it will prominently feature many of the classes of techniques we have covered in this book, such as feature attributions, example-based explanations, and counterfactuals.

What to Look Forward To in Explainable AI

What most excites us about the future of explainability is how it will move from individual, narrowly focused techniques to ones that generate richer explanations with less configuration needed in advance. Broadly, there are three trends to keep an eye on in the future of Explainable AI: natural and semantic explanations, interrogative explanations, and more targeted explanations.

Natural and Semantic Explanations

We often find that many types of explanations remain too technical or abstract for nontechnical users. An array of numbers representing feature attributions isn't necessarily very helpful for these users. Instead, techniques that can either present explanations in a more natural way, perhaps via a generative text model to create fluid sentences for the explanation, or are able to generate explanations based on a semantic understanding of the model and its dataset, will be much more helpful. Imagine if instead of being presented with an array of feature attribution values for a weather prediction, a user could be told, "There is an 80% chance of rain this afternoon because the temperature has dropped significantly and humidity is above 90%."

Semantic explanations, which require an even more innate understanding of the behaviors and concepts in the ML system, will also represent a large change in how we explain AIs. For example, an explainability technique may recognize that many of the similar examples where an ML classified a dog as a cat were due to poor lighting and low resolution in the photos. Instead of trying to highlight the pixels, where, for example, poor resolution or poor lighting may result in vague pixel attributions, it could categorically identify more pervasive causes for why the model failed.

Interrogative Explanations

Today, explanations are a one-way dialogue from the ML to the consumer. It is not unusual for someone to receive an explanation and immediately have more questions (see also the discussion in Chapter 7 on the human-interaction components of building explainable ML systems and "How to Effectively Present Explanations" on page 206). However, with current methods, that often requires an ML practitioner to roll up their sleeves and implement a new type of explanatory technique or perform additional types of explanations that were not expected before. A better ecosystem of Explainable AI tools will help make this job easier but will not solve the problem.

Instead, we expect a wave of second-generation explanation AI techniques that enable a richer experience where a user can query the ML for further information about a

prediction or behavior and can even guide the user in improving their understanding. Imagine this as a conversation between the consumer, perhaps a regulator, and the AI about a credit-rating ML:

> *Regulator*: Why was this individual given a credit rating of 520?
>
> *AI:* The most influential features were the high limits on the individual's credit cards, which caused the model to decrease its rating; their history of missed loan payments further drove the rating down.
>
> *Regulator:* Are individuals who live in similar zip codes (often an indirect variable for race in the US) with missed loan payments penalized as much as others?
>
> *AI:* No. Also, examining what the model considers to be 1,000 most similar people to this individual, there is no correlation with the zip code. Similar individuals who paid loans on time 25% more often on average had an increase of 50 points in their credit rating.

Targeted Explanations

As of 2022, there has been little work performed on assessing whether explanations follow the rule of Occam's razor: that the simplest explanation is the best. We expect that more robust explanations will be those that are more concise and targeted. For example, Local Explanations via Necessity and Sufficiency,[2], lays the foundation for these types of explanations by demonstrating how the minimal amount of perturbation necessary to flip a prediction provides an optimal explanation. These types of explanations will do much to address the brittleness problem described in Chapter 7, and will also take us a step further toward causal explanations.

Summary

In this chapter, we discussed how to design ML solutions with explainability in mind to build more reliable ML systems and provided a look toward the future of XAI. We've seen how Explainable AI techniques can be incorporated into each step of the ML life cycle, from discovery to development to deployment, assisting in building more robust ML solutions. We encourage you to think about XAI as a toolkit for better understanding machine learning models. We also provided a glimpse into what the future of XAI might hold and current research efforts.

Now, with these techniques and an understanding of how and where to apply them, you can improve both the models themselves and how your consumers work with them. Explainability is a rapidly changing field; we encourage you to view new techniques with optimism, but also give them some time to prove their worth in the

2 David Watson et al., "Local Explanations via Necessity and Sufficiency: Unifying Theory and Practice," arXiv, 2021.

CHAPTER 5

Future Trends and Challenges in XAI

Abstract

This chapter delves into the future trends and challenges in explainable artificial intelligence (XAI). It discusses the advances in XAI research, focusing on emerging techniques and methodologies aimed at improving the interpretability and transparency of AI systems. Ethical and regulatory considerations related to XAI are explored, highlighting the importance of addressing issues such as fairness, accountability, and privacy in AI-driven decision-making processes.

Furthermore, the chapter outlines the road ahead for XAI, emphasizing the need for interdisciplinary collaboration, stakeholder engagement, and responsible AI deployment. It explores potential opportunities and challenges in advancing XAI technologies across various domains, including healthcare, finance, autonomous systems, and recommender systems. Overall, the chapter provides insights into the evolving landscape of XAI and its implications for the future of artificial intelligence and society [1].

Keywords: Explainable AI, interpretability, transparency, ethical considerations, regulatory compliance, future trends

5.1 Advances in Explainable AI Research

Explainable artificial intelligence (XAI) has emerged as a critical area of research aimed at improving the transparency, interpretability, and

trustworthiness of AI systems. As AI technologies continue to evolve and permeate various aspects of society, the ability to understand and interpret AI-driven decisions becomes increasingly important. This article explores recent advancements in XAI research, highlighting innovative techniques, methodologies, and applications that enhance transparency and interpretability in AI systems [1].

5.1.1 Advancements in explainable AI research

1. Model-specific interpretability techniques: Recent research has focused on developing modelspecific interpretability techniques tailored to different types of machine learning models, including deep neural networks, decision trees, and support vector machines. These techniques aim to elucidate the internal workings of AI models, providing insights into how they arrive at their decisions. For example, visualization methods such as saliency maps and activation maximization techniques help visualize the features and patterns learned by deep neural networks, enabling stakeholders to understand the factors influencing model predictions [1].
2. Model-agnostic interpretability approaches: Model-agnostic interpretability approaches aim to provide transparency and interpretability for a wide range of machine learning models, regardless of their underlying architecture or complexity. Techniques such as feature importance analysis, permutation importance, and partial dependence plots help identify the most influential features and their impact on model predictions. By decoupling interpretability from specific model architectures, model-agnostic approaches offer flexibility and generality, enabling stakeholders to interpret AI-driven decisions across diverse applications and domains [1].

