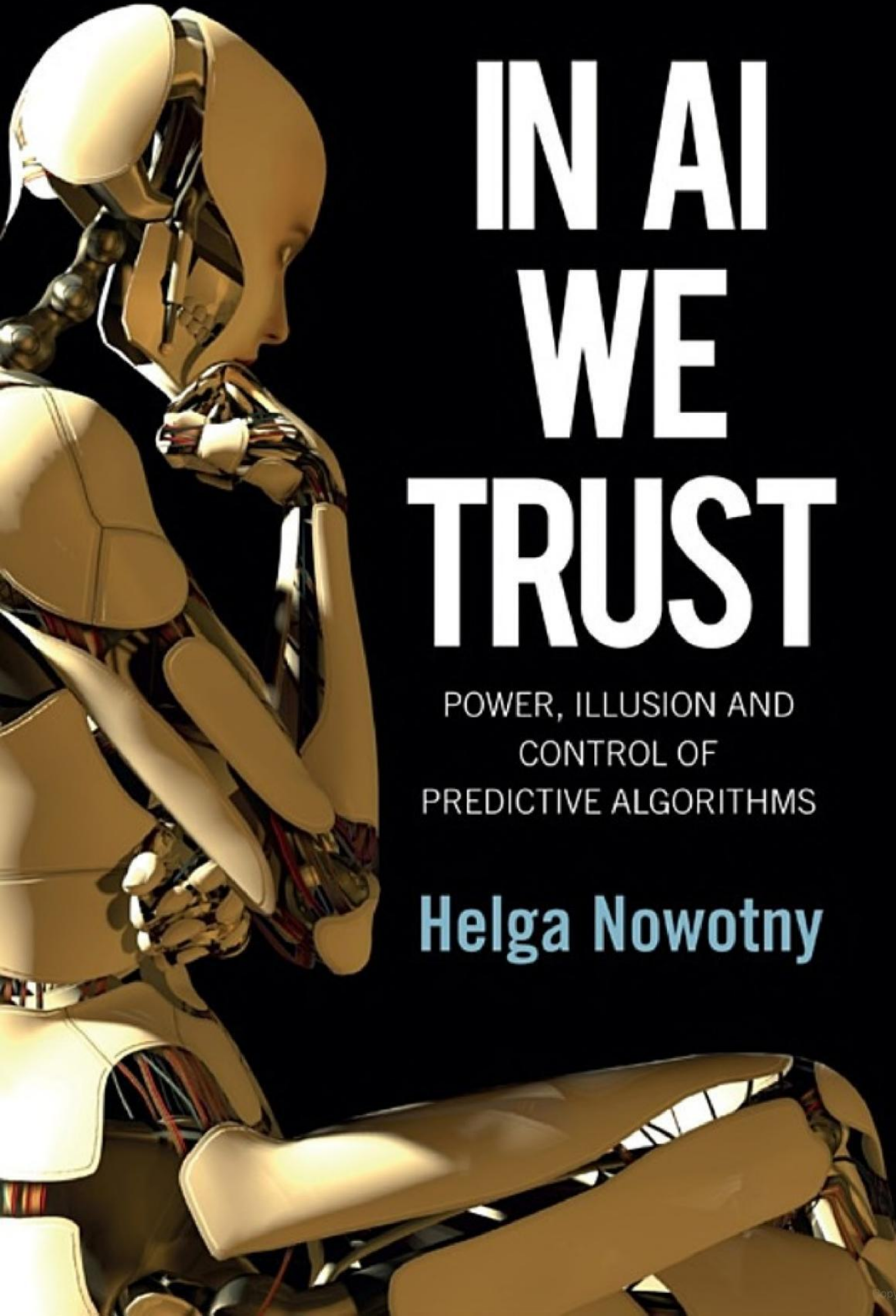

IN AI WE TRUST

POWER, ILLUSION AND CONTROL OF PREDICTIVE ALGORITHMS

Helga Nowotny

Table of Contents

In AI We Trust

Power, Illusion and Control of Predictive Algorithms

Helga Nowotny

polity

Copyright Page

First published in 2021 by Polity Press

Polity Press

65 Bridge Street

Cambridge CB2 1UR, UK

Polity Press

101 Station Landing

Suite 300

Medford, MA 02155, USA

ISBN-13: 978-1-5095-4881-1

A catalogue record for this book is available from the British Library.

Library of Congress Cataloging-in-Publication Data

Names: Nowotny, Helga, author.

Title: In AI we trust : power, illusion and control of predictive algorithms / Helga Nowotny.

Description: Medford, MA : Polity Press, 2021. | Includes bibliographical references and index. | Summary: “A highly original account of the nature of artificial intelligence and its implications for our future”-- Provided by publisher.

Identifiers: LCCN 2021003322 (print) | LCCN 2021003323 (ebook) | ISBN 9781509548811 (hardback) | ISBN 9781509548828 (epub)

Subjects: LCSH: Artificial intelligence.

