Existential Risk and the Technological Understanding of Being

Kim Caspar Hecker

St Edmund's College, Cambridge

This dissertation is submitted for the degree of Doctor of Philosophy

Department of Politics and International Studies
University of Cambridge

November 2018

Page intentionally left blank

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

It does not exceed the prescribed word limit for the relevant Degree Committee.

Page intentionally left blank

Abstract

Existential Risk and the Technological Understanding of Being

Kim Caspar Hecker

‘Existential risk research’ or ‘existential risk studies’ is an emerging, interdisciplinary genre that seeks to provide an integrative, scientific framework for the study of existential dangers to humanity. Having been introduced by Oxford Philosopher Nick Bostrom in the early 2000s, existential risk research over the past ca. 15 years has become increasingly popular amongst scientists from a wide variety of academic disciplines and recent years have seen the foundation of research institutes, dedicated exclusively to the study of existential risk, at some of the most prestigious research universities in Europe and the United States. In spite of its interdisciplinary character, neither history nor political theory play a prominent role in existential risk research. This dissertation argues that this is a regrettable state of affairs and presents the first systematic attempt to survey and frame existential risk research from a political thought perspective. Drawing on three authors who wrote about deeply related questions in the post-war decades - Martin Heidegger, Hannah Arendt and Günther Anders - it contextualises existential risk studies in the light of long-standing discussions about the interrelations between modern technology, human value and human agency under existential conditions. At the heart of these discussions the dissertation identifies a range of ontological complications. It demonstrates that, despite the fact that existential risk scholarship tends to side-line the type of ontological problems that have been uncovered by Heidegger, Arendt and Anders, it cannot escape this dimension altogether but instead highlights the imminent relevance of these authors’ analyses. One instance in which this becomes particularly salient is in the context of existential fears surrounding artificial intelligence. The dissertation therefore closes with a discussion of the issue of artificial intelligence in existential risk research, bringing together insights from the preceding chapters.

Page intentionally left blank

Acknowledgments

I want to express my gratitude to the Friedrich Naumann Foundation for Freedom and the German Federal Ministry for Education and Research for providing the funds which allowed me to embark on and realise this project. I equally want to thank Cambridge University's Economic and Social Research Council Doctoral Training Partnership for their generous support over the past four years.

There are many people whose help and criticism has been essential in writing this dissertation. I would like to thank my supervisor, David Runciman, for his detailed guidance, his support, and many inspiring conversations throughout this journey. I also want to thank my examiners Casper Sylvest and Duncan Bell for their highly instructive and encouraging feedback. Furthermore, I am very grateful that I had the opportunity to discuss my ideas with many inspiring people from a wide variety of disciplinary backgrounds over the past years. Specifically, I would like to express my gratitude to Duncan Kelly, Yasmin Shearmur, Baruch Malewich, Rodrigo Garcia-Velasco, Jens van't Klooster, and Hardy Schilgen for reading and discussing my work with me.

My greatest debt is to those who made this dissertation possible in the most fundamental sense. I would like to thank my wonderful family, Bettina Lange-Hecker, Gereon Hecker, and Felix Hecker, whose intellectual curiosity has always been an inspiration for me and on whose encouragement and support I could always count, not only whilst writing this dissertation but ever since I can remember. I am forever grateful for that. And, finally, I would like to thank Charlotte, without whose sharp mind, continuous, challenging feedback, and positive presence in my life quite simply none of this would have been possible.

Page intentionally left blank

Contents

Note on transliteration and text

This thesis uses double-inverted commas (“”) for quotations. Indirect quotations are marked by single inverted commas (‘’). Further, single inverted commas (‘’) are used to emphasise that concepts are invoked which have a specific meaning within the works of authors covered in this thesis.

Several of Günther Anders’ works have not yet been translated into English. For this reason, I translated some passages myself. Wherever I did so, this is indicated in the footnotes with the remark ‘translated by the author’. Typically, the original German text passage is not quoted in the footnotes. This is only done where the translated passage is longer than three lines.

Abbreviations

Existential risk institutes

BERI:	Berkeley Existential Risk Initiative
CHAI:	Center for Human Compatible Artificial Intelligence
CSER:	Centre for the Study of Existential Risk
FHI:	Future of Humanity Institute
FLI:	Future of Life Institute
FRI:	Foundational Research Institute
GCF:	Global Challenges Foundation
GCRI:	Global Catastrophic Risk Institute
GPI:	Global Priorities Institute
IEET:	Institute for Ethics and Emerging Technologies
LCFI:	Leverhulme Centre for the Future of Intelligence
MIRI:	Machine Intelligence Research Institute
OPP:	Open Philanthropy Project
PFHF:	Project for Future Human Flourishing

Introduction

Over the past 15 years approximately, a new genre of scientific study has emerged, which will hereafter be referred to as ‘existential risk research’ or ‘existential risk studies’. In the field, an existential risk is commonly defined as “one that threatens the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development”.[1] Developed by Oxford philosopher Nick Bostrom in the early 2000s, the concept was gradually adopted by a growing community of scholars. Over time a movement emerged, comprising a growing range of institutes and activist groups, some outside of academia, others affiliated to globally leading universities, all united by a drive to study the most serious hazards threatening humanity along the lines pioneered by Bostrom. In short, we appear to be witnessing the emergence of a self-declared, new research community that converges around a range of shared concerns, concepts, canonical works and convenes at shared platforms such as the Cambridge Conference on Catastrophic Risk, or the Colloquium on Catastrophic and Existential Risk at the University of California, Los Angeles.[2]

This thesis sets out to explore existential risk studies as a subject of critical reflection from a political thought perspective. The existential risk community is a highly diverse and interdisciplinary community, or “eco-system”, as one of its many prominent members, Skype founder Jaan Tallinn, refers to it.[3] It enjoys the support, financial and otherwise, of some of the world’s most prominent technology tycoons such as the above-mentioned Jaan Tallinn, but also Elon Musk, Bill Gates, Reid Hoffmann, or Peter Thiel. Furthermore, several of the academics who are or have been active in the field are comparably well known, including for instance physicists such as the late Stephen Hawking, Astronomer Royal Martin Rees, Oxford physicist David Deutsch, or Nobel laureate Frank Wilczek, philosophers such as Huw Price and Peter Singer, or leading computer scientists such as Berkeley’s Stuart Russell, Murray Shanahan, robotics professor at Imperial College London, and MIT’s Eric Brynjolfsson. This list of but a few of its most prominent members illustrates the degree of academic interdisciplinarity which is characteristic of the field. However, after a closer look, it quickly becomes clear that scholars who contribute to the study of existential risk bring together expertise from public policy scholarship, computer science, law, economics, astrophysics, philosophy of science, moral philosophy, mathematics, statistics, ethics, synthetic biology, and many other disciplines.

In spite of its interdisciplinary character, neither history nor political theory appear to play a prominent role in existential risk research. To an extent this may not be surprising. The mission of

[1] See Bostrom, N. (2002), p. 2 or Bostrom, N. (2013), p. 15.

[2] See respectively CSER (2016), Lin, F. (2017).

[3] Jaan Tallinn as quoted in CSER (2016), and Torres, P. (2017b), p. 2.

existential risk research is an avowedly practical one. It seeks to study and raise attention for the worst possible threats, present and future, humanity might be exposed to, provide an integrative, scientific framework for their analysis, and to make their concerns heard in the upper echelons of the corporate world, of academia, civil society, as well as on the highest levels of national government and international governmental institutions, in an effort to minimise the overall threat level. Political theory and history, being largely reflective disciplines, do not necessarily yield themselves to such pragmatic purposes.

This thesis contends, however, that the lack of historical and political-theoretical perspective in contemporary discourses about existential risk is a nonetheless regrettable state of affairs. A side-effect of the relative absence of history and political theory in existential risk research is that authors in the field believe they are virtually alone with their concerns. Martin Rees (2017) for instance claims that existential risks "have hitherto been seriously addressed by only a small community of serious thinkers" and "that there needs to be a much-extended research program, involving natural and social scientists, to compile a more complete register of possible 'x-risks'".[4] Nick Bostrom (2009) argues in a similar vein that "existential risks have not received as much scholarly attention as they deserve. In recent years, there have been three serious books and one major paper on this topic".[5] Simply put, the aim of this thesis is to develop a better understanding of such claims. It asks if existential risk research is as new as authors in the field appear to believe it is, and, if so, in what respects.

Bringing a historically informed political theory perspective to bear on existential risk research allows us to answer that question in two ways. First, the analytical use of historical concepts allows us to identify and outline a range of, as of yet underarticulated, central theoretical problems and puzzles at its heart. Existential risk research, in spite of its interdisciplinary character, appears to be curiously isolated within the wider academic world. That is, even though the scholars who contribute to existential risk research have a wide variety of academic backgrounds, the output of existential risk research, which hereafter will be referred to as *existential risk theory*, is not widely discussed outside of the existential risk ecosystem itself. Whilst existential risk research imports methods and insights from a multitude of different disciplines for its practical aims, it has itself not yet been uncovered as a subject of scholarly attention and interlocution from an external perspective. For instance, apart from two references to existential risk theory by international relations (IR) scholars,[6] who peripherally touch upon the literature in the context of discussions about challenges to IR theory presented by the Anthropocene concept, existential risk theory has not yet been discussed in political science scholarship. Similar observations can be made for most other disciplines in the humanities and social sciences. A result of this absence of external interlocution is that existential

[4] Rees, M. (2017a), p. iv.

[5] Bostrom, N. (2009), p. 196.

[6] See Harrington, C. (2016) and Mitchell, A. (2017).

risk research has been subjected to little theoretical scrutiny. Once we expose it to such scrutiny from a historically informed political theory perspective, it becomes clear that, beneath its pragmatic, technical, argumentative surface, existential risk research presents us with an interesting and rather distinct take on the problematic interface between modern technology, human agency and human value.

This interface, however, has been the subject matter of political-theoretical and, widening the focus beyond that, philosophical reflection for centuries. From a historically informed perspective, it quickly becomes clear that the puzzles with which existential risk research confronts us, puzzles in relation to the above-mentioned interface, can be meaningfully connected to long standing discussions in political theory and philosophy. Situating existential risk research within such temporally-extended debates allows us to develop a better understanding of existential risk research as occupying a historically and conceptually contingent position within them. It further allows us to see that the idea that existential risk research presents a fundamentally new and neglected field of inquiry, is intelligible only from within the applications-focused framework of existential risk research itself, i.e. when judged by its own standards. When one looks beneath the surface and takes the deeper questions of its core concerns as a benchmark, then the idea that existential risk research occupies an otherwise unpopulated space of intellectual activity speaks, if anything, to a lack of knowledge about itself. A historical perspective on the topic thus allows us to frame existential risk research as a historically and theoretically contingent phenomenon, as a type of response to a range of concerns, where both the type of response and the concerns have precursors in history. At the same time, however, it allows us to identify facets in existential risk research, in its language, its concepts, its methodology, etc., that indeed do appear to be new and which offer partially new perspectives on the deeper problems and puzzles at its core.

In sum, the relative absence of history and political theory in existential risk research is regrettable for at least two reasons. Firstly, because existential risk research and our understanding of the problems it has uncovered can gain in nuance by being connected to older debates in political theory and, secondly, because the same holds in the opposite direction. Existential risk theory provides us with an opportunity to re-examine and re-appreciate the striking relevance of older debates in a new light, highlighting their lasting insightfulness in present circumstances.

This thesis suggests that one suitable starting point for the historical and theoretical embedding of existential risk research can be found in the post-World War II works of three authors: Martin Heidegger, Hannah Arendt and Günther Anders. Emerging from within its pragmatic framework, existential risk research presents us with a distinctive perspective on the future of humanity, namely one in which the future is transformed into a technological optimisation problem. At a later point in the thesis, this perspective on the future of humanity will be referred to as one in which technology emerges as 'humanity's destiny' in the sense that the space of technological possibilities becomes our benchmark, the defining lens through which we envision potential futures.

What opens up beneath the language of risk and risk management is thus an argument about, or rather a range of assumptions and propositions on, the status of human agency and of the human being under conditions of modern technology. However, in existential risk research itself, due to the relative absence of history and political theory, this complex of assumptions, propositions, etc., and their deeper significance in the context of existential risk tends to remain unquestioned, unexplored, and largely abstracted from. The works of Heidegger, Arendt and Anders provide us with a rich conceptual repertoire to not only contour this complex, but also to pinpoint the puzzles and philosophical problems underpinning it. They therefore help to identify and offer new perspectives on the underlying issues in existential risk research. Writing against the backdrop of the emergence of the arguably first human-made existential risk – the ever present possibility of an all-out nuclear war during the cold war period – and in the shadow of the horrors of the second World War, Heidegger, Arendt and Anders identified in technology a fateful force, one impacting the very fundamentals, physical and ideational, of human existence and that was in the process of transforming the human condition on its own terms. A non-negligible proportion of their post-War works can be read as philosophical efforts to understand the roots and implications of that process – in Heidegger's case from a metaphysical perspective, in Arendt's from a political perspective, and in Anders' from what he refers to as a "philosophical anthropological" perspective.[7]

This thesis seeks to demonstrate that Heidegger, Arendt and Anders were, broadly speaking, already observing the same phenomenon, the same transformational process that existential risk researchers today seek to raise attention for, but that they were approaching and writing about it from an inverse perspective, namely an interpretive one. Where existential risk research's analysis seeks to stress and direct the imminent need for action in response to the existential threats humanity is facing, Heidegger, Arendt and Anders were first and foremost engaged in an effort to understand the transformation of the human condition which this novel type of need for action was expressive of. What they identified in what we now call existential risk was, roughly speaking, one of the starkest manifestations of an ontologically rooted, alienating dynamic, whereby modern technology was placing humanity in an increasingly schizophrenic condition, left to occupy two fundamentally different realities: one of everyday sense-experience, language, thought, commitments, common-sense, intuition, etc., and one of abstract, technical knowledge, which becomes physical reality through modern technological objects and of which the possibility of collective nuclear self-annihilation was both a consequence and an instance. The problematic relationship between these two realities is the common theme of Heidegger's, Arendt's and Anders' thought on modern technology and science, and their political and philosophical ramifications. At its centre is the problem that, in the midst of this alienating dynamic, the human being emerges as an increasingly paradoxical figure: simultaneously omnipotent and hopelessly outdated; both becoming aware of the

[7] See Anders, G. (1992), p. 9.

unique preciousness of life on Earth, and attaining the capability to annihilate it; at once separated from and elevated above the natural realm by abstract, objectifying knowledge of it and reduced to it by that very same knowledge.

Existential risk research presents us with several such paradoxes and dilemmas. Most importantly, it appears to want to save something, humanity, based on a logic that cannot make sense of this concept to begin with. By infusing existential risk research with history and political theory we are in a position to identify and unravel these deeper puzzles at its heart. This allows us both to develop a more comprehensive and more nuanced perspective on the extremes of our contemporary technological predicament as sketched out by existential risk research, as well as to re-examine and reappreciate the actuality of Heidegger's, Arendt's and Anders' thought in the context of a new genre of scholarship.

The basis for the systematic integration of existential risk research into older debates in political theory and philosophy, consists in a survey of 'existential risk literature'. By surveying texts from monographs, to academic articles, working papers, conference papers, blog posts, whitepapers, newspaper and magazine articles, web pages, etc., of existential risk researchers and institutes and connecting them to one-another, the thesis seeks to extract the central themes, arguments, claims and assumptions, and to identify a shared theoretical framework which can then be connected and related to older traditions of thought, in this case central writings of Hannah Arendt, Günther Anders and Martin Heidegger.