3. Explainable deep learning: Deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have achieved remarkable success in various tasks, including image recognition, natural language processing, and speech recognition. However, their black-box nature poses challenges for understanding and interpreting their decisions. Recent research in explainable deep learning focuses on developing techniques to improve the transparency and interpretability of deep neural networks. For example, layer-wise relevance propagation (LRP) decomposes the network's output to attribute relevance scores to input features, providing insights into the regions of input data that influence model predictions [1].
4. Counterfactual explanations: Counterfactual explanations offer a novel approach to XAI by providing alternative scenarios or explanations for AI-driven decisions. These explanations highlight how changes in input features would affect model predictions, enabling stakeholders to understand the sensitivity of AI models to different inputs. Counterfactual explanations are particularly useful in sensitive domains such as healthcare and finance, where understanding the factors driving model predictions is critical for decision making. For example, in medical diagnosis, counterfactual explanations can help physicians understand why a certain diagnosis was made and explore alternative treatment options based on hypothetical scenarios [1].
5. Human–computer interaction: Advancements in XAI research also focus on improving the interaction between humans and AI systems to facilitate better understanding and trust. Interactive visualization tools, user-friendly interfaces, and natural language explanations enable
6. stakeholders to interact with AI models intuitively and explore the underlying rationale behind model predictions. Human–computer

interaction techniques such as user feedback and iterative refinement help bridge the gap between AI systems and end-users, fostering collaboration and trust in AI-driven decision-making processes [1].

5.1.2 Applications of explainable AI research

The advancements in XAI research have significant implications for various applications and domains, including:

1. Healthcare: Explainable AI techniques enable physicians to interpret medical imaging results, diagnose diseases, and recommend treatment options with confidence. By providing transparent insights into AI-driven decisions, XAI enhances trust and collaboration between healthcare professionals and AI systems, leading to better patient outcomes and improved healthcare delivery.
2. Finance: In the finance industry, explainable AI research helps financial institutions interpret credit decisions, assess risk factors, and comply with regulatory requirements. By providing transparent explanations for AI-driven decisions, XAI enhances regulatory compliance, reduces bias and discrimination, and improves accountability in financial decision-making processes.
3. Autonomous systems: XAI techniques enhance the transparency and interpretability of AI-driven algorithms used in autonomous systems such as self-driving cars, drones, and robots. By providing insights into the factors influencing decision-making processes, XAI enables stakeholders to understand, validate, and trust autonomous systems, leading to safer and more reliable operation in real-world environments.
4. Recommender systems: Explainable AI research improves the transparency and interpretability of recommender systems used in e-

commerce, social media, and content platforms. By providing transparent explanations for recommendation decisions, XAI enhances user trust, satisfaction, and engagement, leading to better personalized recommendations and user expe-riences.

5.1.3 Challenges and future directions

While advancements in explainable AI research have made significant progress, several challenges and future directions remain:

1. Scalability and efficiency: XAI techniques must be scalable and efficient to handle large-scale datasets and complex AI models effectively. Addressing scalability and efficiency challenges requires developing computationally efficient algorithms and frameworks for XAI that can scale to real-world applications and deployment scenarios [1, 2, 3].
2. Interpretability–accuracy trade-offs: There is often a trade-off between the interpretability and accuracy of AI models, with more interpretable models sacrificing predictive performance or generalization. Balancing the trade-offs between interpretability and accuracy requires developing hybrid approaches that combine the transparency of interpretable models with the predictive power of complex AI models.
3. Human–AI collaboration: Enhancing human–AI collaboration is essential for realizing the full potential of XAI in real-world applications. Future research should focus on designing humancentric XAI systems that empower users to interact with AI models effectively, provide meaningful feedback, and make informed decisions based on transparent explanations.
4. Regulatory and ethical considerations: Addressing regulatory and ethical considerations is critical for ensuring responsible and ethical

deployment of XAI technologies. Future research should focus on developing ethical guidelines, standards, and frameworks for XAI that promote fairness, transparency, accountability, and privacy in AI-driven decision-making processes [4].

Advancements in explainable AI research hold great promise for enhancing transparency, interpretability, and trust in AI systems across various applications and domains. By providing transparent explanations for AI-driven decisions, XAI enables stakeholders to understand, validate, and trust AI models, leading to better decision-making processes, improved user experiences, and enhanced societal impact. As the field continues to evolve, interdisciplinary collaboration, regulatory alignment, and stakeholder engagement will drive the development and adoption of XAI, ultimately shaping the future of artificial intelligence and society [5, 6, 7].

5.2 Ethical and Regulatory Considerations

As artificial intelligence (AI) technologies continue to advance and permeate various aspects of society, ethical and regulatory considerations have become increasingly important. AI systems have the potential to bring about significant benefits, but they also raise complex ethical dilemmas and regulatory challenges. This chapter explores the ethical and regulatory considerations surrounding AI, examining key issues, guidelines, and frameworks aimed at promoting responsible AI deployment and mitigating potential risks [1].

5.2.1 Ethical considerations (Figure 5.1)

Figure 5.1: Ethical considerations.

Fairness and bias: AI systems can inadvertently perpetuate biases present in the data used for training, leading to unfair treatment and discrimination against certain groups. Addressing fairness and bias in AI requires careful attention to data collection, algorithm design, and evaluation methods to mitigate biases and ensure equitable outcomes for all individuals. Ethical considerations also extend to the allocation of resources, opportunities, and benefits generated by AI systems, ensuring that they are distributed fairly and transparently across diverse populations [7].

Accountability and transparency: AI systems operate as black boxes, making it challenging to understand how they arrive at their decisions. Ensuring accountability and transparency in AI requires mechanisms for

explaining and justifying AI-driven decisions to stakeholders, enabling them to understand, validate, and trust AI systems. Ethical considerations also include establishing clear lines of responsibility and accountability for AI systems, delineating roles and obligations for developers, operators, and users to promote responsible AI deployment and usage [7].