Classification: LCC Q335 .N685 2021 (print) | LCC Q335 (ebook) | DDC 006.3--dc23

LC record available at https://lccn.loc.gov/2021003322

LC ebook record available at https://lccn.loc.gov/2021003323

by Fakenham Prepress Solutions, Fakenham, Norfolk NR21 8NL

The publisher has used its best endeavours to ensure that the URLs for external websites referred to in this book are correct and active at the time of going to press. However, the publisher has no responsibility for the websites and can make no guarantee that a site will remain live or that the content is or will remain appropriate.

Every effort has been made to trace all copyright holders, but if any have been overlooked the publisher will be pleased to include any necessary credits in any subsequent reprint or edition.

For further information on Polity, visit our website: politybooks.com

Acknowledgements

This book has been in the making for a long time, the strands of thought interrupted by travel and other obligations. There were several unsatisfactory starts and abandonments before what I was looking for came into focus. The actual process of writing benefited, perhaps paradoxically, from several COVID-19 lockdown periods, with the obligatory and, in my case, productive solitude they imposed. As an unprecedented event for everyone living in the twenty-first century, the pandemic also led me to reflect on its many unintended consequences, including some that have a direct bearing on the themes of this book.

My thanks go to all those who supported me in different ways on this long journey. Jean-Luc Lory offered me again a brief stay of hospitality at the Maison Suger in Paris, a wonderful place of calm in the centre of a buzzing metropolis. Another brief stay that I looked forward to, at the Wissenschaftskolleg zu Berlin, had to be cancelled due to abruptly imposed travel restrictions. Despite the missed occasion I was able to remain in fruitful exchange with Elena Esposito. I also want to thank Vittorio Loreto, of the Sony Lab Paris, for an especially inspiring conversation over dinner in Vienna. While the meal was frugal due to constraining circumstances, the discussion was rich. Stefan Thurner, who heads the Complexity Science Hub in Vienna, has been a continued source of inspiration and productive criticism. I received valuable feedback from him on parts of the manuscript, for which I am very grateful. Special thanks go to Michele Lamont at Harvard University for intellectual and moral support throughout the long stretch of the gestation period. Ever since we first talked about the book over a delicious lunch in New York City, Michele has responded to all my queries and provided me with continued encouragement.

When I approached John Thompson of Polity Press at the beginning of summer 2020 with an outline of the book, he responded without hesitation. Ever since, the process of getting the book published has unfolded in an ideal spirit of cooperation. I also want to thank the anonymous reviewers organized by Polity. One of them especially provided me with precious and concise feedback. I am also extremely lucky to have a splendid granddaughter, Isabel Frey, who volunteered to be my personal editorial assistant. Isabel is not only a wonderful singer of Yiddish protest songs, but an engaged activist and researcher who represents the younger generation acting to change the world. Barbara Blatterer, my long-term personal assistant, took meticulous and efficient care in getting the manuscript through several versions to the final and decisive cuts, advising me also on the book's cover. To all of you, my heartfelt thanks.

Carlo Rizzuto acted again as my very special personal advisor, reading, commenting and encouraging me when I was doubtful whether I would ever finish. His unfailing humour provided me with the necessary distance from what I was doing. My thanks to him are as much for the ways in which he interrupted my writing as for everything else he gave me during this period.

Vienna and Bonassola, January 2021

Introduction: A Personal Journey into Digi-land

Origins: time and uncertainty; science, technology and society

This book is the outcome of a long personal and professional journey. It brings together two strands of my previous work while confronting the major societal transformations that humanity is undergoing right now: the ongoing processes of digitalization and our arrival in the epoch of the Anthropocene. Digitalization moves us towards a co-evolutionary trajectory of humans and machines. It is accompanied by unprecedented technological feats and the trust we put into Artificial Intelligence. But there are also concerns about continuing losses of privacy, what the future of work will be like, and the risks AI may pose for liberal democracies. This creates widespread feelings of ambivalence: we trust in AI as a bet on our future, but we also realize that there are reasons for distrust. We are learning to live with the digital devices we cheerfully interact with as though they were our new relatives, our digital others, while retaining a profound ambivalence towards them and the techno-corporate complex that produces them.

The process of digitalization and datafication coincides with the growing awareness of an environmental sustainability crisis. The impact of climate change and the dire state of the ecosystem upon which we depend for survival call for urgent action. But we are equally in thrall to or anxious about the digital technologies that are sweeping across our societies. Rarely, however, are these two major transformations – digitalization and the transition towards sustainability – thought together. Never before have we had the technological instruments and the scientific knowledge to see so far back into the past and ahead into the future, nor the techno-scientific capabilities for action. And yet, we feel the need to reconsider our existence in this uncanny present that marks a transition towards an unknown future that will be different from what has been promised to us in the past. This widespread feeling of anxiety has only been exacerbated by the COVID-19 pandemic, itself a major disruptive event with long-term consequences at a global scale.