To be more precise, in developing the above arguments, I proceed in four steps. Chapter 1 provides an overview and a preliminary analysis of the emerging genre of existential risk research. It draws on key texts in the field, most importantly the publications of Nick Bostrom who is largely recognised as the central figure in the field, both because of his role as founder and director of Oxford University's Future of Humanity Institute as well as because of the central position his publications assume in the writings of most other existential risk scholars. Bostrom's writings are complemented with texts published by several other authors in the field in order to provide a more comprehensive overview of existential risk research and its theoretical framework. The chapter discusses existential risk research's ethics, its methodology and mission, its policy recommendations, and traces the scope and shape of its 'eco-system' in form of research institutes, individual researchers, etc. The second half of the chapter presents the first analytical step of the thesis. It seeks to bring out the deeper arguments and ramifications of existential risk research and highlights the pivotal position technology assumes in it, demonstrating how, by framing the future in terms of potential end-time scenarios, technology emerges as humanity's destiny.

Having established that the problematic relationship between technology and human agency can be identified as the central theme in existential risk research, chapter 2 proceeds to connect existential risk research to Martin Heidegger's work. Heidegger's thought is a particularly suitable vantage point for contextualising existential risk research for a variety reasons. From a historical perspective, it seems sensible for at least two reasons. First, Heidegger was amongst the very first thinkers who began to systematically study technology as a philosophical problem in its own right.[8] Secondly, Heidegger attributed a comparably pivotal role to technology as existential risk research does today. Like existential risk research, he found in technology both a potential destiny of humanity as well as a source of gravest danger, providing us with a basis to relate existential risk research to older traditions of thought, thinking about technology along, broadly speaking, existential lines. In a way, we might argue that both occupy the same intellectual space when it comes to their thinking about technology, a space where technology is framed as perhaps the single most important force in human history. Historically speaking, then, Heidegger's philosophy of technology provides us with a touchstone to locate existential risk research in a spectrum of traditions of thought surrounding the role of technology in human life and develop a better idea of what might be new or distinctive about it.

Conceptually, Heidegger's philosophy is a suitable starting point both because his phenomenologically and metaphysically rooted critique of technology provides us with highly critical perspectives on existential risk research, allowing us to philosophically 'unpack' the notion of 'technology' and thus to uncover some of the defining intellectual puzzles existential risk research confronts us with. It is also a fruitful starting point because he is at the origin of a tradition of critical thinking about technology, continued by Hannah Arendt and Günther Anders, whose work forms the bedrock of the political-theoretical discussion of existential risk in chapter 3. Specifically, Heidegger's phenomenological method shows that the notion of technology itself cannot be employed 'neutrally', or purely pragmatically, as existential risk researchers tend to, but that it is inherently tied up with ontological problems which inevitably confront us with puzzles regarding the status of human agency and the idea of value under modern (technological) conditions.

Heidegger's ontological grounding of the problem of technology serves to introduce a useful juxtaposition, which will function as a theoretical bracket for the further political-theoretical embedding of existential risk theory – the apparent irreconcilability between phenomenal reality, as it appears 'naturally' to our consciousness on the one hand, and abstract, technical, scientific knowledge, which he calls the 'technological understanding of being', as well as its physical manifestation in the form of modern technological instruments and systems, on the other. A central argument this thesis advances is that existential risk research, in its lack of interest in historical and political-theoretical perspectives, tends to abstract from the normative complications that arise from the ontological tension Heidegger had uncovered. The thesis holds that the field ultimately cannot escape these complications for precisely the reasons presented by Heidegger and that were further developed by his students Hannah Arendt and Günther Anders. However, as we will see, the connection between existential risk research and Heidegger's philosophy of technology leads into an

[8] Viz. Mitcham, C. (1994).

aporetic ending. Where existential risk researchers stress the urgent need for action, thus highlighting the role of agency in the context of existential risk, Heidegger argues that any kind of action under the paradigm of 'the technological understanding of being' can only serve to aggravate the dangers humanity finds itself in. For Heidegger, the very concern for the survival of the species turned out to be expressive of the technological will for mastery. As a result of Heidegger's ontotheological perspective on technology we thus find a curious lack of interest in the problem of human extinction as a philosophical and ethical problem unique in its own right. Whilst Heidegger provides us with a basis to demonstrate that existential risk research's particular theoretical framework can be meaningfully connected to old controversies about the interrelations between technology, human agency and value, his holistic, ontotheological perspective on this complex of issues leads us into an impasse should we want to develop a better understanding of the particular political and philosophical ramifications the problem of human extinction entails, and to what extent existential risk research may or may not offer new perspectives on them.

Chapter 3 therefore turns to Heidegger's students Hannah Arendt and Günther Anders. Arendt and Anders can be shown to occupy a middle ground between Heidegger and existential risk research. Uniting fundamental insights of Heidegger's phenomenological critique of technology with a concern for the survival of the species and thus an awareness for the need for political action, Anders' and Arendt's work allow us to approach the problem of human extinction as a complex, multi-dimensional, political problem of uniquely transformative qualities, emerging from changing technological realities intertwined with problematic constellations of attitudes, irrational hopes and desires, and anachronistic conceptions of technology that determine how we respond to these realities in thought and action. As in the case of Heidegger, the study of Arendt's and Anders' work demonstrates that existential risk research is not a qualitatively new response to the increasing powers of technology. However, their interest in the complexities of human psychology and political life meant that they were in an arguably much better position than Heidegger to investigate this type of response as part of a world reconfiguring itself around the spectre of extinction, the emergence of which they witnessed as part of the nuclear conundrum of their days. Arendt's and Anders' post-war works thus are particularly insightful in the context of existential risk research because they add a level of analysis, demonstrating how the ontological puzzles we can identify at its core and the roots of which Heidegger had uncovered, are reflected in politically highly problematic imbalances between different human faculties – the capacity for action on the one hand and capacities such as understanding and imagination on the other hand side – and what these imbalances in turn might imply for our temporal consciousness, our self-understanding, our hopes, fears, and values, under conditions of existential risk.

In the final chapter of the thesis my discussion of existential risk in the light of Heidegger's, Arendt's and Anders' thought, is brought to bear on contemporary discussions surrounding artificial intelligence (AI), which presently occupy a prominent position in existential risk research. These debates surrounding AI, in particular so called ‘superintelligence’, it is argued, also do not appear to be qualitatively new. In a way they echo almost uncannily closely Heidegger’s concerns regarding the ‘thoughtless’ nature of the technological understanding of being. Furthermore, since AI is seen both as a potential threat to humanity as well as a potential solution to all of our problems, it also highlights some of the pathologies that characterise our daily interaction and thinking about technology, which Günther Anders had uncovered. In other words, in the context of AI the deep puzzles concerning the relationship between technology, human agency, and human value are present in the context of discussions of one particular technology. To the reader, however, this may not come as a surprise since speculations about the ultimate possibilities of AI are in effect no less than speculations about whether or not technology and humanity are interchangeable and therewith about whether or not (the reality, illusion, awareness, or experience of) agency, freedom, and value, etc., can be technologically reproduced.

Before embarking on the first chapter, a final preliminary remark. The scope of this thesis could clearly have been much wider. Since this thesis constitutes, to the author’s knowledge, the first attempt to uncover existential risk research as a research topic in its own right, there is no secondary literature it could have drawn on and used as orientation in its efforts to historically and conceptually embed existential risk research. It therefore is bound to miss out on interesting connections that could have been established from other perspectives, drawing on other schools of thought in political science and related disciplines. Without doubt, for instance existential risk research could have been approached from an International Relations perspective, integrating it into debates about global public bads and the becoming of a global political consciousness as a response to planetary threats, it could also have been approached from a security studies perspective,[9] a futures research perspective,[10] a risk studies perspective,[11] or a social theory perspective, most obviously, perhaps, by relating it to Ulrich Beck’s risk society concept.[12] However, given the limited scope of this thesis, it clearly would have been unfeasible to try and provide a comprehensive survey of interesting historical and conceptual connections between existential risk research and older debates in political theory and related disciplines.

The aim of this thesis is modestly to begin this conversation and to uncover existential risk research as a field of inquiry through the lens of one particular intellectual tradition. It holds that one of arguably many promising ways to approach this task is to conceive of existential risk research as

[9] For instructive discussions of the topic of catastrophe and emergency from a security studies or a democratic theory perspective see for instance Honig, B. (2011), Aradau, C. & Munster, R. (2011), Albertson, B. & Gadarian, S. (2015). All of these works would provide promising starting points to begin situating existential risk research from a political theory perspective.

[10] A particularly suitable starting point for situating existential risk research from this angle would be the recent work of Andersson, J. (2012, 2018) on the history of futures research. A similarly interesting touchstone could be found in the work of Amadae, S. M. (2016, 2018).

[11] See for instance Taleb, N. (2007), Sunstein, C. (2005, 2009), Dupuy, J. P. (2012).

[12] See e.g. Beck, U. (1992, 2007), Adam, B., et al. (2000).

part of an ongoing conversation about the problematic, multi-faceted, puzzling relationship between technology, human value, and human agency. It further holds that a good basis to bring out these deeper puzzles and to develop a better idea of what might be new and distinctive about existential risk research can be found in the post-war works of Martin Heidegger, Hannah Arendt and Günther Anders, even though, without doubt, many other authors could have been chosen as starting points. However, if the thesis succeeds in presenting existential risk research as a rich, new subject of intellectual inquiry from this particular perspective and manages to throw light on further puzzles along the way - puzzles which it does not itself address - it will already have achieved part of what it was intended to do.

1. Existential risk studies – an introduction to the field

Introduction

Since ca. 2010 the spectre of human extinction appears to have featured particularly prominently in the newspaper headlines of the Anglo-Saxon world.[13] *The Atlantic* told its readers that "We're underestimating the risk of human extinction";[14] the *BBC* asked "How are humans going to become extinct?"[15], *Science Magazine* diagnosed a "Denial of catastrophic risk";[16] *The Irish Times* asked "How long will the human species survive on Earth"?[17] and *The Express* cried "Humanity will go EXTINCT in 100 years".[18] Leaving aside the media's proclivity for sensationalist headlines, doomsday-heavy language of this kind does seem to reflect a form of collective existential fear that appears to have taken hold not only parts of the academic community, but also of the wider public.

Global warming is evolving from a distant, somewhat abstract phenomenon into a problem with immediate, real, and tangible consequences. Until recently, its effects were observed almost exclusively by experts, leaving their mark only in scientific models and in remote glacial regions of the planet. Now global warming is directly interfering with people's lives through rising sea levels, hurricanes, wild fires, and prolonged periods of drought.[19] Trends suggest worse is yet to come. As it stands, according to Nicholas Stern, "Bangladeshi farmers and Cairo city-dwellers are at severe risk of flooding and storms; southern Europe and parts of Africa and the Americas are threatened by desertification. Perhaps hundreds of millions of people may need to migrate as a result, posing an immense risk of conflict".[20]

And despite being the focus of our fears of climate change, its direct effects on human civilisation in the form of flooding, storms, draught and desertification, etc., are only the tip of the iceberg. The 2015 Pulitzer Prize for General Non-Fiction was awarded to the journalist Elizabeth Kolbert for her book on *The Sixth Extinction: An Unnatural History*, where she links global warming to sharply increasing losses in global biodiversity. Kolbert claims that we are entering a sixth mass extinction event - episodes in the Earth's history when the diversity of life plummets so sharply that, in a short geological interval, over three-quarters of all living species die out - a fear that is widely

[13] The focus of the thesis is the Anglo-Saxon world. To an extent this decision was made in order to narrow down the scope of the thesis. However, it also appears to be the case that the existential risk eco-system is largely limited to Great Britain and the Unites States. The focus on the Anglo-Saxon world therefore appears to be sufficiently wide to provide a representative portrayal of the existential risk research landscape.
[14] Andersen, R. (2012).
[15] Coughlan, S. (2013).
[16] Rees, M. (2013).
[17] Reville, M. (2016).
[18] Martin, S. (2017).
[19] IPCC (2014), pp. 4-7.
[20] Stern, N. (2016a).

shared within the scientific community.[21] This, Kolbert argues, constitutes not only a threat to the "living things with which we share Earth", it threatens to disrupt crucial "ecosystem services such as crop pollination and water purification" on which human survival depends.[22] "By disrupting these systems" Kolbert argues, "we're putting our own survival in danger".[23] The underlying concern is that humankind will not be able to isolate itself from the dynamics it unleashed on the world.

Contributing to this tense and fearful atmosphere, the world has recently experienced a comeback of fears of nuclear war. After years of relative tranquillity and stability following the end of the Cold War, nuclear programs in Iran and North Korea have raised fears of nuclear conflict to such an extent that the Bulletin of the Atomic Scientists in 2018 decided to shift its doomsday clock forwards to "two minutes to midnight", as close to midnight as it has been since 1953, the year in which the USSR succeeded in developing thermonuclear weapons.[24] According to the Bulletin, the doomsday clock should be seen as something akin to a barometer of apocalyptic fears, "a universally recognized indicator of the world's vulnerability to catastrophe from nuclear weapons, climate change, and new technologies emerging in other domains".[25] With midnight symbolising global catastrophe, the clock's minute handle conveys how close "we are to destroying our world". [26] According to the Bulletin's Science and Security board, humanity's situation in 2018 has not been this precarious since 1953: "world leaders failed to respond effectively to the looming threats of nuclear war and climate change, making the world security situation […] as dangerous as it has been since World War II".[27]

As if to encapsulate this gloomy zeitgeist, the past two decades have witnessed the steady rise in prominence of a new field of scientific inquiry: existential risk research, also sometimes referred to as existential risk scholarship or existential risk studies (in the following these appellations will be used interchangeably). The aim of this chapter is to scope this emerging field of inquiry, to survey its self-understanding, its methodology, and its aims, in order to arrive at a better understanding of what might be distinctive and perhaps new about it. Anticipating a little, what appears to be genuinely new about existential risk research is its attempt to open up the problem of human extinction as a field of scientific inquiry in its own right. Phil Torres, one of the field's most influential authors, for instance argues that existential risk scholarship "uses the tools and methods of rational empiricism to map out the obstacle course of risks that civilization must navigate in the coming centuries – and beyond".[28] That is, whilst stark warnings of human extinction or civilisational

[21] See for instance Ceballos, G. Ehrlich, P. et al. (2015), Barnodsky, A. et al. (2011), Wake, D. & Vredenburg, V. (2008).
[22] Kolbert, E. (2014), p. 472.
[23] Ibid.
[24] Bronson, R. (2018), p. 3.
[25] Ibid, p. 3.
[26] See Benedict, K. (2018).
[27] Bronson, R. (2018), p. 2.
[28] Torres, P. (2017b).

collapse such as those voiced by Elizabeth Kolbert, Paul Ehrlich, or the International Panel on Climate Change (IPCC), are pervasive and arguably have been for a long while, they typically are connected to and arise from observations of specific ecological, geological, technological, social, or geo-political trends. Existential risk research, however, inverts this perspective and takes a deductive approach. It makes the problem of human extinction its starting point for thinking about human affairs, present, and future and screens and analyses contemporary developments under this single aspect. Doing so, existential risk studies introduces a range of new concepts and methods and ultimately arrives at a distinctive set of ethical principles and policy recommendations. Rather than merely providing a descriptive analysis of the genre, however, this chapter seeks to establish to what extent this rather peculiar perspective on humanity and its future presents us with something qualitatively new. In laying the groundwork for the more detailed and interpretive analysis to follow, this chapter distils the underlying assumptions and implications that have remained – until now – implicit or underexplored.