Privacy and data protection: AI systems rely on vast amounts of data for training and decision making, raising concerns about privacy, consent, and data protection. Protecting privacy and data rights in AI requires robust data governance frameworks, encryption techniques, and access controls to safeguard sensitive information from unauthorized access or misuse. Ethical considerations also include respecting individuals' autonomy and privacy preferences, ensuring transparency and informed consent for data collection, storage, and usage in AI applications [7].

Safety and security: AI systems have the potential to pose risks to safety and security if deployed without adequate safeguards and risk mitigation strategies. Ensuring safety and security in AI requires rigorous testing, validation, and certification processes to assess AI systems' reliability, robustness, and resilience to adversarial attacks. Ethical considerations also include designing AI systems with fail-safe mechanisms, ethical AI principles, and human oversight to prevent unintended consequences and ensure responsible AI deployment in safety-critical domains such as healthcare, transportation, and defense [7].

5.2.2 Regulatory considerations (Figure 5.2)

Figure 5.2: Regulatory considerations.

1. Regulatory frameworks: Regulatory frameworks play a crucial role in governing the development, deployment, and usage of AI technologies. Governments and regulatory bodies worldwide are increasingly focusing on establishing guidelines, standards, and regulations to address ethical, legal, and societal concerns related to AI. Regulatory considerations include defining AI terminology, classification, and taxonomy; setting ethical principles and guidelines for AI development and deployment; and establishing accountability mechanisms and enforcement mechanisms to ensure compliance with AI regulations [7].

2. Data governance and protection: Data governance and protection regulations govern the collection, storage, processing, and sharing of data used in AI applications. Regulatory considerations include data privacy laws such as the General Data Protection Regulation (GDPR) in Europe and the California Consumer Privacy Act (CCPA) in the United States, which impose strict requirements for data protection, consent, transparency, and accountability. Compliance with data governance regulations is essential for ensuring ethical AI deployment and protecting individuals' privacy and data rights [7].
3. Algorithmic accountability: Algorithmic accountability regulations aim to hold AI developers and operators accountable for the impacts of their algorithms on individuals, communities, and society at large. Regulatory considerations include establishing mechanisms for auditing, transparency, and explainability of AI systems to assess their fairness, bias, and discriminatory effects. Algorithmic accountability regulations also include provisions for redress mechanisms, oversight bodies, and regulatory enforcement actions to address harmful or discriminatory AI practices and promote responsible AI deployment [7].
4. Safety and certification: Safety and certification regulations govern the safety, reliability, and quality of AI systems deployed in safety-critical domains such as healthcare, transportation, and defense. Regulatory considerations include establishing safety standards, certification requirements, and regulatory approval processes for AI systems to ensure compliance with safety regulations and industry best practices. Safety and certification regulations also include provisions for risk assessment, hazard analysis, and mitigation strategies to address potential risks and vulnerabilities in AI systems [7].

Ethical and regulatory considerations are paramount in ensuring responsible development, deployment, and usage of artificial intelligence (AI) technologies. By addressing issues such as fairness, bias, transparency, privacy, accountability, and safety, ethical and regulatory frameworks help mitigate potential risks and promote trust, confidence, and acceptance of AI systems in society. As AI continues to evolve and impact various sectors and domains, ongoing dialogue, collaboration, and engagement among stakeholders are essential for developing robust and adaptive ethical and regulatory frameworks that uphold ethical principles, protect societal values, and foster innovation and progress in AI.

5.3 The Road Ahead for XAI

As the adoption of artificial intelligence (AI) continues to accelerate across various domains, the need for transparency, interpretability, and accountability in AI systems has become increasingly apparent. Explainable AI (XAI) has emerged as a critical area of research aimed at addressing these challenges and enhancing trust, understanding, and acceptance of AI-driven decisions.

This article explores the road ahead for XAI, examining key opportunities, challenges, and future directions shaping the evolution of transparent and interpretable AI systems [1, 7].

5.3.1 Opportunities

1. Advancements in XAI techniques: The road ahead for XAI is marked by continued advancements in techniques and methodologies aimed at enhancing transparency and interpretability in AI systems. Research in XAI encompasses a wide range of approaches, including modelspecific interpretability techniques, model-agnostic explanations, counterfactual

reasoning, and human–computer interaction methods. By developing innovative XAI techniques, researchers can unlock new opportunities for understanding and improving AI-driven decision-making processes across diverse applications and domains [1].

2. Interdisciplinary collaboration: Interdisciplinary collaboration is essential for advancing XAI research and addressing complex challenges at the intersection of AI, ethics, psychology, and human–computer interaction. Collaborative efforts between computer scientists, ethicists, psychologists, legal experts, and domain specialists can foster holistic approaches to XAI that integrate technical, ethical, and societal perspectives. By leveraging diverse expertise and insights, interdisciplinary collaboration can drive innovation and promote responsible AI deployment that aligns with societal values and aspirations [1].
3. Regulatory alignment: Regulatory alignment is critical for ensuring consistent and harmonized approaches to XAI governance and oversight across different jurisdictions and sectors. As AI technologies transcend geographical boundaries and impact global markets, regulatory frameworks must evolve to address ethical, legal, and societal concerns related to transparency, fairness, accountability, and privacy in AI systems. Regulatory alignment efforts involve international cooperation, standardization initiatives, and policy harmonization to promote responsible AI deployment and mitigate potential risks [7].

5.3.2 Challenges

1. Scalability and complexity: Scalability and complexity pose significant challenges for implementing XAI techniques in real-world AI systems, particularly in large-scale, high-dimensional, and dynamic

environments. Addressing scalability and complexity requires developing scalable XAI algorithms, frameworks, and tools that can handle diverse data sources, complex models, and real-time decision-making processes. Research in scalable XAI aims to overcome computational bottlenecks, optimize resource utilization, and enable XAI techniques to scale to massive datasets and complex AI systems [1].

