them and thereby estimate the associated levels of hazard, exposure and vulnerability. These forms of uncertainty are hardly quantifiable and represent a significant challenge in assessing AI-related risk, but cannot be overlooked in a rigorous and complete discussion of AI technologies and their societal implications.

4 Conclusion

We discussed some possible difficulties in assessing the risks associated with the use of AI systems. Starting from a focus on the components of risk, namely, hazard, exposure, and vulnerability, we highlighted that traditional risk analysis often relies on probabilistic information, which may not be always readily available or reliable for the outcomes of AI systems' deployment. We suggested that incorporating the concept of uncertainty into AI-related risk analysis is beneficial not only when uncertainty is quantifiable but also, and more importantly, when it is not quantifiable. This is particularly relevant in cases of severe forms of uncertainties. We explored general-purpose AI systems as an illustrative example of technology where severe uncertainty may play a pivotal role in risk assessment. Among other things, this uncertainty arises due to the considerable flexibility in these systems' potential applications.

In future lines of research, we will investigate the role of multi-risk analysis related to AI, wherein various risks may interact mutually, potentially producing domino or cascade effects⁷. To this end, we will draw upon the rich literature on engineering safety, risk assessment and uncertainty (e.g., Burton, Mcdermid, Freng, 2023), in particular in the context of AI (e.g., NIST, 2023). We will also explore the impact of unforeseeable events, sometimes referred to as “unknown unknowns,” on AI-related risks. These events can be challenging not only to quantify but also to predict accurately and are typically associated with socio-technical systems, which may pose wicked problems to society – complex issues often intertwined with policy and planning (Rittel & Webber, 1973; Nordström, 2022). Such problems are difficult to address and even analytically define. A rigorous epistemological analysis of uncertainty in AI, however, will hopefully put us in a better position to deal with them.

Acknowledgments. This study was funded by (1) the Italian Ministry of University and Research under the PRIN Scheme (Project BRIO, no. 2020SSKZ7R; Project NAND no. 2022JCMHFS); (2) RETURN, Extended Partnership, Multi-risk science for resilient communities under a changing climate, European Union Next-GenerationEU (National Recovery and Resilience Plan – NRRP, Mission 4, Component 2, Investment 1.3 – D.D. 12432/8/2022, PE0000005); (3) PNRR-PE-AI FAIR-NextGeneration EU program.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

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

⁷ In fact, increasing attention has been devoted to the so-called “Natech” risks. Natech stands for “natural hazard triggering technological accident” (Mesa-Gómez, Casal, Muñoz, 2020). A notorious example of a Natech accident was the nuclear one that took place in Ōkuma, in the Japanese prefecture of Fukushima, in March 2011, when a tsunami hit the Fukushima Daiichi nuclear power plant causing a failure in the electric grid and damaging backup generators, which ultimately resulted in a leak of radioactive contaminants.

- Bommasani, R., et al.: On the opportunities and risks of foundation models (2022). arXiv preprint [arXiv:2108.07258](https://arxiv.org/abs/2108.07258)
- Burton, S., Mcdermid, J., Freng, O.: Closing the gaps: complexity and uncertainty in the safety assurance and regulation of automated driving (2023). <https://www.iks.fraunhofer.de/content/dam/iks/documents/whitepaper-closing-the-gaps.pdf>
- Chiffi, D., Moroni, S., Zanetti, L.: Types of technological innovation in the face of uncertainty. *Philos. Technol.* **35**(4), 94 (2022)
- Dencœur, T., Dubois, D., Prade, H.: Representations of uncertainty in artificial intelligence: probability and possibility. In: Marquis, P., Papini, O., Prade, H. (eds.) *A Guided Tour of Artificial Intelligence Research*, pp. 69–117. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-06164-7_3
- Dencœur, T., Dubois, D., Prade, H.: Representations of uncertainty in AI: beyond probability and possibility. In: Marquis, P., Papini, O., Prade, H. (eds.) *A Guided Tour of Artificial Intelligence Research*, pp. 119–150. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-06164-7_4
- Gutierrez, C.I., Aguirre, A., Uuk, R., Boine, C.C., Franklin, M.: A proposal for a definition of general purpose artificial intelligence systems. *Digit. Soc.* **2**(3), 36 (2023)
- Mesa-Gómez, A., Casal, J., Muñoz, F.: Risk analysis in Natech events: state of the art. *J. Loss Prev. Process Ind.* **64**, 104071 (2020)
- Miyagawa, M., et al.: Consideration of safety management when using Pepper, a humanoid robot for care of older adults. *Intell. Control Autom.* **11**, 15–24 (2019)
- Hansson, S.O.: From the casino to the jungle: dealing with uncertainty in technological risk management. *Synthese* **168**(3), 423–432 (2009)
- Hansson, S.O.: Managing risks of the unknown. In: Gardoni, P., Murphy, C., Rowell, A. (eds.) *Risk Analysis of Natural Hazards*, pp. 155–172. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-22126-7_10
- Hansson, S.O.: Representing uncertainty. In: Hansson, S., Hendricks, V. (eds.) *Introduction to Formal Philosophy*, pp. 387–400. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-77434-3_19
- Hansson, S.O.: Can uncertainty be quantified? *Perspect. Sci.* **30**(2), 210–236 (2022)
- NIST AI 100-1: Artificial Intelligence Risk Management Framework (AIRMF 1.0) (2023). <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>
- Nordström, M.: AI under great uncertainty: implications and decision strategies for public policy. *AI Soc.* **37**(4), 1703–1714 (2022)
- Rittel, H.W., Webber, M.M.: Dilemmas in a general theory of planning. *Policy. Sci.* **4**(2), 155–169 (1973)
- Regulation 2022/2065 of the European Parliament and of the Council of 19 Oct. 2022, on a Single Market for Digital Services and Amending Directive 2000/31/EC (Digital Services Act – DSA). <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32022R2065>. Accessed 04 Apr 2024
- Tanaka, F., Isshiki, K., Takahashi, F., Uekusa, M., Sei, R., Hayashi, K.: Pepper learns together with children: development of an educational application. In: 2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids), Seoul, pp. 270–275. IEEE (2015)
- United Nations Office for Disaster Risk Reduction (UNISDR): UNISDR Annual Report 2015: 2014–15 Biennium Work Programme Final Report, Geneva (2015). https://www.unisdr.org/files/48588_unisdrannualreport2015evs.pdf. Accessed 04 Apr 2024
- Zanotti, G., Chiffi, D., Schiaffonati, V.: AI-related risk: an epistemological approach. *Philos. Technol.* **37**, 66 (2024)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