1.1 The concept of existential risk

The concept of existential risk was formalised by Oxford philosopher Nick Bostrom in a 2002 paper entitled 'Existential risks and related hazards', which, in hindsight, can be regarded as the historic nucleus of 'existential risk research'.[29] In 'Existential risks and related hazards', Bostrom defines existential risks as "threats that could cause our extinction or destroy the potential of Earth-originating intelligent life".[30] Climate change and nuclear war, as we have seen, are often considered threats which could cause our extinction and are therefore typically included in the list of such risks. Bostrom goes further than that, however, and argues that we are presently living in the perhaps most critical phase of human history: "one might argue […] that the current century, or the next few centuries, will be a critical phase for humanity".[31] Increasing technological powers, he argues, are expected to multiply the potential sources of existential risk. Indeed, in addition to nuclear war and climate change, Bostrom argues humanity will be exposed to unprecedented existential risks from emerging technologies, such as nano-technology, synthetic biology, artificial intelligence, or geo-engineering: "Advances in biotechnology might make it possible to design new viruses that combine the easy contagion and mutability of the influenza virus with the lethality of HIV. Molecular nanotechnology might make it possible to create weapons systems with a destructive power dwarfing

[29] Viz. Torres, P. (2017b).

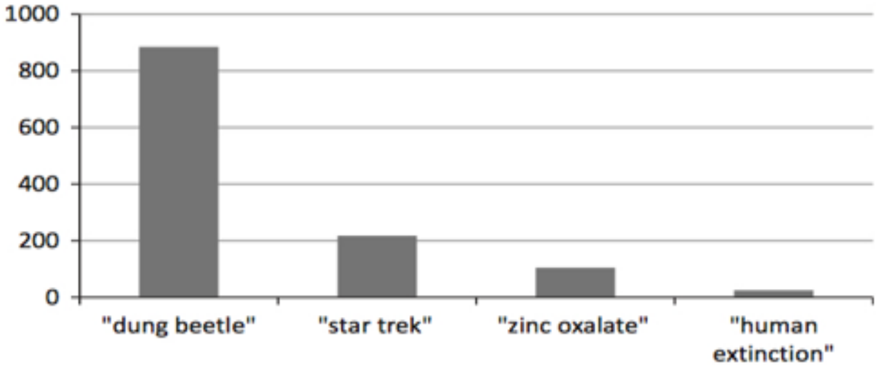
[30] Bostrom, N. (2002), p. 1.

[31] Bostrom, N. (2009), p. 211. This position is widely shared in the existential risk movement. Stephen Hawking, who was affiliated to both CSER and the FLI, to give just one other example at this point, for instance argued that the present century is likely the most dangerous in human history. See: Hawking, S. (2016).

that of both thermonuclear bombs and biowarfare agents. Superintelligent machines might be built and their actions could determine the future of humanity - and whether there will be one".[32]

Against this backdrop, one might be inclined to think of existential risk theory as merely another strand of literature seeking to raise attention about the "world's vulnerability to catastrophe," as it were, adding a perhaps superfluous voice to the amalgam of voices prophesying doom. But curiously, scholars in the field sometimes claim that they are rather isolated with their research interests, complaining about a perceived neglect of their concerns by the wider academic community. In a 2009 paper Bostrom argues that "existential risks have not received as much scholarly attention as they deserve. In recent years, there have been three serious books and one major paper on this topic".[33] In a 2013 piece, Bostrom again states that existential risk receives comparatively little scholarly attention relative to more profane topics: "In light of [the] very high value in studying existential risks and in analysing potential mitigation strategies, it is striking how little academic attention these issues have received, compared to other topics that are less important".[34] He then presents a figure in which the number of 'Scopus'[35] search results for the key-word 'human extinction' is compared to the number of search results for key words such as 'dung beetles', 'star trek' or 'zinc oxalate': [36]

Figure 1: Number of academic papers on various topics (listed in Scopus, August 2012)



Source: Bostrom (2013), p. 26, fig. 6

The figure suggests there is far greater academic interest in dung beetles than in human extinction. Bostrom takes this as a starting point to speculate about potential reasons for such ostensible neglect by the wider research community and argues that

[32] Bostrom, N. (2009), p. 198.
[33] Bostrom, N. (2009), p. 196.
[34] Bostrom. N. (2013), p. 26.
[35] Scopus is an abstract and citation database of peer reviewed journals, books and conference proceedings provided by Elsevier. See: https://www.elsevier.com/solutions/scopus.
[36] Bostrom (2013), p. 26, fig. 6.

> "many factors conspire against the study and mitigation of existential risks. Research is perhaps inhibited by the multidisciplinary nature of the problem, but also by deeper epistemological issues. The biggest existential risks are not amenable to plug-and-play scientific research methodologies. Furthermore, there are unresolved foundational issues, particularly concerning observation selection theory and population ethics, which are crucial to the assessment of existential risk; and these theoretical difficulties are compounded by psychological factors that make it difficult to think clearly about issues such as the end of humanity".[37]

Jason Matheny, another author in the field, similarly speculates about the perceived lack of scholarly research on human extinction, arguing it might be because "human extinction seems impossible, inevitable, or, in either case, beyond our control; maybe human extinction seems inconsequential compared to the other social issues to which cost-effectiveness analysis has been applied; or maybe the methodological and philosophical problems involved seem insuperable".[38] The contention that the problem of human extinction is not being taken seriously enough appears to be a common theme in the existential risk research community.[39]

My aim in this chapter is to develop a better idea of what authors such as Bostrom and Matheny might have in mind when they allege a lack of scholarly attention to the problem of human extinction. Quite clearly, if superficially, there appears to be no scarcity of collective existential fear; the spectre of human extinction is and has been on the mind of many scholars for decades. Countless numbers of books and journal articles have been written about global environmental crises, nuclear war, or pandemics and their potentially apocalyptic implications. What is it then that sets existential risk theory apart from long-standing anxieties about the fate of humanity? Are existential risk researchers justified in their belief that they are virtually alone in caring about the problem of human extinction, whilst the rest of the world is busy researching dung beetles?

My discussion of this question is bipartite. My central claim in this chapter is that existential risk researchers truly appear to be *approaching* the problem of human extinction in a new way, introducing a range of concepts and a new, integrative perspective to study the topic. Whilst their research certainly is emblematic of a zeitgeist that is characterised by a deep sense of the 'world's vulnerability to catastrophe', and as such can be conceived of as an organic outgrowth of the plethora of existential fears outlined above – they are neither alone in worrying about the possibility of human extinction, nor in discussing specific risks - existential risk theory sets out to provide a rigorous analytical framework for structuring this landscape of fears: it seeks to infuse it with conceptual clarity by studying risks to the survival of the human species as one 'integrated field'.[40] As Pamlin, Armstrong, et al. (2015) argue, existential risk research is "a scientific assessment about the

[37] Ibid, p. 26.

[38] Matheny, J. (2007), p. 1335.

[39] See also Rees, M. (2017a).

[40] Baum, S. (2015); Bostrom, N. & M. Ćirković (2008).

possibility of oblivion" *in general*.[41] It seeks to study end-time scenarios in an overarching, non-domain-specific way and to devise overarching strategies for risk minimisation based on such an integrative approach. This generates a genuinely new perspective on the future of humanity. When Bostrom suggests a neglect of the problem of human extinction, he thus does not mean that there is a disregard for the problem of the possibility of human extinction in specific contexts. Rather, he refers to a lack of rigorous, general, non-domain-specific attention to the problem. The claim that existential risk theory is isolated therefore is intelligible only in this highly specific sense.

The second part of my discussion shifts the focus from the descriptive to the interpretive, from the question what existential risk theory *is*, to the question of what the new conceptual toolkit it provides implies for our perspective on humanity. My main argument here is that existential risk theory, albeit ostensibly concerned with the study of threats to the survival of the human species, quickly evolves into a story about the relationship between technology and humanity. That in fact, the framing of the future of humanity in terms of survival means that technology emerges as humanity's destiny. This, I argue, is a direct consequence of the generalised perspective on human extinction scenarios it introduces. Albeit new in the conceptual sense described above, a more substantive reading of existential risk theory resonates with a rich history of thought in philosophy and political theory, echoing both high hopes and deeply rooted anxieties about modern science and technology and its implications for human life. The perspectives on that relationship, on how existential risk theory relates to older narratives, will be discussed in subsequent chapters.

In accordance with these two dimensions of my discussion, this chapter proceeds in two steps. In the first sections, a brief overview of the emerging genre of existential risk theory and the associated research movement is provided. They map out the evolving network of research institutes and scholars working in the field and are intended to carve out the theoretical scaffolding of existential risk theory – shared assumptions, shared narratives (normative and otherwise) and policy recommendations. The final section of the chapter then transitions to my discussion of existential risk theory's substantive implications, namely the role of technology in existential risk theory and demonstrating how technology emerges as the decisive determinant of human destiny.

1.2 Existential risk scholarship – basic ideas

Existential risk theory is predicated on the belief that human extinction is not receiving sufficient attention by the academic community. Per Bostrom: "existential risks have not received as much scholarly attention as they deserve".[42] According to many authors in the field, the problem begins with the fact that people rarely fully appreciate what human extinction would actually entail, i.e. that people tend not to be fully aware of the stakes involved. A common starting point in the literature on

[41] Pamlin, D., Armstrong, S. et al. (2015), p. 6.

[42] Bostrom, N. (2009), p. 196, See also Pamlin, D., Armstrong, S. et al. (2015).

existential risk is a section of Derek Parfit's seminal work *Reasons and Persons*.[43] In a passage about the future prospects of moral and ethical progress, he asks the reader to compare three outcomes:

> "1. Peace. 2. A nuclear war that kills 99 per cent of the world's existing population. 3. A nuclear war that kills 100 per cent. (2) would be worse than (1), and (3) would be worse than (2). Which is the greater of these two differences? Most people believe that the greater difference is between (1) and (2). I believe that the difference between (2) and (3) is very much greater … The Earth will remain habitable for at least another billion years. Civilization began only a few thousand years ago. If we do not destroy mankind, these few thousand years may be only a tiny fraction of the whole of civilized human history. The difference between (2) and (3) may thus be the difference between this tiny fraction and all of the rest of this history".[44]

This view, Parfit argues, must be true for both classical utilitarians as well as for those who attribute value to 'ideal goods' such as the arts and sciences, or moral and ethical progress. Carl Sagan, who is also often invoked by existential risk theorists, made effectively the same point in a 1983 article in Foreign Affairs:

> "Some have argued that the difference between the deaths of several hundred million people in a nuclear war (as has been thought until recently to be a reasonable upper limit) and the death of every person on Earth (as now seems possible) is only a matter of one order of magnitude. For me, the difference is considerably greater. Restricting our attention only to those who die as a consequence of the war conceals its full impact. If we are required to calibrate extinction in numerical terms, I would be sure to include the number of people in future generations who would not be born. A nuclear war imperils all of our descendants, for as long as there will be humans. Even if the population remains static, with an average lifetime of the order of 100 years, over a typical time period for the biological evolution of a successful species (roughly ten million years), we are talking about some 500 trillion people yet to come. By this criterion, the stakes are one million times greater for extinction than for the more modest nuclear wars that kill "only" hundreds of millions of people. There are many other possible measures of the potential loss—including culture and science, the evolutionary history of the planet, and the significance of the lives of all of our ancestors who contributed to the future of their descendants. Extinction is the undoing of the human enterprise".[45]

These two passages contain much of what must be known about existential risk theory. The starting point of existential risk theory is essentially an ethical one, akin to the dictum 'quantity begets a quality of its own'. It is about making explicit and raising attention for the categorically different level of horror human extinction would amount to as compared to all large-scale catastrophes humanity has endured throughout its history, from pandemics, to world wars, to natural disasters.

The basic conviction of existential risk theory is that, once reflected upon 'soberly',[46] human extinction is even worse than our first emotional reaction to the idea might suggest. As the philosopher Peter Singer puts it in a recent article, co-authored by two philosophers from Bostrom's Future of Humanity Institute: "One very bad thing about human extinction would be that billions of

[43] See for instance Bostrom, N. (2013), Matheny, J. (2007), Farquhar, S. et al. (2017).
[44] Parfit, D. (1984), p. 453.
[45] Sagan, C. (1983), p. 275.
[46] Bostrom, N. (2002), p. 4.

people would likely die painful deaths. But in our view, this is, by far, not the worst thing about human extinction. The worst thing about human extinction is that there would be no future generations".[47] Or, as another group of authors has it "what makes existential catastrophes especially bad is that they would "destroy the future".[48] The basic position is that, once we take heed of the fact that human extinction would imply a virtually infinitely negative impact, all other types of concerns, even catastrophes on the scale of the Spanish flu, fade in comparison.[49]

Now, this might at first seem rather trivial, until one remembers that existential risk theory can be seen as the attempt to infuse the amalgam of anxieties permeating society and parts of the academic community with analytical rigour. To do so, the first step is to define what it is one is afraid of, to be clear about one's anxieties, as it were. What existential risk researchers want to raise awareness for is that human extinction is a wholly different category of horror which should not uncritically be lumped together with other kinds of large scale catastrophes: "Tragic as such events are to the people immediately affected", Bostrom argues, "in the big picture of things – from the perspective of humankind as a whole – even the worst of these catastrophes are mere ripples on the surface of the great sea of life.[50] They haven't significantly affected the total amount of human suffering or happiness or determined the long-term fate of our species".[51] Human extinction, on the other hand, is terminal and precludes recovery. From the utilitarian perspective espoused by Bostrom, given the virtually infinite stakes involved,[52] the prevention of human extinction should be morally paramount.[53]

Taking human extinction seriously, making it the focus of one's attention in its own right means taking *all* scenarios that could possibly result in human extinction into account. Bostrom acknowledges that the problem of global catastrophes, of civilisational collapse and, as an upper limit

[47] Singer, P., et al. (2013).

[48] Farquhar, S. et al. (2017), see also Baum, S. & Barrett, A. (2017).

[49] Most existential risk researchers take a pan-generational utilitarian perspective which is contested on a variety of grounds. The basic claim that human extinction would be an extraordinarily bad outcome, however, seems uncontroversial, even if one does not take a pan-generational utilitarian perspective. Several authors have shown that it would be worse than any other outcome even when considered from a non-utilitarian perspective and if one abstracts entirely from effects on future generations in one's normative calculus [Scheffler, S. (2016), Dasgupta, P. (2017)].

[50] In his original paper, Bostrom uses three dimensions to describe the magnitude of a risk: scope, intensity and probability. Scope refers to the size of the group of people that are at risk, intensity to the severity of the impact expected to affect each individual in the respective group and probability denotes the best current subjective estimate of the probability of the adverse outcome. In terms of scope, Bostrom distinguishes between personal, local, and global risks. In terms of intensity, he distinguishes between endurable and terminal risks. 'Endurable' means that the inflicted damage may cause great destruction in the short to medium term but, ultimately, is recoverable. 'Terminal' means that the inflicted damage is so intense that it is effectively unrecoverable: "the targets are either annihilated or irreversibly crippled in ways that radically reduce their potential to live the sort of life they aspire to". Based on this initial categorisation, Bostrom identifies six qualitatively distinct types of risk, comprised of the six possible combinations of types of scope and intensity (probability is superimposed). An existential risk in this scheme is one that is terminal in intensity and global in scope. See Bostrom, N. (2002), p. 4.

[51] Bostrom, N. (2002), p. 2.

[52] Pamlin, D. & Armstrong, S. (2015), p. 31.