2. Interpretability–accuracy trade-offs: The interpretability–accuracy trade-off is a fundamental challenge in XAI, where more interpretable models often sacrifice predictive performance or accuracy. Balancing the trade-offs between interpretability and accuracy requires developing hybrid approaches that combine the transparency of interpretable models with the predictive power of complex AI models. Research in interpretable machine learning focuses on designing hybrid models, ensemble methods, and post-hoc explanations that strike a balance between interpretability and accuracy in AI-driven decision-making processes [1].
3. Ethical and societal implications: Ethical and societal implications are central to the road ahead for XAI, as AI technologies increasingly shape our social, economic, and political landscapes. Addressing ethical and societal implications involves navigating complex trade-offs between competing values, interests, and stakeholders in AI deployment. Research in ethical AI aims to develop frameworks, guidelines, and principles for responsible AI development, deployment, and usage that uphold ethical values, protect human rights, and promote societal well-being [7].

5.3.3 Future directions

1. Human-centric XAI: Human-centric XAI focuses on designing AI systems that prioritize human values, preferences, and perspectives in decision-making processes. Future research in humancentric XAI aims to develop AI systems that are transparent, interpretable, and accountable to users, enabling meaningful human–AI collaboration and interaction. By integrating user feedback, preferences, and trust into AI systems, human-centric XAI can enhance user experiences, foster trust, and promote acceptance of AI-driven decisions in society [1].
2. Explainability across AI lifecycle: Explainability across the AI lifecycle involves providing transparent explanations for AI-driven decisions at various stages of the AI development, deployment, and usage lifecycle. Future research in explainability across the AI lifecycle aims to develop end-to-end XAI solutions that provide interpretable insights into data collection, model training, decision-making, and feedback mechanisms. By ensuring transparency and accountability throughout the AI lifecycle, explainability across the AI lifecycle can enhance trust, reliability, and fairness in AI systems [1].
3. Responsible AI governance: Responsible AI governance involves developing robust and adaptive governance frameworks that promote ethical, legal, and societal values in AI development, deployment, and usage. Future research in responsible AI governance aims to address emerging challenges and opportunities in AI governance, including regulatory alignment, stakeholder engagement, and accountability mechanisms. By fostering responsible AI governance, researchers can contribute to shaping a future where AI technologies are deployed and used in ways that benefit society while respecting human rights, dignity, and autonomy [1].

The road ahead for explainable artificial intelligence (XAI) is paved with opportunities, challenges, and future directions that shape the evolution of transparent and interpretable AI systems. By advancing XAI techniques, fostering interdisciplinary collaboration, promoting regulatory alignment, and addressing ethical and societal implications, researchers can navigate the road ahead for XAI and realize the potential of AI technologies to enhance transparency, accountability, and trust in decision-making processes across diverse applications and domains.

5.4 Summary

This chapter delves into the future trends and challenges in explainable artificial intelligence (XAI). It explores the advancements in XAI research, focusing on emerging techniques and methodologies aimed at improving the interpretability and transparency of AI systems. Ethical and regulatory considerations related to XAI are discussed, emphasizing the importance of addressing fairness, accountability, and privacy in AI-driven decisionmaking processes. The chapter outlines the road ahead for XAI, highlighting opportunities for interdisciplinary collaboration, regulatory alignment, and responsible AI deployment. It explores potential challenges and future directions in advancing XAI technologies across various domains, emphasizing the need for ongoing dialogue, collaboration, and engagement among stakeholders to shape the future of transparent and interpretable AI systems [1, 7].

References

1. van der Velden, B.H.M. *Explainable AI: current status and future potential* *Eur Radiol* 34, 1187–1189 (2024). https://doi.org/10.1007/s00330-023-10121-4

2. R.S. Peres, X. Jia, J. Lee, K. Sun, A.W. Colombo, J. Barata Industrial Artificial Intelligence in Industry 4.0 -Systematic Review, Challenges and Outlook, *Ieee*, 4 (2016).
3. D. Castelvecchi Can we open the black box of AI? *Nat News*, 538 (2016), p. 20.
4. European Commission, "*Building trust in human-centric AI*," 2018 [Online]. Available: https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html
5. Rai, A. (2020). Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science*, 48(1), 137–141.
6. Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gaševi, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 3, 100074.
7. Stahl, B.C. (2021). Ethical Issues of AI. In: Artificial Intelligence for a Better Future. *SpringerBriefs in Research and Innovation Governance*. Springer, Cham. https://doi.org/10.1007/978-3-030-69978-9_4

Chapter 7
Looking Further

"Watch out for arguments about future technology that is magical."

Rodney Brooks

Where we look at the societal impact of AI now and in the not too far future, considering, in particular, the future of jobs and education, and the potential risks associated with AI and the possibility of super-intelligence.

7.1 Introduction

In his article "Seven Deadly Sins of AI Predictions",[1] Rodney Brooks quotes Amara's Law: *"We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run."* In its relatively short history, AI has been overestimated several times, in the 1960s, in the 1980s, leading to the so-called, AI winters, and underestimated at least as many times. I've been working on AI since the late 1980s and during most of my working life I have needed to justify why was I doing research in such an esoteric field, unlikely to ever produce any useful product. But the current developments in AI and machine learning would not have been possible without the work laid out in those cold winter years.

Nevertheless, today's media and business interest would almost make one believe that AI is a new technology which 'suddenly' has taken over the world. The current rise of AI is often compared to the Industrial Revolution. In 2017, Steven Cave, executive director of the Leverhulme Centre for the Future of Intelligence, referred to the AI revolution as "...likely to happen even faster

[1] See https://www.technologyreview.com/s/609048/the-seven-deadly-sins-of-ai-predictions/.

V. Dignum, *Responsible Artificial Intelligence*, Artificial Intelligence: Foundations, Theory, and Algorithms, https://doi.org/10.1007/978-3-030-30371-6_7

– so the potential damage is even greater."[2] The pace may be increasing, but actually, humankind has always been concerned with the pace of technological change.

Moreover, machines have been making decisions for us for quite some time, and are doing it autonomously in many cases. My thermostat decides whether to turn the central heating on or off based on the information it has about my preferred room temperature; the electronic gate in my local train station decides whether to grant me access to the platforms based on the information it gets on my travel credit and on constraints given to it by the local travel authorities; Google decides which of the trillions of Web pages I am more likely to want to see when I search for something; and Netflix tells me what I might want to watch tonight, based on the information it has about my preferences and those of people it thinks are similar to me.