My journey leading up to this book was long and full of surprises. My previous work on time, especially the structure and experience of social time, led me to inquire how our daily exposure to and interaction with AI and the digital devices that have become our intimate companions alter our experience of time once again. How does the confrontation with geological timescales, long-term atmospheric processes or the half-life of the dissolution of microplastic and toxic waste affect the temporalities of our daily lives? How does AI impinge on the temporal dimension of our relationship with each other? Are we witnessing the emergence of something we can call 'digital time' that has now intruded into the familiar nested temporal hierarchy of physical, biological and social times? If so, how do we negotiate and coordinate these different kinds of time as our lives unfold?

The other strand of my previous work, on uncertainty, directed my inquiry towards ways of coping with and managing old and new uncertainties with the help of the powerful computational tools that bring the future closer into the present. These tools allow glimpses into the dynamics of complex systems and, in principle, enable us to identify the tipping points at which systems transition and change the state they are in. Tipping points mark further transformation, including the possibility of collapse. As science begins to understand complex systems, how can this knowledge be harnessed to counteract the risks we face and strengthen the resilience of social networks?

Not surprisingly, I encountered several hurdles on my way, but I also realized that my previous long-standing interest in the study of time and the cunning of uncertainty – which, I argued, we should embrace – allowed me to connect aspects of my personal experience and biographical incidents with empirical studies and scientific findings. Such personal links, however, no longer seemed available when confronting the likely consequences of climate change, loss of biodiversity and the acidification of oceans, or issues like the future of work when digitalization begins to affect middle-class professionals. Like many others confronted with media images of disastrous wildfires, floods and rapidly melting arctic ice, I could see that the stakes had become very high. I kept reading scientific reports that put quantitative estimates on the timelines when we would reach several of the possible tipping points in further environmental degradation, leading to the collapse of the ecosystem. And, again like many others, I felt exposed to the worries and hopes, the opportunities and likely downsides, connected with the ongoing digitalization.

Yet, despite all these observations and analyses, a gap remained between the global scale on which these processes unfolded and my personal life which, fortunately, continued without major perturbations. Even the local impacts were being played out either in far-away places or remained local in the sense that they were soon to be overtaken by other local events. Most of us are cognizant that these major societal transformations will have huge impacts and numerous unintended consequences; and yet, they remain on a level of abstraction that is so overwhelming it is difficult to grasp intellectually in all its complexity. The gap between knowing and acting, between personal insight and collective action, between thinking at the level of the individual and thinking institutions globally, appears to shield us from the immediate impact that these far-reaching changes will have.

Finally, it struck me that there exists an entry point that allows me to connect curiosity-driven and rigorous scientific inquiry with personal experience and intuition about what is at stake: the increasingly important role played by prediction, in particular by predictive algorithms and analytics. Prediction, obviously, is about the future, yet it reacts back on how we conceive the future in the present. When applied to complex systems, prediction faces the non-linearity of processes. In a non-linear system, changes in input are no longer proportional to changes in output. This is the reason why such systems appear as unpredictable or chaotic. Here we are: we want to expand the range of what can be reliably predicted, yet we also realize that complex systems defy the linearity that still underpins so much of our thinking, perhaps as a heritage of modernity.

The behaviour of complex systems is difficult for us to grasp and often appears counter-intuitive. It is exemplified by the famous butterfly effect, where the sensitive dependence on initial conditions can result in large differences at a later stage, as when the flapping of a butterfly's wings in the Amazon leads to a tornado making landfall in Texas. But such metaphors are not always at hand, and I began to wonder whether we are even able to think in non-linear ways. Predictions about the behaviour of dynamic complex systems often come in the garb of mathematical equations embedded in digital technologies. Simulation models do not speak directly to our senses. Their outcome and the options they produce need to be interpreted and explained. Since they are perceived as being scientifically objective, they are often not questioned any further. But then predictions assume the power of agency that we attribute to them. If blindly followed, the predictive power of algorithms turns into a self-fulfilling prophecy – a prediction becomes true simply because people believe in it and act accordingly.

So, I set out to bridge the divide between the personal, in this case the predictions we experience as being addressed to us as individuals, and the collective as represented by complex systems. We are familiar and at ease with messages and forms of communication at the inter-personal level, while, unless we adopt a professional and scientific stance, we experience everything connected with a system as an external, impersonal force that impinges on us. Might it not be, I wondered, that we are so easily persuaded to trust a predictive algorithm because it reaches us on a personal level, while we distrust the digital system, whatever we mean by it or associate with it, because it is perceived as impersonal?

In science, we speak about different levels, organized in hierarchical ways, with each level following its own rules or laws. In the social sciences, including economics, the gap persists in the form of a micro-level and macro-level divide. But none of the epistemological considerations that follow seemed to provide what I was looking for: a way of seeing across these divides, either by switching perspectives or, much more challenging, by trying to find a pluri-perspectival angle that would allow me access to both levels. I have therefore tried to find a way to combine the personal and the impersonal, the effect of predictive algorithms on us as individuals and the effects that digitalization has on us as societies.