[53] Bostrom, N. (2013).

of such catastrophes, the problem of human extinction, has been on the mind of many authors writing throughout the 20th century. He references for instance Rachel Carson's *Silent Spring*, Paul Ehrlich's *Population Bomb,* and the Club of Rome's *Limits to Growth* in the context of concerns regarding the potentially apocalyptic consequences of environmental degradation.[54] However, in all of these cases the risk of human extinction is invoked only peripherally as a potential worst-case scenario and, even where it is discussed in a more focused manner, only within a specific context. The same could be said about Derek Parfit and Carl Sagan, as well as for the many other authors, who discussed the possibility of human extinction in the context of nuclear weapons throughout the 20th century – from Bertrand Russell and Albert Einstein,[55] to Jonathan Schell and Jeff McMahan, to name but a few.[56] According to Bostrom and Cirkovic (2008) this is no longer sufficient: "If we treat risks singly, and never as part of an overall threat profile, we may become unduly fixated on the one or two dangers that happen to have captured the public or expert imagination of the day, while neglecting other risks that are more severe or more amenable to mitigation".[57] Taking human extinction seriously, according to these authors, thus logically requires one "to take on board more generalised concerns" about human extinction and to develop an 'overall threat profile'. This notion is the defining and distinctive feature of existential risk theory.[58]

The first definition of existential risk can be found in Bostrom's original piece on the topic 'Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards'.[59] Bostrom here defines existential risks as "threats that could cause our extinction or destroy the potential of Earth-originating intelligent life".[60] The definition, in particular its second component referring to Earth-originating intelligent life, will be a matter of discussion at a later point - defining existential risk is not as straightforward as it might seem. For the moment, however, suffice it to say that the concept of existential risk may be understood narrowly as a category that allows for the unified discussion of '*all threats that could cause our extinction*' (**Definition 1**), or, in Sagan's words, of all threats that could mean "the undoing of the human enterprise". It is also in this narrow sense that the concept is usually employed throughout the literature. When Bostrom argued in 2009 that "only three serious books and one paper" had been written about the topic of human extinction, he had in mind not any publication where the problem of human extinction is invoked or serves as a backdrop for reflections about specific environmental, technological, or political developments, but the kind of systematic,

[54] Bostrom, N. (2009), p. 198.
[55] Russell, B. & Einstein, A. (1955).
[56] See Schell, J. (1982), McMahan, J. (1986).
[57] Bostrom, N. & Cirkovic, M. (2008), p. 2.
[58] Bostrom, N. (2013), p. 27.
[59] In a search for pre-2002 mentions of "existential risk" on Google Scholar results showed that the concept was almost exclusively employed to refer to life-threatening on an individual. Searches on Scopus yielded similar results.
[60] Bostrom, N. (2002), p. 1.

non-domain-specific reflection about human extinction in general, which the concept of existential risk is intended to allow for.[61]

1.3 The existential risk eco-system

Over the past decade the number of research institutes and individual scholars that chose to make existential risk their focus has grown rapidly. The emerging movement, or eco-system,[62] had considerable success in raising attention for its concerns amongst the general public as well as in the upper echelons of the corporate world, academia, civil society and politics. In this section, a brief overview of the institutional structure of the existential risk movement is provided to convey an idea of the movement's shape and reach.

In 2005 Bostrom founded the Future of Humanity Institute (FHI) at Oxford University as part of the Faculty of Philosophy and the Oxford Martin School. The FHI's mission reads as follows: "using the tools of mathematics, philosophy, and science, we explore the risks and opportunities that will arise from technological change, weigh ethical dilemmas, and evaluate global priorities. Our goal is to clarify the choices that will shape humanity's long-term future ".[63] To this day, the FHI remains arguably the most influential research institute in the field and a number of its researchers, such as Anders Sandberg, Toby Ord, Robin Hanson and Stuart Armstrong, will be discussed here. Since the FHI's foundation many other institutes with a similar focus have been founded across Europe and the US.

In 2011 the Global Catastrophic Risk Institute (GCRI) was founded by Seth Baum and Tony Barrett, two other influential scholars in the field. The GCRI's focus is not exclusively on existential risks but on global large-scale risks more generally. However, as will be discussed in greater depth later, this distinction is not entirely straightforward and much of the GCRI's research can therefore be considered as existential risk research.[64] The mission of the Global Challenges Foundation, founded in Sweden in 2012, is "to incite deeper understanding of the global risks that threaten humanity and catalyse ideas to tackle them. Rooted in a scientific analysis of risk, the Foundation brings together the brightest minds from academia, politics, business and civil society to forge transformative approaches to secure a better future for all".[65] The Centre for the Study of Existential Risk, founded in 2013 at the University of Cambridge by Astronomer Royal Martin Rees,

[61] Bostrom, N. & Cirkovic, M. (2008), p. 2.

[62] Torres, P. (2017b), p. 2.

[63] Please compare to FHI (2018a).

[64] The GCR's mission statement reads as follows: "The Global Catastrophic Risk Institute (GCRI) is a nonprofit, nonpartisan think tank. GCRI was founded in 2011 by Seth Baum and Tony Barrett. GCRI studies the full range of GCRs and GCR topics in order to answer the big questions: Which risks should society be most worried about? How do the different risks affect each other? And above all, what are the best ways to reduce the risk?" Please see: GCRI (2018).

[65] See Baum, S. et al. (2016), p. 3.

philosopher Huw Price and Skype founder Jaan Tallinn, describes itself as "a multidisciplinary research centre dedicated to the study and mitigation of risks that could lead to human extinction. Our goal is to steer a small fraction of Cambridge's great intellectual resources […] to the task of ensuring that our own species has a long-term future".[66] The Future of Life Institute, which was founded in 2014 by MIT physicist Max Tegmark as well as Jaan Tallinn, considers itself a "research and outreach organization working to mitigate existential risks facing humanity" with the mission "to catalyse and support research and initiatives for safeguarding life and developing optimistic visions of the future, including positive ways for humanity to steer its own course considering new technologies and challenges".[67]

Several more recently formed institutes can also be identified, with either an explicit focus on existential risk or at least list existential risk research among their top research priorities. The Institute for Ethics and Emerging Technologies (IEET), which was founded by Nick Bostrom in 2004;[68] the Foundational Research Institute (FRI), founded in 2013; [69] The Project for Future of Human Flourishing (FHF), formerly the Existential Risk Institute, founded in 2017;[70] The Global Priorities Project (GPP), founded in 2014;[71] The Berkeley Existential Risk Initiative, founded in 2017;[72] and the Global Priorities Institute (GPI), a research centre within the University of Oxford's faculty of philosophy, founded in 2018.[73] Several additional institutes have a focus on specific existential risks, in most cases those associated with artificial intelligence, and have close personal and institutional ties to one or more of the above named institutes. The Machine Intelligence Research Institute at the University of Berkeley, or Open AI, an independent research company funded by a range of Silicon Valley entrepreneurs and with close ties to the FLF, is one such example. Both of these institutes have a focus on developing 'benevolent' artificial general intelligence, a topic I will return to in chapter 4.[74] I abstract from these institutes for the moment, however, as my focus here is on general existential risk research.

What stands out, is that the movement was able to gather the intellectual and financial support of many highly prominent public figures with global reach, in particular from academia. Cambridge physicists Martin Rees and the late Stephen Hawking, were both on the board of CSER and the FLI; MIT physicists Max Tegmark, Alan Guth and Nobel laureate Frank Wilczek, are on the

[66] Please compare to the Centre for the Study of Existential Risk's webpage, see CSER (2018a).
[67] Please compare to the Future of Life's webpage, see FLI (2018d).
[68] See IEET (2018).
[69] See FRI (2018).
[70] See FHF (2018), the FHF's mission statement reads as follows: "The Project for FHF aims to bring into conversation a wide range of research on issues relating not just to our near-term prospects, but to humanity's long-term future in the universe. We aim to understand the formidable existential threats before us and to devise effective means for mitigating these threats."
[71] The GPP is a joint initiative of the FHI and the Centre for Effective Altruism. See GPP (2018).
[72] See BERI (2018).
[73] See GPI (2018)
[74] Please see respectively: MIRI (2018a, 2018b) and OpenAI (2015).

scientific advisory board of the FLI;[75] and high-profile Silicon Valley entrepreneurs such as Elon Musk, Jaan Tallinn, Peter Thiel, or Sam Altmann, financially support several of these institutes, such as CSER, the FLI and Open AI.

1.4 Academic disciplines in existential risk research

Organisationally, the existential risk movement is expanding rapidly and the above list of institutes may well continue to expand past the time of writing. The main purpose of this overview was to demonstrate that the existential risk eco-system is expanding rapidly, particularly between 2010 and 2018, and to introduce some of the key research institutes and scholars in the field.

The movement has not only expanded organisationally, but also on a disciplinary level. The origins of the field are, if we take Nick Bostrom as its figurehead, in transhumanist philosophy. Bostrom is one of the founders of the 'World Transhumanist Association', which was founded in 1998 and is now called 'Humanity+', and one of the authors of the 'Transhumanist Declaration'.[76] Many authors currently contributing to existential risk research have a background in transhumanist philosophy, particularly at Oxford's FHI, for instance Stuart Armstrong, Robin Hanson, and Anders Sandberg, or Phil Torres of the FHF. The fact that existential risk research has its roots in transhumanism is interesting and to an extent revealing in its own right (we return to this later). The concept of existential risk has also attracted the attention of an increasing number of prominent scholars from a wide array of more traditional academic backgrounds, from legal scholars such as Richard Posner,[77] Cass Sunstein,[78] or Jonathan Wiener,[79] to economists such as Partha Dasgupta,[80] mathematicians such as Olle Häggström,[81] security scholars such as Frances Flannery or Gary Ackermann,[82] to philosophers of mind, such as Susan Schneider, ethicists such as Oxford professors Hilary Greaves and Toby Ord,[83] biologists, such as Harvard professor of genetics George Church, political scientists, such as Yale professor Allan Dafoe,[84] and several experts in the field of artificial intelligence such as Berkeley professor Stuart Russell, Francesca Rossi of IBM, or Viktoriya Krakovna of Google DeepMind,[85] to name just a few of the better-known personalities.

[75] Please see FLI (2018d).

[76] Please see Humanity+ (2018).

[77] Posner, R. (2006).

[78] Sunstein, C. (2009).

[79] Wiener, J. (2016).

[80] Partha Dasgupta is on the Management committee of Cambridge University's Centre for the Study of Existential Risk. Please see CSER (2018c).

[81] Olle Häggström is a mathematics professor at Chalmers University in Sweden and sits on the board of the recently founded FHF. See FHF (2018b). He also recently published a book on existential risk, see Häggström, O. (2016).

[82] Both are on the board of the FHF. See FHF (2018b).

[83] Hilary Greaves is a professor of philosophy at Oxford University and founder of the GPI.

[84] Allan Dafoe is Director of the FHI's 'Governance of AI Program'. Please FHI (2018b).

[85] George Church, Stuart Russell and Viktoriya Krakovna are on the board of the FLI, see FLI (2018d).

Existential risk research is a rapidly growing and increasingly interdisciplinary field of study, which makes it difficult to provide a comprehensive and exhaustive account of its web of institutes and the associated literature. I therefore focus on some of the most central publications, such as those of Nick Bostrom or Martin Rees, and complement them, where relevant, with publications, reports, and other types of publications that were issued either by the above-named institutes or authors affiliated to them. This textual analysis is informed by insights gathered by means of participant observation. Over a period of 4 years, between October 2014 and October 2018, I immersed myself deeply in the UK existential risk community, regularly attending talks, seminars, workshops, and conferences organised by CSER at Cambridge University, by the FHI at Oxford University, and by the UK Parliament's All Party Parliamentary Group (APPG) for Future Generations at the House of Commons.

I attended for instance the Cambridge Conference on Catastrophic Risk in 2016 and 2018, the CFI Conference 2017, many of CSER's lectures, seminars and workshops, such as David Denkenberger's lecture on 'Feeding Everyone no Matter What', or Toby Ord's lecture on 'Will We Cause our Own Extinction?', or the APPG's session on 'How do We Make AI Safe for Future Generations?', to give but a few examples. Furthermore, I joined the Cambridge University's 'The Future of Sentience' Society (a recently founded student society which serves as a link between CSER and Cambridge University's student body) for several of its meetings.

Immersing myself in the community in such a way helped me to develop an overview of the movement, to gain a better understanding of its priorities, its key figures and its core convictions, and to orientate and anchor my textual analysis. What unites the outlined movement, are two core convictions: 1) that we, as the late Stephen Hawking noted, live in perhaps the most dangerous and most decisive period in human history *ever* because the contemporary world faces a growing number of unprecedented existential risks on ever more frontiers.[86] 2) That the chief contributing factor is humanity's rapidly expanding technological powers. Martin Rees sums these two convictions up in one remark: "extending far into the future as well as into the past - the twenty-first century may be a defining moment. It is the first in our planet's history where one species - ours - has Earth' s future in its hands and could jeopardise not only itself but also life's immense potential".[87] These two contentions translate directly into the directive to make existential risk research a global priority.

1.5 The 'science of existential risk'

Two levels of existential risk research can be distinguished. The first deals with existential risks in general, treating them as one analytical category. As such it is concerned with identifying, conceptualising and analysing common features of existential risks and to investigate the ethical and

[86] See for instance Rees, M. (2004), Torres, P. (2016), Häggström, O. (2017).
[87] Rees, M. (2008), p. xi.

political implications of the entire phenomenon without going into the detail of specific risks. This type of research is intended to provide conceptual and methodological clarity, define categories (of different existential risks for instance), think through the ethical status of the far future, and to derive policy recommendations. The FHI refers to this type of existential risk research as "macro-strategic" research, CSER calls it simply the "science of existential risk".[88] The second type of research is domain-specific. Research conducted on this level seeks to identify specific sources of existential risk, analyse their main drivers and discuss potential regulatory or technological solutions to them. Oftentimes research lies somewhere between these two poles and most of the research institutes introduced above work on both frontiers. My interest in the first three chapters lies specifically on the macro-strategic level of existential risk research because it is this dimension that appears to genuinely bring new facets to bear. Macro-strategic existential risk research presents us with is a quite distinctive perspective on the future of humanity. By making potential end-time scenarios its reference point for reflecting about human affairs, present and future, it manages to assume a singularly detached and ostensibly objective perspective on the future of humanity. In a curious way the focus on the *end* of humanity opens the future up as a field of rigorous inquiry whereby the future of humanity emerges as an open-ended "obstacle-course",[89] or a "mine-field", which humanity must navigate.[90] However, at some points in my analysis I will refer to debates on the applied level, as illustrations or clarifications. The discussion on artificial intelligence (AI), for instance, which I expand in chapter four, is particularly interesting because it lies between the two poles of existential risk research. On the one hand, AI is a single technology and therefore debates surrounding it can be considered domain-specific. On the other hand, AI cannot be lumped together with other technologies (if it can be called a technology to begin with). As we shall see, the problem of AI touches upon the very question of what it means to be human and therefore the question of AI overlaps in multiple ways with wider macro-strategic questions about the future of humanity in general.

Macro-strategic existential risk research

With respect to macro-strategic existential risk research the FHI is by far the most active institute. Bostrom, Ord, Armstrong and Sandberg regularly publish on the topic, and their work forms the bulk of my literature review.[91]

[88] Please see FHI (2018c) and CSER (2018a).

[89] Torres, P. (2017b), p. 2.

[90] Häggström, O. (2016), p. 6.