What makes AI decision-making different, and for some people more worrisome, is the opacity of those decisions, and the perceived lack of control over the actions that AI systems take based on those decisions. This view is both stifling and dismissive. Stifling because it may lead to a feeling of powerlessness in our control over machines, and dismissive because it may lead humans to feel less responsibility for the use and outcomes of AI systems.

As we have seen throughout this book, we, humans, are the ones that determine the optimisation goals and the utility functions at the basis of machine learning algorithms; we are the ones that decide what the machine should be maximising. Indeed, even in the famous *paperclip maximiser* example by Nick Bostrom [20], someone once gave this maximisation goal to the unfortunate intelligent factory.

AI does potentially pose many risks and can be used for evil by evildoers. However, AI also brings enormous potential to improve the lives of many, and to ensure human rights for all. It is up to us to decide:

- Are we building algorithms to maximise shareholder profit or to maximise fair distribution of resources in a community? For example, algorithms that can provide solutions to tragedy-of-the-commons situations and support fair distribution of resources;
- Are we using AI to optimise company performance or to optimise crop yield for small farmers around the world? For example, AI systems that provide real-time information on fertiliser levels, planting and harvesting moments and weather conditions;
- Are we building AI systems to emulate and replace people, misleading users about their machine nature, or are we building AI to ensure well-being, participation and inclusion? For example, translation services to improve cross-cultural communication, tools to provide access to information and education for all, and that curb the distribution of fake news.

[2] See https://www.telegraph.co.uk/business/leaders-of-transformation/horizons/intelligence-revolution/

It is up to us to decide. AI development can be motivated by money and shareholder value, or by human rights and well-being, and societal values. It is up to us to decide. Is AI going to enhance our facilities and enable us to work better, or is it going to replace us? The impact of AI on jobs is perhaps one of the most-discussed aspects of the potential technological advance that AI brings. We will discuss this issue in Section 7.2.1.

The power of decision is the power of all of us. Researchers, developers, policymakers, users. Every one of us.

But to use this power we need to be informed and involved in the policy and strategy discussions around AI. All of us. In the same way that war is too important to be left to the generals, and democracy is too important to be left to the politicians, AI is too important to be left to the technocrats. This implies that we need a multidisciplinary approach to AI, but most of all that we need a different approach to education that enables all to be involved. We discussed the importance of education with respect to inclusion and diversity in Section 6.4, and will reflect on the consequences of AI for education in Section 7.2.2.

We need also to consider the risks associated with AI technology. These can be either intended or unintended but in both cases may lead to profound negative consequences. We will discuss the main risks in Section 7.2.3. In Section 7.2.4 we reflect on the many beneficial uses of AI, the field of AI for Good.

Finally, a book about AI cannot be complete without some reflection on the possibility and plausibility of super-intelligence. We will look at this in Section 7.3.

7.2 AI and Society

As Artificial Intelligence technology becomes better at many tasks that have so far been done by people, we can expect to see an increase in efficiency and wealth. However, many concerns have been voiced as to how these benefits are going to be guaranteed and shared by the whole of humankind and not remain the privilege of a few. Concerns about the impact of Artificial Intelligence on society as we know it, are currently widely voiced by experts, media and policymakers alike. These concerns mostly follow two main directions. On the one hand, worries about the impact of AI on our current society: the number and nature of jobs, privacy, (cyber)security and deployment of autonomous weapons are some of the most-cited issues. On the other hand, the possible existential risks associated with super-intelligence, or singularity, that is, the dangers for humankind in the event that AI systems surpass human intelligence.

It is an often-heard claim that Artificial Intelligence has the potential to disrupt all areas of society and business. Many initiatives, news reports and

projects are concerned with the societal impact of AI and the means to ensure that this impact is a positive one.

7.2.1 Future of Jobs

AI's impact may be be most noticeable on how jobs and work will look in the not too far future. When AI systems will replace people in many traditional jobs, we must rethink the meaning of work. Jobs will change but more importantly the character of jobs will change. Meaningful occupations are those who contribute to the welfare of society, the fulfilment of oneself and the advance of mankind. These are not necessarily equated with our current understanding of 'paid jobs'. Used well, AI systems can free us to care for each other, engage in arts, hobbies and sports, enjoy nature, and meditate, i.e. those things that give us energy and make us happy. But used without concern for societal impact and human well-being, AI can lead to massive job losses, increased inequality, and social unrest. Defining proper incentives and rewards to ensure well-being and sustainability, and a new definition of wealth and meaningful occupation are needed to ensure that AI is used for the good of humanity and the environment. In parallel, many new jobs will appear for which skilled human workers are needed with a set of skills that combine technical education with the humanities, arts and social sciences.

AI will perform hard, dangerous or boring work for us; it will help us to save lives and cope with disasters; and it will entertain us and make our daily life more comfortable. In fact, current AI systems are already changing our daily lives, almost entirely in ways that improve human health, safety and productivity. In the coming years we can expect AI systems to be increasingly be used in domains such as transportation, service robots, healthcare, education, low-resource communities, public safety and security, employment and the workplace, and entertainment. These systems must be introduced in ways that build trust and understanding, respect human and civil rights, and are in line with cultural and social context.

It is urgent to anticipate and prepare for the impact of our digital future. Next to the imperative technological skills, increasingly the human workforce of the future will be challenged to cooperate, adapt to an ever-changing world, and maintain a questioning mind. AI applications will be participating in the digital ecosystem alongside us, and whereas these systems can help us in many ways, co-existing alongside AI systems will necessarily bring with it changes in the way we, people, interact, learn and work.

In general, experts believe that AI technology is both creating and destroying jobs, and notably also that it's unlikely to cause a major reduction in the number of jobs in the future. We know from historical trends that in the long term, technological advances have always been beneficial for the quality and quantity of jobs. However, in the short term, disruptions can be expected to

specific occupations and specific demographic groups. Studies also show that regions where job opportunities are diverse will likely adjust better to change [59].