Although most of this book was written before a new virus wreaked havoc around the globe, exacerbated by the uncoordinated and often irresponsible policy response that followed, it is still marked by the impact of the COVID-19 pandemic. Unexpectedly, the emergence of the coronavirus crisis revealed the limitations of predictions. A pandemic is one of those known unknowns that are expected to happen. It is known that more are likely to occur, but it is unknown when and where. In the case of the SARS-CoV-2 virus, the gap between the predictions and the lack of preparedness soon became obvious. We are ready to blindly follow the predictions algorithms deliver about what we will consume, our future behaviour and even our emotional state of mind. We believe what they tell us about our health risks and that we should change our lifestyles. They are used for police profiling, court sentencing and much more. And yet we were unprepared for a pandemic that had been long predicted. How could this have gone so wrong?

Thus the COVID-19 crisis, itself likely to turn from an emergency into a more chronic condition, strengthened my conviction that the key to understanding the changes we are living through is linked to what I call the paradox of prediction. When human behaviour, flexible and adaptive as it is, begins to conform to what the predictions foretell, we risk returning to a deterministic world, one in which the future has already been set. The paradox is poised at the dynamic but volatile interface between present and future: predictions are obviously about the future, but they act directly on how we behave in the present.

The predictive power of algorithms enables us to see further and to assess the various outcomes of emergent properties in complex systems obtained through simulation models. Backed by vast computational power, and trained on an enormous amount of data extracted from the natural and social world, we can observe predictive algorithms in action and analyse their impact. But the way we do this is paradoxical in itself: we crave to know the future, but largely ignore what predictions do to us. When do we believe them and which ones do we discard? The paradox stems from the incompatibility between an algorithmic function as an abstract mathematical equation, and a human belief which may or may not be strong enough to propel us to action.

Predictive algorithms have acquired a rare power that unfolds in several dimensions. We have come to rely on them in ways that include scientific predictions with their extensive range of applications, like improving weather forecasts or the numerous technological products designed to create new markets. They are based on techniques of predictive analytics that have resulted in a wide range of products and services, from the analysis of DNA samples to predict the risk of certain diseases, to applications in politics where the targeting of specific groups whose voting profile has been established through data trails has become a regular feature of campaigning. Predictions have become ubiquitous in our daily lives. We trade our personal data for the convenience, efficiency and cost-savings of the products we are offered in return by the large corporations. We feed their insatiable appetite for more data and entrust them with information about our most intimate feelings and behaviour. We seem to have embarked on an irreversible track of trusting them. Predictive analytics reigns supreme in financial markets where automated trading and fintech risk assessments were installed long ago. They are the backbone of the military's development of autonomous weapons, the actual deployment of which would be a nightmare scenario.

However, the COVID-19 pandemic has revealed that we are far less in control than we thought. This is not due to faulty algorithms or a lack of data, although the pandemic has revealed the extent of grossly underestimating the importance of access to quality data and its interoperability. There was no need for predictive algorithms to warn of future epidemics; epidemiological models and Bayesian statistical reasoning were sufficient. But the warnings went unheard. The gap between knowing and doing persists if people do not want to know or offer many reasons to justify their inaction. Thus, predictions must also always be seen in context. They can fall on fallow ground or lure us into following them blindly. Predictive analytics, although couched in the probabilities of our ignorance, comes as a digital package that we gladly receive, but rarely see a need to unpack. They appear as refined algorithmic products, produced by a system that appears impenetrable to most of us, and often jealously guarded by the large corporations that own them.

Thus, the observations made during my patchy journey began to converge on the power of prediction and especially the power exerted by predictive algorithms. This allowed me to ask questions such as 'how does Artificial Intelligence change our conception of the future and our experience of time?' I could return to my long-standing involvement with the study of social time, and in particular the concept of *Eigenzeit*, which was the subject of a book I wrote in the late 1980s. A few years ago I followed up with 'Eigenzeit. Revisited', in which I analysed the changes introduced through our interaction with digital media and devices that had by then become our daily companions (Nowotny 2017). New temporal relationships have emerged with those who are physically distant but digitally close, so that absence and presence as well as physical and digital location have converged in an altered experience of time.

Neither I nor others could have imagined the meaning that terms like physical and social distancing would acquire only a few years later. In the midst of the COVID-19 pandemic, I saw my earlier diagnosis about an extended present confirmed. My argument had been that the line separating the present from the future was dissolving as the dynamics of innovation, spearheaded by science and technology, opened up the present to the many new options that were becoming available. The present was being extended as novel technologies and their social selection and appropriation had to be accommodated. Much of what had seemed possible only in a far-away future now invaded the present. This altered the experience of time. The present was becoming both compressed and densified while extending into the immediate future (Nowotny 1989).

What I observe now is that the future has arrived. We are living not only in a digital age but in a digital time machine. A machine fuelled by predictive algorithms that produce the energy to thrust us beyond the future that has arrived into an unknown future that we desperately seek to unravel. Hence, we scramble to compile forecasts and engage in manifold foresight exercises, attempting to gain a measure of control over what appears otherwise uncontrollable because of its unpredictable complexity. Predictive algorithms and analytics offer us reassurance as they lay out the trajectories for future behaviour. We attribute agency to them and feel heartened by the messages they deliver on the predictions that concern us most. Such is our craving for certainty that even in cases when the forecast is negative, we feel relieved that we at least know what will happen. In offering such assurance, algorithmic predictions can help us to cope with uncertainty and, at least partly, give us back some control of the future.