[91] CSER has taken up work only relatively recently and thus has not yet produced many academic publications. Martin Rees's book *Our Final Hour* [Rees, M. (2004)] can be seen as an early contribution to macro-strategic existential risk research. The FLI in the long run wants to contribute to a better understanding of existential risk in general but currently focuses on risks associated with artificial intelligence and therefore focuses on research on the applied level, see FLI (2018d). Others publishing on topics in the field are the researchers linked to the

The most basic question underpinning macro-strategic existential risk research is how high the overall level of existential risk facing humanity at any given point might actually be. Drawing on John Leslie's *The End of The World: The Science and Ethics of Human Extinction,* Bostrom distinguishes between direct and indirect methods of approximating the total level of existential risk. Direct estimates "analyse the various specific failure-modes, assign them probabilities, and then subtract the sum of these disaster-probabilities".[92] Most direct existential risk estimates start by distinguishing between *natural* existential risks, such as the risks associated with volcano eruptions, gamma-ray bursts, or asteroid impacts, and *anthropogenic* existential risks, i.e. risks that have their origin in human activity, for instance risks stemming from nuclear technology, or human interferences with the environment, such as anthropogenic climate change, which has its roots in carbon dioxide emissions from fossil fuel combustion.[93] There is a wide consensus that natural existential risks constitute only a small share of overall existential risk since geological data suggests that natural events which could conceivably pose an existential threat to the human species are extremely rare occurrences.[94] Existential risk researchers argue that, given that humanity (defined as the species of *homo sapiens*) has existed for only 200,000 years, whereas the median duration of mammalian species is about 2.2 million years, the probability of natural extinction events occurring is regarded as extremely low for any given century.[95] In other words, it is with relatively high confidence that one can assume that we and the generations of humans who will succeed us will likely not be exposed to a civilisation threatening asteroid impact or a super volcano eruption, provided that the distribution of natural disasters does not change.[96]

Existential risk researchers therefore argue that the overwhelming majority of existential risks we are facing in this century and beyond are likely to be man-made. These self-inflicted threats, however, as Rees (2013) argues, are a very new phenomenon and therefore we have virtually no data based on which we could confidently assume they are equally unlikely to materialise as natural existential disasters.[97] On the contrary, the limited experience mankind has made with anthropogenic existential risks is grounds alone for serious concern. In the 20th century the probability of an all-out nuclear war (arguably the first anthropogenic existential risk), was at times very high. People closest to the situation, such US President John F. Kennedy or John von Neumann, believed nuclear war to

Global Catastrophic Risk Institute, specifically Seth Baum. Apart from that there are also a couple of researchers without any direct affiliation to one of the above institutes, most notably Cirkovic, Jason Matheny, Richard Posner and Cass Sunstein, who regularly contribute to research in the field and will be referenced repeatedly throughout the thesis.

[92] Bostrom, N. (2002), p. 15.

[93] See for instance Bostrom, N. (2002), p.20; Bostrom, N. (2013), p.15; or Rees, M. (2013).

[94] See Ibid, as well as for instance Smil, V. (2008), Matheny, J. (2007); Farquhar, S. et al. (2017); Todd, B. (2017), or Rees, M. (2013, 2014).

[95] Compare this argument to Beckstead, N. & Ord, T. (2014). p. 116; Similar arguments can be found in Matheny, J. (2007), p. 1336, and Rees, M. (2014).

[96] Rees, M. (2014).

[97] Rees, M. (2013), p. 1223; see also Bostrom, N. (2013), p. 15-16.

be extremely likely if not inevitable.[98] Albeit being aware of the necessarily speculative nature of any *precise* figure, Martin Rees claims retrospectively that the "annual risk of thermonuclear destruction during the Cold War was about 10,000 times higher than from asteroid impact".[99] Many existential risk researchers share the opinion that this century must be expected to be no less dangerous than the last. If anything, they predict that humanity is going to face a variety of man-made challenges in the coming decades that might well expose it to even greater levels of existential risk than the Cold War did. Some of these risks, such as the ones associated with runaway climate change, are already quite well known. However, existential risk researchers are concerned that developments in emerging technologies might pose even greater threats. Technological existential risks are in fact expected to account for the 'great bulk' of existential risk humanity will be facing throughout the coming decades and centuries.[100] For this reason most of the existential risk institutes currently focus on studying the risks associated with emerging technologies, rather than other anthropogenic existential risks, such as climate change. The technological focus might be driven by the fact these researchers regard technological risks as understudied, whereas climate change and ecological collapse are already widely discussed and thoroughly analysed 'mainstream' issues.[101]

The emerging technologies that are most frequently named as potential future sources of existential risks are synthetic biotechnology, nanotechnology, artificial intelligence, and geo-engineering.[102] Synthetic biotechnology is expected to make it possible to overcome natural limits on virulence and transmissibility, which then might make it relatively easy to engineer pathogens of extreme lethality, that might cause pandemics of unprecedented scale and severity and pose an extinction risk to humanity.[103] Many authors see biotechnology as a particularly problematic case because the barriers to make use of it are relatively low: "knowledge and equipment needed to engineer viruses is modest in comparison with what is required to create a nuclear weapon…a single undetected terrorist group would be able to develop and deploy engineered pathogens".[104]

The existential risks posed by nanotechnology, defined as atomically precise manufacturing, on the other hand are still largely hypothetical. One scenario that is frequently named in the literature is that experiments in nanotechnology could accidentally lead to the emergence of self-replicating nano-machines, which, in a runaway replication process, could end up consuming the entire

[98] "President Kennedy is said to have at one point estimated the probability of a nuclear war between the US and the USSR to be *"somewhere between one out of three and even*". John von Neumann (1903-1957), who as chairman of the Air Force Strategic Missiles Evaluation Committee was a key architect of early US nuclear strategy, is reported to have said it was *"absolutely certain (1) that there would be a nuclear war; and (2) that everyone would die in it",* as quoted in Bostrom, N. (2002), p. 3, fn. 4.

[99] Rees, M. (2014).

[100] Bostrom, N. (2013), p.16; see also e.g. Beckstead, N. & Ord, T. (2014); p. 118; Rees, M. (2013), p.1223; Baum, S., Farquhar, S., et al. (2016).

[101] See particularly Rees, M. (2014).

[102] See for example Beckstead, N. et al. (2014); Bostrom, N. (2002, 2013); Pamlin, D. & Armstrong, S. (2015); Beckstead, N. & Ord, T. (2014), p. 116.

[103] See for instance Beckstead, N. & Ord, T. (2014), p. 118.

[104] Beckstead, N. & Ord, T. (2014), p. 118.

biosphere and turn it into 'grey goo'.[105] Another, and perhaps more conventional, source of risk associated with nano-technology is comparable to those associated with synthetic biotechnology. Namely that, once nanotechnology has matured and can be used for 'distributed manufacturing', it could give an increasing number of people the means to produce their own arsenals of highly powerful, sophisticated weaponry. Some authors are concerned that it might facilitate the production of nuclear bombs or chemical weapons to unsustainable levels.[106]

Concerns relating to developments in artificial intelligence, the science of engineering intelligent machines, usually focus on the hypothetical moment when so called superintelligence is developed. Very broadly put, superintelligence is defined as an artificial intelligence that supersedes human intelligence in all domains, i.e. in every instrumentally relevant respect. It is envisioned to be better not only at solving specific, well-defined, problems (such as playing chess, driving a car, or diagnosing cancer), but at the very capacity to autonomously identify and specify problems as well as to devise and implement strategies to solve them. For our immediate purpose, let us consider superintelligence simply as superhuman instrumental rationality. The fears surrounding superintelligence, concern the precise *instance* when such a superintelligent agent is first switched on or emerges. The fear is that once such a superhuman instrumental rationality is unleashed onto the world it can no longer be contained, precisely because it exceeds human intelligence and thus must be expected to be able to behave and strategise in ways that humanity cannot predict and prevent. Humanity would then find itself at the mercy of that superhuman will. The existential risk is generally associated with the possibility that the superintelligent agent might pursue goals detrimental to humanity's interests. According to the literature, this could be happening for a variety of reasons and in a variety of forms, either because the AI starts pursuing its own, unforeseen ends, or simply because it was equipped with a set of poorly defined ends and value functions to begin with. [107] Some authors argue that an existential catastrophe might be the *default* outcome of the development of superintelligence, and therefore consider the existential risk associated with the rapid progress we are presently witnessing in the field of AI to be high.[108] The account of the existential risk associated with AI will be refined in chapter 4, but the above discussion should suffice as a working basis.

Finally, geoengineering, typically defined as the deliberate use of technology to alter the world's climate, might pose existential risks because its application could have severe unintended consequences such as draughts, ozone depletion, or acid rain which could, according to some authors, turn out to be so severe that the Earth becomes uninhabitable for humans.[109]

[105] See for instance Phoenix, C. & Drexler, E. (2004); Rees, M. (2004), p.132.

[106] Pamlin, D. & Armstrong, S. (2015), p. 115.

[107] See for instance Müller, V. (2014), p. 297, or Price, H. & Tallinn, J. (2012).

[108] Soares, N. & Fallenstein, B. (2014, 2017).

[109] Beckstead, N. et al. (2014), pp. 6-7, See also Bostrom, N. et al. (2013), or Price, H. & Ó hÉigeartaigh, S. (2014).

Existential risk researchers are of course aware that the attempt to assign quantifiable probabilities to any single of these scenarios is a highly controversial enterprise, to say the least. Predictions about future risks associated with potential future technological capabilities hinge to a large extent on the informed but subjective judgment of individual experts. As opposed to natural existential risks, the assessment of which draws on a wealth of geological and astrological data, direct estimates of the total level of anthropogenic risk humanity, i.e. the attempt to analyse "individual failure-modes, assign them probabilities, and then subtract the sum of these disaster-probabilities",[110] thus faces severe epistemological limitations, a situation further compounded by the multidisciplinary nature of the project.[111]

This is where indirect risk estimators come into play. Indirect existential risk estimates typically rely on thought experiments and abstract, probabilistic theorising as opposed to predictions about specific future technological developments. They rely on reference points external to humanity itself to logically constrain what can be coherently believed about the potential duration of the future of humanity.[112] It is beyond the scope of this thesis to summarise these debates in detail and I will therefore limit myself to one example. The most frequently discussed indirect estimator in the literature is the so called Fermi Paradox.[113] On the one hand, we have reason to assume that there is a non-negligible probability for life to have emerged somewhere else in the universe (given that the universe has existed for billions of years and it is believed that there are millions of planets with broadly earth-like conditions). On the other, hypothetically, the universe could be colonised with relative ease once a civilisation has passed a certain technological threshold.[114] However, there are no signs of it.[115] There are several interesting explanations for this perceived paradox, including one which concludes that we are more likely to live in a simulation than not.[116] The one most widely discussed in the context of existential risk, however, is that *intelligent* life could face a 'Great Filter' at some point in its evolution which prevents it from spacefaring. It may be either, that it is hard for intelligent life to emerge in the first place, in which case the Great Filter could lie in our past, or it could be the case that almost every intelligent species destroys itself once it has reached a certain level of technological development, in which case the Great Filter could be ahead of us.[117] There is a long-standing debate about what to make of the Fermi Paradox,[118] whether it is logically valid to begin with and, if it is valid, whether it can tell us anything about the likelihood of human extinction

[110] Bostrom, N. (2002), p. 15.

[111] See Bostrom, N. (2002), p. 16; Rees, M. (2014).

[112] Rees, M. (2014).

[113] See for instance Baum, S. (2010), pp. 594-597; Matheny, J. (2007), pp. 1336 – 1337; Armstrong, S. & Sandberg, A. (2013); Bostrom, N. (2002), p. 16; Rees, M. (2004).

[114] Much of this debate goes back to Brin (1983).

[115] See Ibid, Armstrong, S. & Sandberg, A. (2013), p. 1.

[116] Bostrom, N. (2003).

[117] Armstrong, S. & Sandberg, A. (2013), p.1; Bostrom, N. (2002), p. 16; Scharf, C. (2016).

[118] For instance, Freitas, R. (1985) argues that it is a logical fallacy, whereas Baum, S. (2010) holds the opposite.

events.[119] In addition to the Fermi Paradox, there are other indirect estimators of human extinction risk, such as observation selection effects, which are central to John Leslie's Doomsday Argument,[120] as well as indirect estimators that try to correct for psychological biases in the way people (including researchers) judge the likelihood of large scale disasters and draw inferences about the plausibility of direct existential risk estimates.[121]

Based on direct and indirect estimates, several of the leading voices in existential risk theory arrive at rather gloomy predictions about the chances of humanity surviving the decades and centuries to come. Martin Rees believes that there is a 50 percent chance that humanity will not survive the 21st century,[122] Nick Bostrom believes that "setting this probability lower than 25% would be misguided" and "that the best estimate may be considerably higher",[123] the 'Stern Review' estimated the risk of extinction at 10 per cent for this century,[124] and philosopher John Leslie at 30 per cent.[125] In an informal survey, conducted with participants of the 2008 Global Catastrophic Risk Conference, the median respondent assigned a 19 per cent probability for humanity going extinct in this century.[126]

Of course, these estimates should be taken with a pinch of salt. As indicated above, they rely entirely on the informed but ultimately speculative guesses of individual experts about future technological capabilities, complemented by their own indirect risk estimators. However, these limitations are effectively inconsequential for existential risk researchers. Given the stakes involved, what matters for existential risk research is solely that existential catastrophe is a non-negligible *possibility*, even if reliable judgements about its probability are impossible to provide. Per Bostrom, "even if the probability were much smaller (say, ~1%) the subject matter would still merit very serious attention because of how much is at stake".[127] What is more, as Ord et al. (2011) show, the chance of getting our estimates of existential risk probabilities wrong *itself* is considered to be a source of existential risk.[128]

The ethics of existential risk

Existential risk researchers, writing in the tradition of Derek Parfit and Carl Sagan, approach the problem of human extinction from a strictly pan-generational, utilitarian ethical framework,

[119] See for instance Scharf, C. (2016) for an overview of the debate; For the latter position, see for instance Pisani as referred to in Matheny, J. (2007), p. 1337.
[120] See Leslie, J. (1996), pp. 187 – 237.
[121] For an overview, see Bostrom, N. (2002). pp. 16-19; or Baum, S. (2010), pp. 594-597.
[122] Rees, M. (2004), p. 20.
[123] Bostrom, N. (2002), p. 19.
[124] Matheny, J. (2007).
[125] Leslie, J. (1996).
[126] Sandberg, A. & Bostrom, N. (2008), p. 1.
[127] Bostrom, N. (2002), p. 19.
[128] Ord, T. et al. (2010).

assigning a stable, positive and aggregate value to every individual human life, independent of when such a life might be led.[129] From this point of view, an existential risk not only threatens the lives of the billions of people alive at any given point in time, but of potentially trillions more. Translated into a simple expected value calculation this means that, no matter how small the probability of an existential threat might be, the expected value of the negative impact will nevertheless remain astronomically high if it is assumed that humanity could otherwise expect to exist for several thousands, millions, or even trillions of years more.[130] By implication, the expected value of even the smallest reduction in existential risk exposure turns out to be extraordinarily high: Bostrom calculates that under 'conservative assumptions' "reducing existential risk by a mere one millionth of one percentage point is at least a hundred times the value of a million human lives".[131] Martin Rees similarly argues that the stakes are so high that "those involved in this effort [to reduce existential risk levels] will have earned their keep even if they reduce the probability of a catastrophe by one in the sixth decimal place".[132] In sum, these authors argue that even the smallest probability of existential catastrophe is highly practically significant.[133] This is what distinguishes the category of existential risk, as used by the FHI, the CSER, or the FLI, from even the worst global catastrophes mankind has experienced so far, from famines to plagues, world wars, pandemics and pestilence. Judged from within the ethical framework of existential risk, such catastrophes appear as mere "setbacks" when compared to human extinction.[134]

Within this pan-generational utilitarian framework, existential risk reduction becomes morally paramount.[135] According to Bostrom, existential risk reduction is the most important global public good and he argues that it should become a global priority, serving as a focus for long-term global political efforts.[136] Generally he argues that our political efforts should be guided by a moral principle he refers to as the 'Maxipok rule': "Maximise the probability of an 'OK outcome', where an OK outcome is any outcome that avoids existential catastrophe".[137] This rule of thumb, might seem entirely non-controversial and common-sensical, but existential risk researchers point out that in reality it is rarely followed. According to them, researchers and policy makers tend ignore or underestimate the significance of low-probability, high-impact risks, particularly in the case of

[129] Bostrom, N. (2013), p. 16; Beckstead, N. (2013); Singer, P. et al. (2014).

[130] Pamlin, D. & Sandberg, A. (2015).

[131] "Even if we use the most conservative of these estimates [author's note: estimates of how many descendants we could have in total], which entirely ignores the possibility of space colonisation and software minds, we find the expected loss of an existential catastrophe is greater than the value of 10^16 human lives. This implies that the expected value of reducing existential risk by a mere one millionth of one percentage point is at least a hundred times the value of a million human lives." See Bostrom, N. (2013), p 18-19.

[132] Rees, M. (2014), p.1.

[133] Bostrom, N. (2013); Matheny, J. (2007); Posner, R. (2004); Weitzman, M. (2011).