Education and training are mandatory to ensure that disruptions are minimised. Even though we cannot know for sure which jobs will exist in the future, it is expected that human skills of empathy, caring, creativity and ease to quickly adapt to unforeseen situations will be central. Also, the demand for skills that combine technical education with the humanities, arts and social sciences is likely to increase. We also need to foster the use of AI to support the development of new skills that will enable people to adapt to new types of jobs [56].

There is also an important role for regulation here to ensure that the burden of change is balanced across regions, demographic groups and job areas. Possible regulatory interventions that are currently being considered include taxation (e.g. the infamous robot tax[3]) and 'universal basic income' schemes. This also requires social partners, such as trade unions and professional organisations to be involved in the conversation and to take their own responsibility for the impact of AI.

In conclusion, technological developments in the last century led to mass production and mass consumption. Until very recently, having has been the main goal and competition the main drive: 'I am what I have'. Digital developments, including AI, favour openness over competition: open data, open source and open access. The drive is now quickly shifting to sharing: 'I am what I share'. Combined with the changing role of work, this novel view of wealth requires a new view of economy and finance, but can contribute to a fundamental positive change to society.

7.2.2 Future of Education

The digital transformation of society is possibly the main challenge of this century. By the end of 2013, those that have grown up in a digital world started to outnumber those that had to adapt to it. However, capacity building to ensure that everybody is able to contribute to the digital ecosystem and to fully participate in the workforce is lagging behind, and current education curricula are perhaps not the most suitable to meet the demands of future work.

Considering that *"the tools that we shape, will thereafter shape us"*,[4] the digital ecosystem will bring along a redefinition of fundamental human values, including our current understanding of work and wealth. In order to ensure the skills needed for resilient and sustainable capacity building for the digital

[3] See https://www.techemergence.com/robot-tax-summary-arguments/ for a summary of the main arguments for and against robot taxation.

[4] Quote attributed to media theorist Marshall McLuhan.

ecosystem, the following aspects must be central in education curricula across the world:

- **Collaborate**: The digital ecosystem makes possible and assumes collaboration across distance, time, cultures and contexts. The world is indeed a village, and all of us are the inhabitants of this village. Skills are needed to interact, build relationships and show the self-awareness needed to work effectively with others in person and virtually, across cultures.
- **Question**: AI systems are great at finding answers, and will do this increasingly well. It is up to us to ask the right questions, and to critically evaluate results in order to be able to contribute to responsible implementation of solutions.
- **Imagine**: Skills to approach problem-solving creatively, using empathy, logic and novel thinking, are needed. For this, humanities education is paramount and should be included in all technology curricula.
- **Learn to learn**: The ability to adapt and pick up new skills quickly is vital for success, requiring us to continuously learn and grow, and adapt to change. Being able to understand what it is necessary to know, and knowing when to apply a particular concept as well as knowing how to do it, are key to continuous success.

The digital age is a time for reinvention and creativity. Capacity building must embrace these skills alongside technological expertise. This shows that the traditional separation between humanities, arts, social sciences and STEM (Science, Technology, Engineering and Mathematics) is not suitable for the needs of the digital age. More than multidisciplinary, future students need to be transdisciplinary: creating a unity of intellectual frameworks beyond the disciplinary perspectives. In fact, I would say that Artificial Intelligence is not a STEM discipline. It is in essence trans-disciplinary and requires a spectrum of capabilities that is not covered by current education curricula. It is urgent to redesign studies. This also gives a unique opportunity to truly achieve inclusion and diversity across academic fields.

7.2.3 Dealing with Risks

AI brings in itself enormous potential to improve the lives of many, and to ensure human rights for all. It is for all our sakes that we should ensure that risks are minimised. As discussed in Chapter 6, a combination of regulation, certification, education and self-awareness are important towards this endeavour.

Whether risks are intended or unintended, they can have profound consequences for safety, democracy and the very meaning of being human. In the following, we discuss some of these issues. Note that this discussion aims at

presenting an overall idea and does not pretend to be an exhaustive list of all potential risks.

Dealing with risk requires that people are equipped with the competencies to control the systems they interact with, rather than be controlled by them. This again requires education and the willingness of societal institutions to take responsibility for the ways we are allowing AI to shape society.

Safety

Ensuring that AI systems are safe, is a main condition for trust.

Safety is about being sure that the system will indeed do what it is supposed to do, without harming users, resources or the environment. Also it is about knowing that the purpose of the system is beneficial and according to human rights and values.

Unintended consequences are of many sorts. They can be errors in programming, but also incorrect application of resources, regulations or algorithms. Biased results, breaches of privacy, or erroneous decisions are some of the results. As we have discussed in Chapter 4 a structured, open and value-centred design process is of utmost importance to mitigate and correct unintended risks. Moreover, formal mechanisms are needed to measure and direct the adaptability of AI systems and to ensure robust processes. Safeguards that enable fall-back plans in case of problems with the AI system are also necessary. In some cases this can mean that the AI system switches from statistical to rule-based procedures; in other cases it can mean that the system will request an human operator to take over control.

Worse is the case of intended malicious uses of AI. These range from minor nuisances such as email spam, to full-fledged cyber warfare [117]. A recent report [24] surveys the landscape of potential security threats from malicious uses of artificial intelligence technologies, and proposes ways to better forecast, prevent and mitigate these threats. The report recommends a combination of regulation and taking responsibility to address this issue. In particular, researchers and developers are urged to consider carefully the dual-use purpose of their work. Moreover, developing AI techniques that are able to assess and deflect malicious objectives of other AI systems can also contribute to minimise the effects of misuse of AI.

Democracy

It is increasingly clear that AI will contribute to a fundamental change in the way in which we organise the economy and society. But is AI the end of democracy as we know it, as some have predicted [72]? The current ideal of democracy is grounded on the individual's right to self-determination [32].

By affecting people's self-determination, AI is potentially affecting the democratic process, for better or for worse.