My background in science and technology studies (STS) allowed me to bridge the gap between science and society and reach a better understanding of the frictions and mutual misunderstandings that beset this tenuous and tension-ridden relationship. STS opens up the possibility of observing how research is actually carried out in practice and allows us to analyse the social structures and processes that underpin how science works. The pandemic has merely added a new twist, albeit a largely unfortunate one. While at the beginning of the pandemic science took centre-stage, combined with the expectation that a vaccine could soon be developed and therapeutic cures were in the pipeline, science soon became mired in political opportunism. A nasty ‘vaccine nationalism’ arose, while science was sidestepped by COVID-19 deniers and conspiracy theories that began to flourish together with anti-vax and extreme-right political movements. After a brief and bright interlude, the interface between science, politics and the public became troubled again.

The pandemic offered an advanced testing ground, especially for the biomedical sciences, whose recourse to Artificial Intelligence and the most recent digital technologies proved to be a great asset. It allowed them to sequence the genomes of the virus and its subsequent mutations in record time, with researchers sharing samples around the world and repurposing equipment in their labs to provide added test facilities. It enabled the COVID-19 High Performance Consortium, a public-private initiative with the big AI players and NASA on board, to aggregate the computing capability of the world’s fastest and most advanced computers. With the help of Deep Learning methods it was possible to reduce the 1 billion molecules analysed for potential therapeutic value to less than a few thousand.

The response to the pandemic also brought a vastly increased role for data. The pressure was enormous to proceed as quickly as possible with whatever data was available, in order to feed it into the simulation models that data scientists, epidemiologists and mathematicians were using to make forecasts. The aim was to predict the various trajectories the pandemic could take, plotting the rise, fall or flattening of curves and analysing the implications for different population groups, healthcare infrastructure, supply chains and the expected socio-economic collateral damage. Yet, despite the important and visible role given to data throughout the COVID-19 pandemic, no quick quantitative data-fix emerged that would provide a solid basis for the measures to be taken. If the data quality is poor or the right kind of data does not exist, a supposed asset quickly turns into garbage that contaminates simulation models and radically reduces their usefulness for society.

To some extent, the COVID-19 crisis has overshadowed the ongoing discussion about innovation and how scientific findings are transferred into society. It is therefore appropriate to recall the work of STS scholars who have extensively analysed the social shaping of technologies. Their findings show that technologies are always selectively taken up. They are gendered. They are appropriated and translated into products around which new markets emerge that give another boost to global capitalism. The benefits of technological innovation are never equally distributed, and already existing social inequalities are deepened through accelerated technological change. But it is never technology alone that acts as an external force bringing about social change. Rather, technologies and technological change are the products and the outcome of societal, cultural and economic preconditions and result from many co-productive processes.

Seen from an STS perspective, what is claimed to be entirely novel and unique calls for contextualization in historical and comparative terms. The current transformation can be compared to previous techno-economic paradigm shifts that also had profound impacts on society. In the age of modernity, progress was conceived as being linear and onedirectional. Spearheaded and upheld by the techno-sciences, the belief was that continued economic growth would assure a brighter and better future. It came with the promise of being in control, manifest in the overconfidence that was projected into planning. This belief in progress has, however, been on the wane for some time, and more recently many events and developments have injected new doubts. The destruction of the natural environment on a global scale confronts all of us with an ‘inconvenient truth’, reconfirmed by the Fridays for Future movement that has galvanized the younger generation. In addition, the pandemic has demonstrated the helplessness of many governments and the cynicism of their responses, while coping with the long-term consequences will require a change in direction.

The remarkable speed of recent advances in AI and its convergence with the sustainability crisis invites the question: What is different this time? We are already becoming conscious of the limitations of our spatial habitat, and face multiple challenges when it comes to using the available resources in a sustainable manner. These range from managing the transition to clean energy, to maintaining biodiversity and making cities more liveable, to drastically curbing plastic pollution and managing the increasing amount of waste. No wonder there is a growing concern that the control we can exert will be further diminished. The machines we have created are expected to take over many jobs currently performed by humans, but our capacity for control will shrink even further because these machines will monitor and limit our actions and possibilities. For these reasons much wisdom will be needed to better understand how AI affects and limits human agency.

I soon realized that I had touched only the surface of deeper transformational processes that we will have to think about together. The future will be dominated by digital technologies while we simultaneously face a sustainability crisis, and both of these transitions are linked with changes in the temporal structures and regimes that shape our lives and society. Digital technologies bring the future into the present, while the sustainability crisis confronts us with the past and challenges us to develop new capabilities for the future. Whatever solutions we come up with must integrate the human dimension and our altered relationship to the natural and technologically transformed environment. These were some of the underlying questions that kept me going, humming quietly but persistently in the background while I continued my search. My journey took me to a number of international meetings, workshops and conferences where some of these issues were discussed. For example, there were meetings on how to protect rights to privacy, which received special legal status in Europe through the General Data Protection Regulation (GDPR). Europe is perceived to play only a side role in the geopolitical competition between the two AI superpowers, the United States and China, a competition sometimes referred to as the digital arms race for supremacy in the twenty-first century, and which has recently been rekindled in alarming ways. Many Europeans take solace in the fact that they at least have a regulatory system to protect them, even if they acknowledge that neither the GDPR nor other forms of vigilance against intrusion by the large transnational corporations are sufficient in practice.