[134] Matheny, J. (2007), p. 1337.

[135] Ord, T. et al. (2010), p. 1.

[136] Bostrom, N. (2013), p. 1.

[137] Ibid, p.19. Similar recommendations can be found in Beckstead, N. (2013), Beckstead, N. et al. (2014), Rees, M. (2003, 2013, 2014).

technological risks.[138] Rees laments the fact that most people tend to worry disproportionally about minor risks, such as carcinogens in food, or air crashes, while global catastrophic risks are largely ignored, including by industry and politicians.[139] This diagnosis is supported by several prominent legal and economic scholars such as Weitzman, Posner, or Sunstein, who similarly identify low-probability, high-impact risks, or 'fat-tail risks', as a systematically neglected and ignored category of problem and list a variety of psychological heuristics and politicial forces that underly the neglect of such risks.[140] Wiener (2016) refers to this problem as the "tragedy of the uncommons" and argues that in the case of existential risks precautionary action should be taken not because of the uncertainty involved, but because of the inability to learn from the catastrophe, should it materialise. At the heart of the failure to take catastrophic events seriously and implement precautionary measures, Bostrom identifies a psychological bias he terms, (referencing Voltaire's *Candide*), the 'Panglossian view': the idea that "the past record of success gives us grounds for thinking that evolution (whether biological, memetic, or technological) will continue to lead in desirable directions".[141]

Existential Risk Policy

Based on these ethical considerations, existential risk researchers and institutes formulate policy recommendations. Given that they expect most existential risks to originate from emerging technologies rather than from natural events or environmental degradation, most of their policy recommendations, too, focus on technological developments and on providing policy makers with guidelines on how to increase the chances that these developments lead to 'OK' outcomes. I will not attempt to provide a comprehensive overview of these recommendations at this point, but limit my analysis to the presentation of a few representative examples and brief summary of their core ideas.

The distinction between macro-strategic existential risk research and domain-specific existential risk research unsurprisingly permeates the policy recommendations, which also fit these categories. On a macro-strategic level, the key public policy we should adopt is to turn our default approach to managing technological change upside down, from a reactive approach to a proactive, precautionary approach.[142] The dominant approach to regulating scientific and technological developments, at least in sensitive areas of research, can no longer be based on 'learning by doing', or 'trial and error', because, as Bostrom puts it, one cannot learn from errors if there is no ex post.[143] Historically, a 'learning by doing' approach was a feasible and even a highly effective way to handle

[138] See for instance Wiener, J. (2016).
[139] Rees, M. (2013).
[140] See for instance Weitzman, M. (2011); Posner, R. (2006); Sunstein, C. (2005, 2009).
[141] Bostrom, N. (2004), p. 339.
[142] Bostrom, N. (2013), p. 27.
[143] Bostrom, N. (2002), p. 2.

technological progress because the downsides of new technologies were not only compensable but usually “small compared to their benefits”.[144] However, in a situation, where the adverse effects of technology, whether caused by error or terror, pose an existential threat, this ratio surely is inverted and a reactive approach rendered unsustainable.

This is why, according to existential risk research, humanity is entering a new era where our relationship to technological change must be fundamentally rethought. Rees claims that “there is too little planning, too little horizon-scanning, too little awareness of long-term risks” and that “the balance of effort in technology needs redirection – and to be guided by values that science itself can’t provide”.[145] Bostrom similarly argues that contemporary policies and attitudes are ill prepared for technological existential risks in that “we have no evolved mechanisms, either biologically or culturally, for managing such risks”.[146]

Existential risk theorists therefore call for global stewardship, a new approach of regulating technological developments, based on foresight and precautionary action. As a general guideline Bostrom suggests the adoption of a ‘principle of differential technological development’. This principle holds that society should “retard the development of dangerous and harmful technologies, especially ones that raise the level of existential risk, and accelerate the development of beneficial technologies, especially those that reduce the existential risks posed by nature or by other technologies”. The principle of differential technological development will be discussed at greater length below.[147]

More specific policy recommendations can be found in the institutes’ policy briefs. One such report, published jointly by CSER and the FHI, proposes that present day policy should focus on two sub-categories of policies. Firstly, on policies aimed at improving the state of knowledge about existential risk by funding and initiating more private and public research projects on both levels of existential risk research (macro-strategy and domain-specific). This is to help identify potentially beneficial and/or hazardous technologies as well as priority areas for precautionary governmental action. Secondly, policies should be put in place that help building safety into institutions, for instance by creating governance structures with decision-making processes that explicitly take into account future generations. In the long run, furthermore, policies should be adopted, which reduce the risk that dangerous technologies will be misapplied. Such policies could include for instance government oversight over the total amount of funding spent on research in high risk areas, regulations that require all researchers to register on a central database in such areas, setting “up an initiative to give developing countries access to safe technologies in exchange for setting up safety and monitoring systems to protect against accidents and terrorism”, etc.[148] In a contribution to the

[144] Beckstead, N. & Ord, T. (2014), p. 116.
[145] Rees, M. (2014).
[146] Bostrom, N. (2002), p. 2.
[147] Ibid.
[148] Beckstead, N. et al. (2014).

2014 annual report of the UK government's chief scientific adviser a researcher of the FHI recommends institutionalising "horizon-scanning efforts, foresight programs, risk and uncertainty assessments and policy-oriented research" as well as putting in place a "special intelligence service to ensure that we know what misuse some technologies are being put to".[149] As a final example, a recent report on 'Existential Risk: Diplomacy and Governance' was published jointly by the Global Priorities Project (GPP), the FHI and the Finnish Ministry of Foreign Affairs. Policy recommendations ranged from institutionalising the political representation of future generations, to making 'existential risk negligence' a crime against humanity, to the foundation of a UN Office of Existential Risk Reduction.[150] Pamlin and Armstrong (2015) suggest the foundation of a Global Risk Organisation (GRO), of a 'Global Risk and Opportunity Indicator', of global 'early warning systems', and systematically representing future generations in existing policy-making structures.[151]

Currently, there is little evidence to suggest that the existential risk movement has succeeded in raising awareness for such far-reaching, macro-strategic considerations, either in government or amongst prominent non-governmental organisations.[152] The story is very different on the domain-specific levels, however, where the existential risk movement has been remarkably successful in raising awareness for their concerns and focusing the attention of policy makers, industry leaders, and civil society, on potential black swan events.

The GPP's report on 'Unprecedented Technological Risks' was widely shared in the UK government: a section by Séan Ó hÉigeartaigh and Huw Price of CSER on risks associated with geo-engineering technologies eventually became a chapter of the Government Chief Scientific Adviser's 2014 annual report.[153] Martin Rees and Partha Dasgupta, also of CSER, wrote a joint statement with the Pontifical Academy of the Sciences and Social Sciences on 'Climate Change and the Common Good', in which the Catholic Church is urged to take a leading role in combating climate change by leveraging its unique influence on public opinion and mobilising public funds.[154] The area in which existential risk considerations are featuring most prominently, however, is undoubtedly artificial intelligence, to which we return in chapter 4.

[149] See Beckstead, N. & Ord, T. (2014), p. 116.

[150] See Farquhar, S. et al. (2017).

[151] See for instance Pamlin, D. & Armstrong, S. (2015).

[152] That said, in 2018 an All-Party-Parliamentary-Group for Future Generations was formed in the UK House of Commons, a process which was supported and encouraged by CSER, which will also provide the secretariat for the group for the coming years (see https://www.appgfuturegenerations.com, last checked on July 26, 2018). Furthermore, Jason Matheny, a former researcher at Oxford University's Future of Humanity Institute, since 2015 serves as director of the Intelligence Advanced Research Projects Activity, an organisation within the US government's Office of the Director of National Intelligence (see: https://www.iarpa.gov/index.php/about-iarpa/leadership, last checked on May 15, 2018). As director, Matheny has overseen an increase in investment in research in biotechnology, cybertechnology, and artificial intelligence, which he, according to a recent interview in the Bulletin of the Atomic Scientists, considers to be the "three areas of defense technology" that concern him "most as potential threats" [Eaves, E. (2018)]. This is a position he has voiced at earlier occasions, for instance during his time as a researcher at the FHI, see Sandberg, A., Matheny, J. & Cirkovic, M. (2008).

[153] See Annual Report of the Government Chief Scientific Adviser (2014), p. 117.

[154] See Dasgupta, P., Ramanathan, V., Raven, P. et al. (2015).

In none of these instances can the material effects, if any, of the contribution of existential risk researchers be positively established. But the fact that the existential risk research community seeks and finds a governmental audience shows that it is not an ivory tower community, but one with the ambition to inform government policy and regulation and that its concerns are, to some extent, being taken seriously.

1.6 Existential risk theory – an analysis

My aim in the first section of this chapter was to tease out the common foundations of the emerging genre of existential risk theory. What I have established so far is that existential risk research appears to be onto something new with its integrative take on human extinction scenarios. Macro-strategic existential risk research presents us with a distinctive perspective on the future of humanity. By making potential end-time scenarios its reference point for reflecting about human affairs, present and future, it manages to assume a singularly detached and ostensibly objective standpoint on the topic. In a curious way, by making potential *endpoints* of humanity its reference point for thinking about the future, the future opens up as a field of scientific inquiry and emerges as an open-ended "obstacle-course",[155] or mine-field, which humanity must navigate.[156] Beyond that I have set out to provide a brief overview of propositions that appear to be associated with and/or follow from this form of theorising. The three main propositions can be summarised as follows:

> P(1) That the issue of existential risk is widely ignored and neglected by policy makers, academics, and the wider public, leading to a situation whereby the severe challenges humanity is facing are largely neglected, and existential risks not only underestimated, but poorly understood.
>
> P(2) That mankind has entered a new era in which it is exposed to unprecedented technological risks on ever more frontiers. Torres (2016) summarises this conviction thus: "existential risks are more likely to kill you than terrorism".[157] This, according to Bostrom, Hawking, Ord, Rees, Baum, et al. marks the beginning of a critical phase, perhaps the most critical phase in human history, whereby it is the actions of whatever generation that happens to be alive that determine whether humanity has a future or not.[158] Since existential risks have a virtually infinite negative expected value, technological existential risk mitigation is morally paramount and should become a global priority issue.

[155] Torres, P. (2017b), p. 2.
[156] Häggström, O. (2016), p. 6.
[157] Torres, P. (2016).
[158] Rees, M. (2004), p. 20.

> P(3) That the realities of this new era fundamentally challenge traditional ethical and political approaches to dealing with technological change, i.e. the ways in which technological developments are traditionally managed by public authorities and private actors. In order to live up to the challenges posed by the emergence of technological existential risks, public authorities must invest heavily in research on the subject, and support the development of a new 'science of existential risk'. Furthermore, it is recommended that policies should be adopted which will facilitate horizon-scanning, the identification, monitoring and regulation of high-risk areas of technological developments, and the taking of necessary precautionary action. Taken together, it is hoped these measures will help to better understand, evaluate, and manage technological progress and to steer humanity safely through the dangerous waters of the future.

In this section my focus shifts from the descriptive to the interpretive, from the question what existential risk theory *is* to the question what the new conceptual toolkit it provides us with *implies* from a critical perspective. As it stands, existential risk research appears to be rather straightforward and uncontroversial, perhaps due to the intuitive appeal and plausibility of the generalised outlook on existential risk.

However, if one takes a closer look at the literature, it becomes clear that existential risk research is predicated on a set of tacit, more problematic assumptions. In the remainder of this chapter, I discuss some of these complications and tease out underlying assumptions. My aim is to spell out more clearly what existential risk research implies if one looks beyond the technical language. My main argument here is that existential risk theory, though ostensibly concerned with the study of threats to the survival of the human species, quickly turns into a story about the relationship between technology and humanity; that frames technology as humanity's destiny. I argue that this appears to be a direct consequence of existential risk research's quest to make the problem of human extinction the benchmark for thinking about the future of humanity. Albeit new in the narrow sense described above, existential risk theory, in a more substantive reading, therefore resonates with a rich tradition of thought in philosophy and political theory, echoing both the high hopes and deeply rooted anxieties about modern science and technology's role in human affairs. How the perspectives on that relationship relate to such older questionings will be discussed in subsequent chapters.

Defining existential risk

In order to develop a better understanding of the complications and assumptions characteristic for existential risk research one need not look far. Complications in fact begin to emerge with the search for a suitable definition of ‘existential risk’ and ‘existential catastrophe’. Cotton-Barrat and Ord (2015) of the FHI argue that, to begin with, it might seem most straightforward to use the following definition of ‘existential catastrophe’:

Definition 1: "An existential catastrophe is an event which causes the end of existence of our descendants".[159]

This definition is essentially equivalent to **Definition 1** of existential risk (p. 17), and according to which existential risks are ‘all threats that could cause our extinction’. Under this definition, an existential catastrophe is simply the materialisation of a thus defined existential risk, i.e. an event which causes the human species to go extinct. However, as Cotton-Barratt and Ord point out, limiting the use of the category of existential catastrophe to human extinction events would mean to exclude events that might indirectly lead to human extinction, in which case it "would seem sensible to refer to these events as existential catastrophes too, rather than only the event that ultimately triggers extinction physically".[160]

To illustrate this point, the authors ask the reader to consider the following thought experiment: "A totalitarian regime takes control of Earth. It uses mass surveillance to prevent any rebellion, and there is no chance for escape. This regime persists for thousands of years, eventually collapsing when a super-volcano throws up enough ash that agriculture is prevented for decades, and no humans survive".[161] Cotton-Barratt and Ord claim that it would be misleading to conceive only of the volcano eruption as an existential catastrophe because "the worst of the damage was done earlier", namely when the totalitarian regime rose to power. After this point they argue it was only a question of time "until something or other would finish things off", whether it is the eruption of a super volcano, a meteor strike or the implosion of the sun, and therefore Cotton-Barratt and Ord argue that one should be "able to talk about entering this regime as the existential catastrophe, rather than whatever event happens to end it".[162] For this reason Bostrom suggests the adoption of the following definition of existential risk:

Definition 2: "An existential risk is one that threatens the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development".[163]

Under Bostrom's definition, threats, such as entering a state of permanent totalitarian control

[159] Cotton-Barratt, O. & Ord, T. (2015), p. 1.
[160] Ibid, p. 2.
[161] Ibid.
[162] Ibid.
[163] Bostrom, N. (2013), p. 15.

resulting in the stagnation of scientific and technological progress, would count as an existential risk because it would imply the destruction of our potential for 'desirable' future development and thus constitute the cause for humanity's eventual extinction. This definition is now the most widely used definition across the field.[164]

Cotton-Barratt and Ord, however, argue that this definition too has limitations because, as the authors point out, 'potential', as opposed to 'extinction', is not a binary term. That is to say, it is hard to determine in advance at what point our potential for desirable future development would be destroyed. The definition makes it necessary to determine ex ante whether or not to include a risk under that category, a difficult task. According to Cotton-Barratt and Ord, it is for instance not clear whether an event that severely curtails the potential for desirable future development, but where there is still a slight chance that humanity may regain that potential, should be considered an existential catastrophe or not. On the one hand, if desirable future development is indeed prevented, we would have to retrospectively conceive of the event as an existential catastrophe, on the other hand if the loss of potential turns out not to be permanent, it would make little sense to consider it an existential catastrophe. It would thus not be entirely clear whether the risk of such an event should be seen as an existential risk or not. Cotton-Barratt and Ord argue that it should be regarded as an existential risk because it includes the *possibility* of leading to extinction and therefore suggest an even more inclusive definition:

Definition 3: "An existential catastrophe is an event which causes the loss of a large fraction of expected value".[165]

This definition works better for Cotton-Barratt and Ord. If, for instance, "we enter into the totalitarian regime [with a slight chance of escape] and then at a later date the hope of escape is snuffed out, that represents *two* existential catastrophes under this definition. We lost most of the expected value when we entered the regime, and then lost most of the remaining expected value when the chance for escape disappeared." An existential risk would thus simply be one that threatens to lead to "the loss of a large fraction of expected value".