What we are currently seeing is that the capability to even out majority sentiment, traditionally a function of democratic institutions, is eroding under the possibilities of AI systems. Increasingly, the Internet has rendered the diversity of citizens' views more salient, and has proven a powerful medium for discontented citizens to put pressure on democratic institutions and force changes in policies. At the same time, manipulating public sentiment towards specific standpoints is made easy by (fake) news targeted at individual sentiments and preferences. The political risks associated with AI refer to the risks of imbalance of power between individuals, groups or values in society, and are made possible by the increased power and speed in collection and processing of information.

Often democracy is taken as given. The current view is that democracy is the great equaliser, something that we need and want to uphold in all cases. However, critical perspectives on democracy show that traditional democracy often privileges a specific kind of individual/citizen as well as rules out more radical notions of democracy.

On the other hand, AI and digital technologies are already disrupting the traditional view of democracy, and not always for the better, ensuring that processes are more inclusive. In fact, if anything, AI is strengthening the link between democracy and economics, supporting the manipulation of information to meet the needs of a few, and enabling the emergence of super economic powers that are outside democratic scrutiny and control. When algorithms are used to decide our access to information about a news item, a political candidate or a business, opinions and votes can shift, and potential government's be made or broken. Because algorithmic censorship is mostly unregulated, large corporations can in principle decide what information we have, outside the traditional democratic processes of governance and accountability. Moreover, these same corporations also own most of our, and our governments, data, by the conditions under which we use their products to share, store and manage information.

Thus, democracy as we know it is changing its forms. It is an open area of research to investigate how democracy is changing through the use and effect of AI.

Another important issue is that of participation and control. AI systems currently are for a large part owned and dominated by large, private corporations, which not only determine how and what AI is being designed and deployed but also control and formally own all the data shared through those systems. As such these corporations have a huge impact on how we create and design our public space, 'the commons' that we use, and that are also used by municipalities, schools, etc. for communicating with citizens, pupils and parents.

In short, the ones that control AI can *de facto* control democratic processes and institutions. Without clear regulation, we have no way to under-

stand what are the values and requirements behind those systems. Moreover, as collective responsibility is perceived to be eroding in many societies, alignment between social and individual values becomes less clear, and each one's notions of what is legally allowed, socially accepted and morally acceptable are no longer shared to the same extent.

Human dignity

As AI systems increasingly replace human decision-making, the risk is that we will become dependent on these systems. The margin between enhancement of capabilities and dependency on technology is narrow. More importantly, how autonomous and self-determining are we when AI systems determine the information we receive, and nudge us to choose healthy choices, to watch certain programs, to vote for given candidates? And how can we justify our own autonomous choices, when those are contrary to the ones our 'know-better' systems propose to us?

Due to their ability to mine data and make decisions, and powered by various abilities to interact with users, AI systems can potentially interfere with users, even if the aim is to contribute to their own good, thus exposing them to a form of paternalism. In a recent report by the AI4People work group, we described the main risks of AI. These have to do with the nature of being human and the human right to self-determination and autonomy, and include the danger of devaluing human skills, removing human responsibility, and reducing human control [56].

Moreover, by interpreting earlier choice behaviour as well as by inspecting users' goals and wishes, AI systems can potentially create preference profiles that are very reliable and that reflect their user's long-term agendas better than the less informed, more context-dependent and more temporarily distorted decisions of users themselves. This raises the issue of whether and in what circumstances such information should guide interactions with users, potentially even when users object to such interaction.

Ensuring human self-determination will impact the way assistive AI systems are designed: to what extent and in what circumstances is it morally justified to act towards another, potentially in freedom-restricting ways, when such action is not a response to that person's request, or is without her consent, or even against her expressed will? Assistive technology that is supposed to assist its user but that does so in ways that the user has not requested or not consented to is arguably ignoring the user's self-determination.

7.2.4 AI for Good

Beyond the risks that AI may pose, AI's impact will be defined by its contribution to human well-being and eco-social sustainability. The idea of AI for Good refers to this vision. Contribution to the achievement of UN Sustainable Development Goals is a main driver of AI for Good initiatives.

AI for Good projects also aim at extending access to AI, in particular to those regions and demographic groups less likely to have easy access to technology. Initiatives targeting developing countries, minorities and people with disabilities are examples of such projects. Other areas of AI for Good efforts are the use of AI technology to solve environmental questions, to support sustainable agriculture, and to promote low-carbon business practices. AI can also be used to help farmers, improve diagnostics, personalise learning or help refugees find jobs.

Nevertheless, and despite many events promoting AI for Good and an increasing number of organisations dedicated to the topic, most of these initiatives are small-scale and prototypical. A possible way forward is through incentives and nudges to bring large corporations to commit part of their work towards 'for Good' projects.

7.3 Super-intelligence

This book cannot end without a reflection on the issue of super-intelligence: the hypothetical capability of machines to surpass the brightest and most-gifted human minds. This relates to the quest to develop machines that not only reproduce but exceed human intelligence, or what is sometimes referred to as 'true AI'. It would suffice here to quote Luciano Floridi and state that *"True AI is not logically impossible, but it is utterly implausible"* [54], but actually I am not so sure about the logical possibility of such systems.

Humankind has always been engaged in developing super-scaled versions of itself. Airplanes can fly better than we can, cars, or even bicycles can move faster than we do, calculators can calculate square roots better than we can. All are superhuman from a specific perspective. However, as attributed to Edsger W. Dijkstra: "the question of whether a computer can think is no more interesting than the question of whether a submarine can swim."

Rather than focusing on the possibility, or not, of building super-intelligent machines, it is probably more relevant to discuss how we feel about it, and, from a responsibility/ethical perspective: should we do it, and what does it mean if we succeed? In the following, I will describe my views about all these issues.

Firstly, how do we feel about super-intelligent artefacts? Even before we think about whether it is possible and whether we should want to build such artefacts if it is possible, already our feelings about this are different from

those we experience about other superhuman artefacts. What is it that makes intelligence different from flying or moving? We have no problem recognising that when we talk of an airplane flying, we mean a very different action than what birds do. That is, artificial flying is different from natural flying. In the same way, Artificial Intelligence is different from natural intelligence. It will, and in fact already does, complement us in many tasks, but it is a fundamentally different phenomenon than human intelligence.