Other items on the agenda of discussion fora about digitalization were concerned with the risks arising from the ongoing processes of automation. Foremost was the burning issue of the future of work and the potential risks that digitalization entails for liberal democracies. It seemed to me that the fear that more jobs would be lost than could be created in time was being felt much more strongly in the United States than in Europe, partly due to still-existing European welfare provisions and partly because digitalization had not yet visibly hit professionals and the middle class. The threats to liberal democracies became more apparent when populist, nationalist and xenophobic waves swept across many countries. They were nurtured by sinister phenomena such as ‘fake news’ and Trojan horses, with unknown hackers and presumed foreign secret services engaged in micro-targeting specific groups with their made-up messages. More generally, they appeared intent on undermining existing democratic institutions while supporting political leaders with authoritarian tendencies. Digital technologies and social media were being appropriated as the means to erode democratic principles and the rule of law, while the internet, it seemed, had turned into an unrestrained and unregulated space for the diffusion of hate and contempt.

My regular visits to Singapore provided a different angle on how societies might embrace digitalization, and a unique opportunity to observe a digitally and economically advanced country in action. I gathered insights into Singapore's much-vaunted educational system, and observed the reliance of the bureaucracy on digital technologies but also its high standards of efficiency and maintenance of equally high levels of trust in government. What impressed me most, however, was the country's delicate and always precarious balance between a widely shared sense of its vulnerability – small, without natural resources and surrounded by large and powerful neighbours – and the equally widely shared determination to be well prepared for the future. Here was a country that perceived itself as still being a young nation, drawing much of its energy from the remarkable economic wealth and social well-being it has achieved. This energy now had to be channelled into a future it was determined to shape. Nowhere else did I encounter so many debates, workshops, reports and policy measures focused on a future that, despite remaining uncertain, was to be deliberated and carefully planned for, taking in the many contingencies that would arise. Obviously, it would be a digital future. The necessary digital skills were to be cultivated and all available digital tools put to practical use.

More insights and observations came from attending international gatherings on the future of Artificial Intelligence. In my previous role as President of the European Research Council (ERC), I participated in various World Economic Forum meetings. The WEF wants to be seen as keenly engaged in digital future building. At the meetings I attended, well-known figures from the world of technology and business mingled with academics and corporate researchers working at the forefront of AI. It was obvious that excitement about the opportunities offered by digital technologies had to be weighed against their possible risks if governments and the corporate world wanted to avert a backlash from citizens concerned about the pace of technological change. The many uncertainties regarding how this would be played out were recognized, but the solutions offered were few.

Other meetings in which I participated had the explicit aim of involving the general public in a discussion about the future of AI, such as the Nobel Week Dialogue 2015 in Gothenburg, or the Falling Walls Circle in Berlin in 2018. There were also visits to IT and robotics labs and workshops tasked with setting up various kinds of digital strategies. I gained much from ongoing discussions with colleagues at the Vienna Complexity Science Hub and members of their international network, allowing me glimpses into complexity science. By chance, I stumbled into an eye-opening conference on digital humanism, a trend that is gradually expanding to become a movement.

Scattered and inconclusive as these conversations mostly were, they nevertheless projected the image of a dynamic field rapidly moving forward. The main protagonists were eager to portray their work as incorporating their responsibility of moving towards a 'beneficial AI' or similar initiatives. There was a notable impatience to demonstrate that AI researchers and promoters were aware of the risks involved, but the line between sincere concern and the insincere attempts of large corporations to claim 'ethics ownership' was often blurred as well. Human intelligence might indeed one day be outwitted by AI, but the discussants seldom dwelt on the difference between the two. Instead, they offered reassurances that the risks could be managed. Occasionally, the topic of human stupidity and the role played by ignorance were touched upon as well. And at times, a fascination with the 'sweetness of technology' shimmered through, similar to that J. Robert Oppenheimer described when he spoke about his infatuation with the atomic bomb.

At one of the many conferences I attended on the future of AI, the organizers had decided to use an algorithm in order to maximize diversity within each group. The AI was also tasked to come up with four different haikus, one for each group. (Incidentally, the first time an AI succeeded in accomplishing such a 'creative' task was back in the 1960s.) The conference was a success and the discussions within each 'haiku group' were rewarding, but somehow I felt dissatisfied with the haiku the AI had produced for my group. So, on the plane on my way back I decided to write one myself – my first ever. With beginner's luck the last line of my haiku read 'future needs wisdom'.