Each step between Definition 1 and Definition 3 is obviously marked by an increased level of inclusivity. Definition 1 is rather narrow, limiting the use of the notion of existential risk to events that *themselves* threaten to annihilate humanity, which is helpful because it is a binary criterion (humanity either goes extinct as the result of a given event or it does not). In Definitions 2 and 3 the defining criteria become much more inclusive, stretching from the destruction of our potential for desirable future development to simply the loss of a large fraction of expected value. This makes it rather difficult to judge whether an event should be counted as an existential catastrophe, or as

[164] See for instance Farquhar, S. et al. (2017), or Singer, P. (2015), p. 165.
[165] Cotton-Barratt, O. & Ord, T. (2015), p. 2.

entailing the risk thereof, or neither.

This difficulty in defining existential risk is reflected in much of the recent existential risk scholarship. Every global catastrophic risk effectively has the potential to turn into an existential risk. But since risk *is defined* as the likelihood of a possible outcome, this means that every global catastrophic risk *is* or at least harbours existential risk. As a result, we end up with highly inclusive definitions such as Cotton-Barratt's and Ord's, because ex ante it is impossible to exclude the possibility that humanity never recovers from such an event.

To an extent, however, such definitions appear to defy the very purpose of existential risk research. Indeed, given that the concept of existential risk is intended to clarify and focus our anxieties, to bring our attention on the 1% difference between the scenarios in Parfit's thought experiment (cf. p. 9), such highly inclusive definitions seem unsuitable. The very purpose of the concept of existential risk is to raise attention to the fact that an existential catastrophe (the materialisation of an existential risk) is not just a massive global catastrophe; that it is not, figuratively speaking, about the 99% but that it is first and foremost a 'catastrophe of time',[166] with almost all of its damage contained in the loss of value that is locked in the future, which in turn is represented by the 1%. Yet, Definition 3 defines an existential catastrophe simply as the 'loss of a large fraction of expected value'.

Technology in existential risk theory

Underlying this struggle to define existential risk appears to be a central observation in existential risk research which will be discussed in subsequent chapters, specifically with reference to Heidegger and Arendt. The observed problem is that our increasing technological powers are inadvertently placing ever more parameters of our existence, across all dimensions, in our own hands; something that was not the case for previous generations. Concepts such as the Anthropocene speak of the perceived transformation in question. The problem is near complete responsibility for ever more aspects of human and natural life, resulting in a situation where no risk can be understood as fully 'natural' any longer. In existential risk research this observation is reflected in five premises that are commonly made in the field:

> (1) That humanity will at some point in the future be confronted with events that would, under normal conditions, cause its extinction – such as massive meteorite strikes, super volcanoes eruptions, the reversal of the poles, the implosion of the sun, etc.
>
> (2) That, in principle, none of these natural events must necessarily result in human

[166] Svetlana Aleksievich employs the phrase 'catastrophe of time' in her book Chernobyl Prayer in order to refer to the reactor explosion in Chernobyl, see Aleksievich, S. (2016), p. 24.

extinction. Humanity, it is argued, could develop means and policies either to prevent the event itself or, if it cannot be prevented, to avoid extinction by adequately preparing for its consequences and taking necessary precautions to survive it, such as colonizing other planets, building shelters on earth, or devising other schemes to keep the survivors of any such catastrophes alive for as long as possible. The implication is that that humanity could theoretically survive for an indefinite amount of time, if only it plays its cards right.

(3) That today and for the foreseeable future, existing technologies, technological developments presently underway, and yet unknown technological developments, constitute the greatest source of existential risk.

(4) That, "if scientific and technological development efforts do not effectively cease, then all important basic capabilities that could be obtained through some possible technology will be obtained".[167]

(5) That the sequence of technological and scientific discoveries can be managed and that there are more and less risky ways of doing this.

Below, each of these premises is discussed in greater detail and specifically how their interplay allows us to develop a better idea of what is new and distinctive about existential risk research. Taken together the first two premises imply that an event such as the rise to power of a Luddite global totalitarian regime that would deliberately prevent humanity from 'playing its cards right' must be considered an existential catastrophe. In sum, *any event that directly (by destroying humanity) or indirectly (by preventing humanity from developing the means to prevent existential catastrophes) prevents humanity from existing for an indefinite amount of time would have to be considered an existential catastrophe*. This can be seen as the macro-strategic core insight of existential risk research. It presents us with a thoroughly new perspective on the future as a technological optimisation problem which results seamlessly from making the generalised perspective on potential endpoints of the species the benchmark for thinking about human affairs, present and future.

Central to this argument is the distinction between natural existential risks and anthropogenic existential risks. Natural existential risks are risks the original cause of which is independent of human action, such as meteor strikes, super volcano eruptions, or the implosion of the sun. Anthropogenic existential risks on the other hand, have their origin in human action. The risk of runaway climate change, of an all-out nuclear war, or of uncontrollable artificial intelligence are examples. It is clear, however, that the above logic renders the distinction between natural and

[167] Bostrom, N. (2009), p. 191.

anthropogenic existential risks obsolete: the idea is that, in principle, any existential catastrophe can be prevented by adequate preparation turns every natural existential risk into an anthropogenic risk. Any given natural existential catastrophe turns into a *failure to prepare or even a failure to prepare for preparation* in one way or another. As a result, all existential risks turn into an anthropogenic problem.

Existential risk researchers distinguish between a variety of possible classification systems for interventions to reduce existential risks. One classification system distinguishes between four types of interventions, focusing on different time points in the development of an existential risk: prevention, response, endurance and recovery.[168] Preventative interventions reduce the likelihood that the event itself occurs, or at least attempt to reduce the likelihood that the risk becomes existential. Response interventions help improve the capacity to manage the immediate impact of an event, to ensure that a catastrophe does *not* turn into an existential one. Endurance-type interventions (making it easier for people to survive the aftermath) and recovery-type interventions (making it easier to "rebuild a flourishing civilisation" after catastrophe has struck) focus on later time points during which, otherwise, a chain of cause and consequence could turn result in an existential catastrophe. [169] Other authors distinguish between cross-cutting and risk-specific, or between direct and capacity-building types of intervention.[170]

What these different ways of thinking about existential risk imply is that all existential risks effectively turn into technological management problems. In technology, humanity finds the means to avert and prepare for what might otherwise turn into natural or anthropogenic existential catastrophes (building meteor shields and refuges on Earth, colonising other planets, etc., all require the further development of our technological capabilities). The potential failure to adequately prepare for any given natural doomsday scenario is therefore equivalent to a failure to develop the technological capacities necessary to avoid extinction. Technological progress or development hence assumes an eschatological quality. Without it, natural existential catastrophes cannot be averted - we must progress or perish.

As we have seen above, however, existential risk researchers typically hold that technology also is the greatest *source* of existential risk.[171] Natural existential threats, they argue, need not overly

[168] Green, B. P. (2016).

[169] Examples are pervasive in the existential risk literature. The civil engineers Denkenberger, D. & Pearce, J.M. (2015), for instance, have devised schemes on how survivors of a nuclear winter or a super-volcano eruption could be provided with sufficient nutrition for at least 5 years by turning otherwise inedible organic material into food-sources. They therewith hope to increase the chances that a global catastrophe does not turn into an existential one. The idea has been widely shared in the existential risk community through the Global Catastrophic Risk Report 2017, see Abe, N., et al. (2017); See also Jebari, K. (2015) in that context. The author provides a comprehensive overview and discussion of the effectiveness of many often proposed resilience strategies in the field - such as building shelters on earth, the Moon or Mars, as well as many other potential existential risk mitigation strategies.

[170] See for instance Farquhar, S. et al. (2017), p. 17.

[171] Beckstead, N. et al. (2014).

concern us because the probability of their occurring in any given century is exceedingly small.[172] When Rees puts the chances of existential catastrophe for this century at 50 per cent, Bostrom at 30 per cent and Musk at 20 per cent, the great bulk of risk must therefore come from technology or other anthropogenic sources. It is commonly argued in the literature that we need to focus on reducing technological existential risk before committing resources to reducing the level of natural existential risk we are exposed to.[173] Existential risk researchers today are predominantly worried about the potential effects of nano-technology, synthetic biology, artificial intelligence, and geo-engineering. However, they also argue that risks associated with these technologies might merely be the tip of the iceberg of risks yet to come. As Bostrom puts it, "as our powers expand, so will the scale of their potential consequences—intended and unintended, positive and negative".[174] We can therefore identify the above introduced third premise:

> (3) That today and for the foreseeable future, existing technologies, technological developments presently underway, and yet unknown technological developments, will constitute the greatest source of existential risk.

Combined, premises one to three create what we might consider the central problem of existential risk: On the one hand, they imply that we need to keep developing our technological capacities in order to be able to mitigate natural existential risk, particularly in the long term. On the other hand, they imply that the very development of our technological capacities constitutes the greatest source of existential risk in the short term and for the foreseeable future, likely even beyond that.

The situation is further complicated when we consider that many, perhaps all, technologies have an ambivalent status regarding their effect on the overall level of existential risk. Geo-engineering technologies might for instance allow us to counteract catastrophic climate change, which we otherwise might fail to prevent. At the same time geo-engineering technologies entail existential risks of their own.[175] Seth Baum and Anthony Barrett of the GCRI summarise the ambivalence problem in two dilemmas: "One dilemma occurs when actions to reduce global catastrophic risk could harm society in other ways, as in the case of geoengineering to reduce catastrophic climate change risk. Another dilemma occurs when reducing one global catastrophic risk could increase another, as in the case of nuclear power reducing climate change risk while increasing risks from nuclear weapons".[176]

An especially ambivalent and interesting case is artificial intelligence. Artificial intelligence is generally conceived of as an aid to our problem-solving capacities, and progress in the development

[172] Matheny, J. (2007), or Bostrom, N. (2003, 2013).
[173] See for instane Price, H. (2013).
[174] Bostrom, N. (2013), p. 16.
[175] Price, H. & O'Heigeartaigh, S. (2014), p. 117.
[176] Baum, S. & Barrett, A. (2017), p. 1.

of artificial intelligence is expected to help humanity in all sorts of ways – no problem, not even the one of existential risk, is believed to be beyond the help of AI. Progress in narrow AI (i.e. domain-specific AI systems) is therefore generally welcomed and expected to help humanity improve its task by task problem solving. At the same time, some researchers claim that severe risks, perhaps even catastrophic risks, are posed by progress in narrow artificial intelligence, for instance by increasingly powerful and independent autonomous weapons systems, warning against the risks associated with global arms races.[177]

The ambivalence, however, of course is particularly pronounced where so-called superintelligence is concerned. As we have seen, the prospect of the introduction of such an intelligence is tied up with fears regarding existential risk. This is why many existential risk researchers caution against rapid, undiscriminating, across-the-board progress in artificial intelligence, including developments which are not explicitly about building artificial general intelligence but merely about building more capable narrow AIs. On the flipside, superintelligence is believed to have the potential to free humanity from most Earthly burdens, including existential catastrophe prevention. Demis Hassabis, the founder of Google DeepMind, encapsulates the underlying mindset neatly when he argues that Google DeepMind's efforts are aimed at "solving intelligence, and then using that to solve everything else".[178] But 'solving intelligence' means that we need to get it exactly right in order to prevent a 'rogue AI' from coming into existence. It is therefore not entirely clear whether we should hasten our quest to find a solution or delay it.[179] These issues are discussed in greater detail in chapter 4 (see specifically section 4.1).

From the perspective of existential risk theory, therefore, the question is not whether we ought to further develop our technological capabilities or not. The problem is framed in terms of progress or perish. The question is *how* we should progress so that we can ensure that technological progress unfolds as safely as possible and the overall amount of existential risk we are exposed to at any given point in time is minimised. The question how to think about and deal with the problem of existential risk thus turns into the question how to think about and deal with the problem of technological progress. Taking existential risk theory seriously implies that mastering and controlling technological progress, making it *safe,* becomes the ultimate imperative for humanity.

However, if 'ought' implies 'can', then before asking how *best* to steer technological progress, we must ask whether humanity *can* steer technological progress. Nick Bostrom has a rather ambiguous position on the question. On the one hand, he introduces another premise, referred to as the *technological completion conjecture*, which holds that

[177] In 2015 the Future of Life Institute published an open letter concerning autonomous weapons, specifically highlighting potential risks associated with an arms race. By now this letter has been signed by almost 4000 robotics and AI researchers. Please see FLI (2015a).

[178] Demis Hassabis, as quoted in Simonite, T. (2016).

[179] Elizier Yudkowsky of Berkeley's Machine Intelligence Research Institute (MIRI) wrote a comprehensive paper on such macro-strategic considerations in the context of artificial intelligence (AI), entitled 'Artificial Intelligence as a Positive and a Negative Factor in Global Risk'. See Yudkowsky, E. (2008).

> (4) “if scientific and technological development efforts do not effectively cease, then all important basic capabilities that could be obtained through some possible technology will be obtained”.[180]

The technological completion conjecture posits that there are certain given basic technological capabilities and thus implies that the kinds of technologies we are going to develop are to some extent predetermined. It is a perspective on technology that is based on the concept of ‘discovery’, which assumes that there is an objective realm of capabilities that can be technologically realised which merely needs to be uncovered, independent from the particularities of human activities. Of course, technological progress is ultimately propelled by humans, and therefore remains dependent on some form of human agency – this is why Bostrom must allow for the possibility that technological developments cease entirely. But the completion conjecture implies two things.

First, that it is ultimately a matter of either or: either technological development will be ‘complete’ at some point, i.e. humanity’s technological capabilities will be developed to their fullest possible extent, or technological development comes to a complete halt. There is no middle-ground. Bostrom underpins that point when he argues that “It [the technological completion conjecture] would be false if some important capability can only be achieved through some possible technology which, while it could have been developed, will not in fact ever be developed even though scientific and technological development efforts continue”.[181] In other words, the completion conjecture holds that humanity cannot decide *not to* realise a specific technological capability if the means for doing so are, in principle, at hand.

Second, the completion conjecture implies that, if humans do act in ways broadly conducive to technological development, technology will develop along lines that are in the grand scheme of things independent from the contingencies of economics, history, culture, politics, etc. If *all* important basic capabilities are going to be developed, this means that, independent of actual choices made by humans, the same *range* of important basic capabilities will be realised. To put it differently, Bostrom’s argument implies that the range of capabilities that can be obtained by technology is predetermined – that there is a bandwidth of technological capabilities that exists independently from the actual technologies in the form of tools, machines, know-how, etc., which factually materialise as time progresses. Technologies are seen as material realisations of a range of capabilities that have an independent, abstract reality of their own, best understood perhaps as analogous to Plato's realm of ideas.

On the other hand, Bostrom argues that whilst in the long run we have no leverage over *whether* a technological capability will be developed or not, we can affect “*when* it is developed, by

[180] Bostrom, N. (2009), p. 191.
[181] Ibid, p. 192.

whom, and in *what context".*[182] His 'principle for differential technological development' suggests that humanity should "retard the development of dangerous and harmful technologies, especially ones that raise the level of existential risk; and accelerate the development of beneficial technologies, especially those that reduce the existential risks posed by nature or by other technologies".[183] As we have seen, one discussion point is how advancing the pace in progress toward advanced artificial intelligence would affect total global existential risk levels. Yudkowsky (2008) argues that 'friendly' superintelligence should be developed before advanced nanotechnology because the former could help reducing the risks associated with the later but not the other way around.[184]

If 'ought' implies 'can', Bostrom thus assumes that technological progress can be controlled in the sense that we can influence the order of arrival of technological innovations. He claims that we should "think of a discovery as an act that moves the arrival of information from a later point in time to an earlier point".[185] In fact, he likens the process of technological innovation to progress in mathematics: "A scientist or a mathematician may show great skill by being the first to find a solution that has eluded many others; yet if the problem would soon have been solved anyway, then the work probably has not much benefited the world".[186] This leaves us with a fifth central claim of existential risk theory regarding technology:

> (5) The sequence of technological and scientific discoveries can be managed and there are more and less risky ways of doing this.