We also tend to put intelligence and consciousness in the same box. What makes us conscious? Who and what are conscious entities? Newborn human babies? Monkeys? Chickens? Trees? Stones? What will make a machine conscious? As an illustration, consider the moment Lee Sedol was beaten by AlphaGo at the game of Go. Most newspapers showed a picture of the moment Lee realised he would lose: powerlessness, disbelief, sadness, all are expressions of his consciousness. Next to him was AlphaGo, a computer. It had no notion of what winning or losing meant, in fact it was not even aware it was playing a game. It was just following instructions and optimising some utility functions. Was it conscious?

Intelligence is not just about knowing, it is about feeling, enjoying, pushing limits. I often run marathons. I don't doubt that it is possible to build a 'running robot' but will it ever experience, and enjoy, what it means to run a marathon, to push through the pain and enjoy it?

On the issue of whether it is possible to build super-intelligent machines, the field is divided. Whereas many top scholars are sure of this possibility, the belief is not shared by all. One of the main supports for the possibility of super-intelligence is the so-called Church-Turing hypothesis [34].

Put simply, this thesis states that a mathematical function is computable if and only if it is computable by a Turing machine, i.e. by manipulation of symbols. This abstract machine can simulate what occurs in any computer, in a similar logical process. This being true, then there is a universal core to all computations which then means that any computation our brains can make can be simulated by a machine. The Church-Turing thesis thus argues that human intelligence can be reproduced by machine, but not that super-intelligence can be so achieved. Till date, we do not have a model more powerful than Turing Machine which can solve problems that Turing Machine cannot. However, it is not proved that a human brain can be represented as a Turing machine.

However, a Turing machine assumes infinite time and resources are available for the computation. Moreover, the (energetic) cost of such computations in an *in silico* machine are probably too large to allow practical development of such systems using the hardware we have available, nor or in the foreseeable future.

Another objection to super-intelligence is its conception of natural intelligence as a literal, single-dimension, linear graph of increasing amplitude, as it is portrayed by Nick Bostrom in his book *Superintelligence* [20]. However, intelligence is not one-dimensional. It is a complex of many types and

modes of cognition, each one a continuum. Therefore, the paradigm of super-intelligence may fail in its assumption that if we manage to make machines intelligent in many different areas, it will entail super-intelligence in general. That is, intelligence is not compositional. Another way of putting it is not about finite vs infinite, but limits. For example, if I jump out of a plane I will reach terminal velocity due to friction. Are there similar frictions acting on intelligence? In his book *"Machines that Think"*, Toby Walsh discusses many more arguments against super-intelligence [133].

Moreover, super-intelligence rests on the assumption that intelligence is infinite. However, there is no other physical dimension in the universe that is infinite, as far as science knows so far. Temperature, space, time, speed are all finite. Only the abstract concept of numbers is infinite. It stands to reason that intelligence would itself also be finite.

Finally, there is the issue of whether we should want to build super-intelligence. Sure. As long as we build it to extend our capabilities to ensure human flourishing and well-being in a sustainable world, it is like extending our other capabilities to a super version of our selves. A main fear about super-intelligence is that those super-intelligent machines would rule out humans. However, remember, they are artefacts that we have built. The reason we need to build AI responsibly is exactly to ensure that the purpose we put into the machines, is the purpose we really need [103]. The goals that an AI system has can always be related to some human actor (that is why humans are the responsible agents). Besides, whatever goals the system has, they don't 'matter' to the system; its goals are not linked to any ulterior 'needs' of the system, they are given to it. So, in the words of Margaret Boden, the machine has no interest in ruling humankind. It will not take over because it couldn't care less [15].

Moreover, it is important to remember that true intelligence is not about 'winning'. It is about social skills, about collaboration and contribution to a greater good, about joining forces in order to survive and prosper. There is no reason to expect super-intelligence will be different.

On the other hand, we can also state that actually super-intelligence is easy. It is already here. It is the combined intelligence of all humans and other reasoning entities put together to work towards a common goal.

Or we can ask, does it matter? The ultimate goal of technology is to improve the human condition in a sustainable way for all of us and for our environment. As long as the motives are pure and we are achieving this goal, we are obliged to do so. This actually combines the different philosophical views on ethics, and brings us back to Chapter 3.

7.4 Responsible Artificial Intelligence

Responsible AI means that AI systems should be designed and implemented in ways that recognise and are sensitive to human interaction contexts without infringing on core values and human rights. Even though AI in itself does not absolve individuals from responsibility for their actions and decisions, the increasing complexity of AI systems renders attribution of responsibilities more difficult. Therefore, methods are needed that clarify responsibilities and make explicit what are the design choices, data and knowledge provenance, process and stakeholders.

Responsible AI means that AI systems must be understood as part of complex socio-technical systems. As such, an empirical/experimental ethics approach that can shape responsible (or good) AI, not only from the outside but also from within AI practices, is needed.

Responsible AI means that design, development and use of AI must take into account societal values and moral and ethical considerations, weigh the respective priorities of values held by different stakeholders in different multicultural contexts, explain its reasoning and guarantee transparency.

Responsible AI is more than the ticking of some ethical 'boxes' in a report, or the development of some add-on features, or switch-off buttons in AI systems. Rather, responsibility is fundamental to autonomy and one of the core stances underlying AI research and development.

But above all,

> **Responsible Artificial Intelligence is about human responsibility for the development of intelligent systems along fundamental human principles and values, to ensure human flourishing and well-being in a sustainable world.**

I look forward to a future when all AI is Responsible AI.

7.5 Further Reading

This book presents my views about the impact of AI. My aim was not to present a vision of the future, but to describe the current developments and the opportunities and challenges we face in ensuring that AI's impact will be a positive one for human well-being. In the last few years, several authors have proposed some interesting, in some cases profound, views of the future. You may want to check these:

- Tegmark, M. *Life 3.0: Being Human in the Age of Artificial Intelligence.* Knopf, 2017

- Harari, Y. N. *Homo Deus: A Brief History of Tomorrow.* Random House, 2016
- Walsh, T. *Machines that Think: The Future of Artificial Intelligence.* Prometheus Books, 2018