A haiku is said to be about capturing a fleeting moment, a transient impression or an ephemeral sensation. My impressions were obviously connected to the theme of the conference, the future of AI. 'Future needs wisdom' – the phrase stuck with me. Which future was I so concerned about? Would it be dominated by predictive algorithms? And if so, how would this change human behaviour and our institutions? What could I do to bring some wisdom into the future? What I have learned on my journey in digi-land is to listen carefully to the dissonances and overtones and to plumb the nuances and halftones; to spot the ambiguities and ambivalences in our approaches to the problems we face, and to hone the ability to glide between our selective memories of the past, a present that overwhelms us and a future that remains uncertain, but open.

The maze and the labyrinth

None of these encounters and discussions prepared me for the surprise I got when I began to scan the available literature more systematically. There is a lot of it out there already, and a never-ending stream of updates that keep coming in. I concluded that much of it must have been written in haste, as if trying to catch up with the speed of actual developments. Sometimes it felt like being on an involuntary binge, overloaded with superfluous information while feeling intellectually undernourished. Most striking was the fact that the vast majority of books in this area espouse either an optimistic, techno-enthusiastic view or a dystopian one. They are often based on speculations or simply describe to a lay audience what AI nerds are up to and how digital technologies will change people's lives. I came away with a profound dissatisfaction about how issues and topics that I considered important were being treated: the approach was largely short-term and ahistorical, superficial and mostly speculative, often espousing a narrow disciplinary perspective, unable to connect technological developments with societal processes in a meaningful way, and occasionally arrogant in dismissing 'the social' or misreading it as a mere appendix to 'the technological'.

Plenty of books on AI and digitalization continue to flood the market. Most of the literature is written in an enthusiastic, technology-friendly voice, but there is also a sharp focus on the dark side of digital technologies. The former either provide a broad overview of the latest developments in AI and their economic benefits, or showcase some recently added features that are intended to alleviate fears that the machines will soon take over. The social impact of AI is acknowledged, as is the desirability of cross-disciplinary dialogue. A nod towards ethical considerations has by now become obligatory, but other problems are sidestepped and expected to be dealt with elsewhere. Only rarely, for instance, do we hear about topics like digital social justice. Finding my way through the copious literature on AI felt at times like moving through a maze, a deliberately confusing structure designed to prevent escape.

In this maze there are plenty of brightly lit pathways, their walls lined with the latest gadgetry, proudly displaying features designed to take the user into a virtual wonderland.

The darker groves in the maze are filled with images and dire warnings of worse things to come, occasionally projecting a truly apocalyptic digital ending. Sci-fi occupies several specialized niches, often couched in an overload of technological imagination and an underexposed social side. In between there are a large number of mundane small pathways, some of which turn out to be blind alleys. One can also find useful advice on how to cope with the daily nitty-gritty annoyances caused by digital technologies or how to work around the system. Plenty of marketing pervades the maze, conveying a sense of short-lived excitement and a readiness to be pumped up again to deliver the next and higher dose of digital enhancement.

At times, I felt that I was no longer caught in a maze but in what had become a labyrinth. This was particularly the case when the themes of the books turned to 'singularity' and transhumanism, topics that can easily acquire cult status and are permeated by theories, fantasies and speculations that the human species will soon transcend its present cognitive and physical limitations. In contrast to a maze with its tangled and twisted features, dead ends and meandering pathways, a labyrinth is carefully designed to have a centre that can be reached by following a single, unicursal path. It is artfully, and often playfully, arranged around geometrical figures, such as a circle or a spiral. No wonder that labyrinths have inspired many writers and artists to play with these forms and with the meaning-laden concept of a journey. If the points of departure and arrival are the same, the journey between them is expected to have changed something during the course of it. Usually, this is the self. Hence the close association of the labyrinth with a higher state of awareness or spiritual enlightenment.

The labyrinth is an ancient cultic place, symbolizing a transformation, even if we know little about the rituals that were practised there. In the digital age, the imagined centre of the digital or computational labyrinth is the point where AI overtakes human intelligence, also called the singularity. At this point the human mind would be fused with an artificially created higher mind, and the frail and ageing human body could finally be left behind. The body and the material world are discarded as the newborn digital being is absorbed by the digital world or a higher digital order. Here we encounter an ancient fantasy, the recurring dream of immortality born from the desire to become like the gods, this time reimagined as the masters of the digital universe. I was struck by how closely the discussion of transcendental topics, like immortality or the search for the soul in technology, could combine with very technical matters and down-to-earth topics in informatics and computer science. I seemed that the maze could transform itself suddenly into a labyrinth, and vice versa.

In practice, however, gaps in communication prevail. Those who worry about the potential risks that digital technologies pose for liberal democracies discover that experts working on the risks have little interest in democracy or much understanding of politics. Those writing on the future of work rarely speak to those engaged in the actual design of the automated systems that will either put people out of work or create new jobs. Many computer scientists and IT experts are clearly aware of the biases and other flaws in their products, and they deplore the constraints that come from being part of a larger technological system. But at heart they are convinced that the solutions to many of the problems besetting society will arise from technology. Meanwhile, humanists either retreat to their historical niche or act in defence of humanistic values. The often-stated goal of interdisciplinarity, it seems, is not yet much advanced in practice.

I came away from the maze largely feeling that it is an overrated marketplace where existing products are rapidly displaced by new ones selected primarily for their novelty value. Depending on the mood of potential buyers, utopian or dystopian visions would prevail, subject to market volatility. The labyrinth, of course, is a more intriguing and enchanting place where deep philosophical questions intersect with the wildest speculations. Here, at times, I felt like Ariadne, laying out the threads that would lead me out from the centre of the labyrinth. One of these threads is based on the idea of a digital humanism, a vision that human values and perspectives ought to be the starting point for the design of algorithms and AI systems that claim to serve humanity. It is based on the conviction that such an alternative is possible.

Another thread is interwoven with the sense of direction that takes its inspiration from a remarkable human discovery: the idea of the future as an open horizon, full of as yet unimaginable possibilities and inherently uncertain. The open horizon extends into the vast space of what is yet unknown, pulsating with the dynamics of what is possible. Human creativity is ready to explore it, with science and art at the forefront. It is this conception of the future which is at stake when predictive algorithms threaten to fill the present with their apparent certainty, and when human behaviour begins to conform to these predictions.

The larger frame of this book is set by a co-evolutionary trajectory on which humankind has embarked together with the digital machines it has invented and deployed. Co-evolution means that a mutual interdependence is in the making, with flexible adaptations on both sides. Digital beings or entities like the robots created by us are mutating into our significant Others. We have no clue where this journey will lead or how it will end. However, in the long course of human evolution, it is possible that we have become something akin to a self-domesticating species that has learned to value cooperation and, at least to some extent, decrease its potential for aggression. That capacity for cooperation could now extend to digital machines. We have already reached the point of starting to believe that the algorithm knows us better than we know ourselves. It then comes to be seen as a new authority to guide the self, one that knows what is good for us and what the future holds.

The road ahead: how to live forward and understand life backwards

Scientific predictions are considered the hallmark of modern science. Notably physics advances by inventing new theoretical concepts and the instruments to test predictions derived from them. The computational revolution that began in the middle of the last century has been boosted by the vastly increased computational power and Deep Learning methods that took off in the twenty-first century. Together with access to an unprecedented and still growing amount of data, these developments have extended the power of predictions and their applicability across an enormous range of natural and social phenomena. Scientific predictions are no longer confined to science.

Ever since, predictive analytics has become highly profitable for the economy and pervaded the entire social fabric. The operation of algorithms underlies the functioning of technological products that have disrupted business models and created new markets. Harnessed by the marketing and advertisement industry, instrumentalized by politicians seeking to maximize votes, and quickly adopted by the shadowy world of secret services, hackers and fraudsters exploiting the anonymity of the internet, the use of predictive analytics has convinced consumers, voters and health-conscious citizens that these powerful digital instruments are there to serve our needs and latent desires.

Much of their successful spread and eager adoption is due to the fact that the power of predictive algorithms is performative. An algorithm has the capability to make happen what it predicts when human behaviour follows the prediction. Performativity means that what is enacted, pronounced or performed can affect action, as shown in the pioneering work on the performativity of speech acts and non-verbal communication by J. L. Austin, Judith Butler and others. Another well-known social phenomenon is captured in the confidence among AI practitioners that work on ethical AI is progressing well. The tacit assumption is that the dark side of digital technologies and all the hitherto unresolved problems will also be sorted out by an ultimate problem-solving intelligence, a kind of far-sighted, benign Leviathan fit to manage our worries and steer us through the conflicts and challenges facing humanity in the twenty-first century.

The other line of thinking insists that theoretical understanding is necessary and urgent, not only for mathematicians and computational scientists, but also for developing tools to assess the performance and output quality of Deep Learning algorithms and to optimize their training. This requires the courage to approach the difficult questions of ‘why’ and ‘how’, and to acknowledge both the uses and the limitations of AI. Since algorithms have huge implications for humans it will be important to make them fair and to align them with human values. If we can confidently predict that algorithms will shape the future, the question as to which kinds of algorithms will do the shaping is currently still open (Wigderson 2019).

Understanding also includes the expectation that we can learn how things work. If an AI system claims to solve problems at least as well as a human, then there is no reason not to expect and demand transparency and accountability from it. In practice, we are far from receiving satisfactory answers as to how the inner representations of AI work in sufficient detail, let alone an answer to the question of cause and effect. The awareness begins to sink in that we are about to lose something connected to what makes us human, as difficult to pin down as it is. Maybe the time has come to admit that we are not in control of everything, to humbly concede that our tenuous and risky journey of co-evolution with the machines we have built will be more fecund if we renew our attempt to understand our shared humanity and how we might live together better. We have to continue our exploration of living forward while trying to understand Life backwards and linking the two. Prediction will then no longer only map the trajectories of living forward for us, but will become an integral part of understanding *how* to live forward. Rather than foretelling *what* will happen, it will help us understand *why* things happen.

After all, what makes us human is our unique ability to ask the question: *Why do things happen – why and how?*