Bostrom, as well as other existential risk researchers, often writes in the future perfect tense. That is, he projects himself into an arbitrarily distant point in the future and asks what conditions would have to have been met should humanity still exist at this future reference point in time. Looking back from such a distant point in the future, he conjectures that, if humanity still exists, technological progress cannot have stopped in the meantime because otherwise humanity would necessarily have perished at some point along the way. Crudely put, there are only two categories of events that can put a halt to technological progress: natural existential catastrophes or anthropogenic existential catastrophes, which include events that end technological progress. Either a natural or an anthropogenic existential catastrophe happens to destroy humanity and therefore technological progress with it. Or humanity stops technological progress, which would merely move the existential catastrophe to an earlier point in time, because it would leave humanity defenceless vis-à-vis natural existential threats. Viewed from a sufficiently distant point in the future, then, there are broadly two options. Either humanity still exists, in which case it must have developed its technological capacities significantly, beyond

[182] Bostrom, N. (2014), p. 142.
[183] Bostrom, N. (2009), p. 193.
[184] See for instance Yudkowsky, E. (2008), p. 36 and Beckstead, N. (2015).
[185] Bostrom, N. (2014), p. 293.
[186] Ibid.

anything conceivable today. Or humanity no longer exists, in which case it has either been destroyed by a catastrophe which technology has caused or which technology has been powerless to prevent because it was not sufficiently far advanced at that point in time. As a result, the notion 'end of technological development' and the notion 'existential catastrophe' appear to be synonymous in existential risk theory.

Given that, from this perspective, all significant technological and scientific discoveries will be made eventually, Bostrom claims that the value of any such discovery does not equal the value of the information discovered but rather the value of having the information available "earlier than it *otherwise* would have been".[187] The 'principle of differential technological development' accordingly states that we should organise the sequence of discoveries in such a manner that the level of existential risk we are exposed to at any given point in time is minimised. In the very long run, Bostrom and other existential risk researchers hope to reach a state referred to as 'technological maturity'. Technological maturity is understood as "the attainment of capabilities affording a level of economic productivity and control over nature close to the maximum that could be feasibly achieved".[188] As Torres (2017) points out, a risk counts as 'existential' if and only if it prevents our species from realising the hypothetical safe state of technological maturity.[189]

This leaves us in an odd and seemingly contradictory position, which Bostrom's theory tries to resolve. On the one hand, under the condition that we manage to stay clear of existential catastrophes, technological progress and indeed 'technological completion' are considered inevitable. A predetermined range of technological capabilities is expected to be realised as time progresses. On the other hand, there is room for human agency because the sequence of technological and scientific 'discoveries' as well as the *design* of technologies can be manipulated. This allows Bostrom to remain optimistic regarding technological futures whilst acknowledging the seemingly unstoppable surge of technological power. To an extent this position is reminiscent of Marxism, the "scientificity" of which, as Carl Schmitt once put it, resides precisely in the "aspiration to change the world without jumping out of history".[190]

A brief detour to one of the more prominent political theorists of technology of the 20th century might help to clarify that point. In his 1979 book *Autonomous Technology*, Langdon Winner argues that discussions of technology are plagued by a deep-seated tension between two seemingly unreconcilable and yet equally widely held conceptions of the nature of technological change. On the one hand, he argues, technology is commonly conceived of as a tool or a means which caters to our needs and allows us to gain increasingly refined control over the environment and its resources - to

[187] Ibid.

[188] Torres, P. (2017b), Bostrom, N. (2013).

[189] Torres, P. (2017b).

[190] This passage was translated by the author. The German original reads as follows: "… Marx, dessen Wissenschaftlichkeit den Anspruch bedeutet, die Welt zu verändern, ohne aus der Geschichte herauszuspringen." See Schmitt, C. (1991), p. 83.

‘make ends meet’, as it were. Winner calls this conception of technology the ‘mastery tradition’, which essentially holds that “technical tools are, by their very nature, mere tools subject to the will of whomever employs them”.[191] Tracing back its roots to philosophers such as Bacon or Aristotle, Winner claims that, implicitly or explicitly, the ‘mastery’ conception is the most commonplace conception of technology. This idea is closely related to what Winner later refers to as the position of ‘voluntarism’ - a position which holds that human beings “have full, conscious choice when it comes to technologies and technological inventions and that they are thus responsible for choices made at each step in the sequence of technological change”.[192] Voluntarists contend that “behind the massive process of [technological] transformation one always finds a realm of human motives and conscious decisions in which actors at various levels determine which kinds of apparatus, technique, and organization are going to be developed and applied”.[193]

On the other hand, Winner argues, there are accounts of technology that point to the exact opposite direction. These accounts hold that people have lost their ability to make choices or exercise control over the course of technological change. According to that view technological development goes forward virtually of its own volition, resists any limitations, and has the character of a self-propelling, self-sustaining, self-determining force. Winner's main point of reference in that regard is French philosopher Jacques Ellul, whose philosophy of technology draws an infamously bleak, fatalistic picture of technological change. Ellul is convinced that technological progress since long has escaped human control and has itself become the dominant force in history: “man participates less and less actively in technical creation, which, by the automatic combination of prior elements, becomes a kind of fate. Man is reduced to the level of a catalyst”, [194] since, Ellul continues, “when all the conditions concur, only minimal human intervention is needed to produce important advances it might almost be maintained that, at this stage of evolution of a technical problem, whoever attacked the problem would find the solution”.[195]

Bostrom’s account of technological change is characterised by the very tension Winner identifies, but attempts to somehow blend the two poles, that of ‘voluntarism’ and that of determinism, into one account. On the one hand he entertains a teleological perspective on technological change, where the path of technological development is in significant respects predetermined and the role of the individual researcher or engineer is reduced to that of the messenger rather than that of the conqueror. Bostrom’s likening of technological innovation to that of progress in mathematics is highly reminiscent of Ellul’s account of the ‘evolution of a technological problem’ where ‘only minimal human intervention is needed’. On the other hand, the entire point of existential risk research is not to succumb to fatalism, to raise awareness for the potentially

[191] Winner, L. (2001), p. 26.
[192] Ibid, p. 46.
[193] Ibid, p. 53.
[194] Ellul, J. as cited in Winner, L. (2001), p. 58.
[195] Ibid, p. 60.

catastrophic effects of technological progress in the hope that the global community will manage to assume mastery and control over that process. This presupposes, at least on some level, scope for human agency and the possibility of human stewardship.

Conclusion - technology as destiny

I have argued in the introduction to this chapter that it should be understood as an exploratory exercise. Existential risk research is still a young genre and, in my research, I have not encountered any publications that have tried to systematically embed it in a humanities context, or reflect about its philosophical and political significance, or its core concepts, from an outside perspective. Most, if not all, of the literature on existential risk is produced by authors who are part of the movement themselves and, as a result, the genre is largely self-referential. The task I set myself is to uncover the deeper philosophical themes that underpin existential risk theory and to establish to what extent the perspectives it opens up might be new by connecting existential risk theory to older debates in political theory. This is of course a rather wide question that needs some calibration. At first sight it may be unclear what debates in political theory this emerging genre may best be connected to, in order to then establish where it can add something to these debates, and, vice versa, what the respective debates can tell us about existential risk. To provide the basis for this was the purpose of this chapter, in which I have sought to tease out central underlying themes and puzzles in existential risk research.

In a superficial reading, existential risk theory tells a story about the precariousness of modern human existence. It highlights the fact that the existence of our species can not only no longer be taken for granted but, on the contrary, should be seen as more vulnerable than perhaps ever before and that it should become a global priority to mitigate existential risk. Beneath that, however, existential risk theory, I argue, should first and foremost be seen as a story about technology, or to be more precise, about the existential role technology is beginning to assume in human life and in contemporary visions of the future. Based on a mixture of empirical and normative arguments, existential risk theory presents us with a case in which prudently managed technological progress – not regress, not stalemate, both of which would have to be considered as existential catastrophes – becomes imperative. In other words, ‘taking human extinction’ seriously along the lines in which existential risk researchers do appears to translate seamlessly into a call for the perfection of technological mastery culminating in a hypothetical state of ‘technological maturity’. It thereby renders the continued extension of technological civilisation, of technological control over ever more aspects of human life and natural processes a matter of necessity. Existential risk theory therefore presents us with a rather distinctive outlook on the future of humanity, where the problems of the future of humanity and the future of technology become in effect indistinguishable. Human destiny becomes a technological optimisation problem and the role of politics and ethics is reduced to a purely instrumental, auxiliary one in this wider endeavour.

It is this perspective on technology that I am interested in and that, I argue, can be meaningfully connected to long-standing debates in political theory. Initially, existential risk theory could be seen as a critical, a cautioning voice in the face of the seemingly overpowering forces of exponentially accelerating technological progress, highlighting its potentially disastrous consequences. At one point, Bostrom for instance claims that speaking of 'technological progress' should best be avoided because 'progress' has an evaluative connotation "of things getting better", which according to him cannot be regarded a 'conceptual truth', given the potentially catastrophic downsides of present and future technologies. Bostrom instead advises to use the less value-laden notion 'technological development'.[196] In that light, one might be inclined to see existential risk research as a criticism of uncritical tech-optimism and a wake-up call vis-à-vis a condition Langdon Winner refers to as "technological somnambulism", according to which humanity is being dragged along by technological change, perhaps not against its will, but also without consciously trying to resist, shape, transform or steer it.[197]

I believe it is in debates such as these that existential risk theory can be embedded in most meaningfully and in the context of which we can further develop our enquiries into what might be new about it. As the following chapters will demonstrate, the perspective on technology presented by existential risk theory opens up interesting new perspectives on old questions in political theory and philosophy of technology, resonating with deeply rooted anxieties and hopes about modern science and technology's role in human life. Given the increasing prominence of existential risk research, it is interesting to see that this emerging genre has so far remained unconnected to such older debates.

[196] Bostrom, N. (2009), p. 192.

[197] Winner, L. (2014), p. 5.

2. The technological understanding of being

Introduction

In the preceding chapter I have argued that the central topic in existential risk theory is technology, or, rather, the inexorable intertwining of human destiny and technology. Based on a mixture of empirical and normative arguments, existential risk theory presents us with a case in which prudently managed technological progress - not regress, not stalemate, both of which would have to be considered existential catastrophes - becomes imperative. In other words, 'taking human extinction' seriously along the lines of existential risk research and the resulting macro-strategic outlook results in a binary perspective on the future of humanity, where humanity either perishes or perfects its technological capabilities. In a strict interpretation of that logic, technology turns into destiny in the sense that the extension of technological control over ever more natural processes and ever more aspects of human life becomes morally imperative.

Perhaps self-evidently, given its name, existential risk theory draws our thinking about technology into what we might call an existential space. With 'existential' I mean to denote first and foremost that the brute fact of our existence, the '*whether'* of human existence, is at stake and the circumstance that, in existential risk theory, this fact is now understood to be entirely dependent on technology. But technology is of course part of the '*how'* of life. The logic of existential risk theory hence entails that the 'how', i.e. the *terms* of human existence, also are drawn into that existential space. In fact, it can be seen as existential risk theory's core point to demonstrate that the 'whether' and the 'how'' of human existence can no longer be meaningfully distinguished, since it now depends entirely on how we live, on how 'we play our cards', as it were, *whether* there will be a future.

Whilst existential risk theory might start out with a basic concern for existence as such, its logic implies that through technology, as arbitrator over life and death, it begins to swamp the spaces in-between too. However, this discussion is largely anticipatory. The exact mechanisms by which technology appears to emerge as the arbitrator over not only the 'whether' but also the 'how' of human existence in existential risk theory will form an integral part of my discussion throughout the following chapters.

Suffice to say, for the moment, that existential risk theory encourages us to think about technology along such existential lines. This space, however, is not unpopulated. In fact, technology has been discussed as an existential question for a long time and there is a large, temporally extended, community of thinkers who, for varying reasons, have been arguing that the existential space is the only appropriate space for discussing modern technology. Since my aim is to develop a better understanding of what might be new about existential risk theory, my attention now turns towards a closer inspection of its conception of technology. Having established that the generalised perspective on human extinction scenarios indeed appears to present us with a new and distinctive framework for thinking about the future of humanity, transforming the future into a technological optimisation problem, my question now becomes whether the resulting perspective on the role of technology in human affairs also is new and, if so, in what respects.

I will commence this discussion by connecting existential risk theory to Martin Heidegger's philosophy of technology. Heidegger is a notoriously elusive and controversial thinker. Heralded by some as one of the most important philosophers of 20th century continental Europe (e.g. Hans Jonas, Hannah Arendt, Hubert Dreyfus or Mark Wrathall), and denounced by others as a charlatan or a 'self-infatuated blowhard' (Richard Rorty),[198] his philosophy has come under additional suspicion due to his membership in the Nazi party, which casts his moral and political judgement in a dubious light, to say the least.[199] As a consequence, invoking Heidegger in any context appear to be a controversial step and requires at least some clarification.

My first reason for drawing on Heidegger is that he is indispensable if one is interested in the history of philosophy of technology, i.e. in the history of philosophical reflection on technology as a subject for serious and systematic consideration in its own right. Heidegger often is regarded as one of the founding fathers of the philosophy of technology and to this day is frequently drawn on and invoked by authors in the field.[200] Of course, as Petrina (2017), or Franssen, Lokhorst, et al. (2018) argue, philosophical reflection on technology is perhaps as old as philosophy itself, listing (in chronological order) for instance thinkers such as Aristotle, Vitruvius, as well as Roger and Francis Bacon.[201] However, as historian of philosophy of technology Mitcham (1994) argues, until the 19th century reflection on technology tended to be subsumed under other aspects of philosophy (for instance in the cases of the above listed authors: causation, architecture, the arts and alchemy, and science and experimentation). Only relatively recently, Mitcham argues, has philosophy of technology emerged as a veritable, cooperative, self-declared genre of its own. The first time the term appeared was in as late as 1877, in Ernst Kapp's book *Grundlinien einer Philosophie der Technik*, and it was not until the 20th century that the genre developed traction.[202] Mitcham claims that one historical complication in the birth of philosophy of technology was that it can mean two very different things. If 'of technology' is understood as a subjective genitive it is "an attempt by technologists or engineers to elaborate a technological philosophy". [203] On the other hand, if 'of technology' is understood as an objective genitive, Mitcham argues, "then philosophy of technology

[198] Rorty, R. (2005), p. 275.

[199] For in-depth discussions of the relationship between Heidegger's support of the Nazi movement and his philosophy as well as of the complications arising from this relationship for our interpretation of his philosophy see amongst many others Farias, V. (1989), Wolin, R. (1990), or Strong, T. (2007), ch. 7.

[200] See for instance Franssen, M., Lokhorst, G., et al. (2018), Mitcham, C. (1994), Ihde, D. (2010).

[201] See Petrina, S. (2017), and Franssen, M., Lokhorst, G., et al. (2018).

[202] Olsen, J., Selinger, E., Riis, S. (2009), p. viii.

[203] Mitcham, C. (1994), p. 17.

refers to an effort by scholars from the humanities, especially philosophers, to take technology seriously as a theme for disciplined reflection".[204] It is in that second sense that Heidegger can be located at the origins of philosophy of technology.

My second reason for drawing on Heidegger is that he is even less dispensable if one is interested in the intermingling of technology and human destiny, i.e. in thinking about technology as an existential question. For Heidegger, technology *is* a 'Geschick', which William Lovitt translates as 'destining' in his seminal 1977 translation of the 'Question Concerning Technology', which I am using here. I will discuss this concept in further detail below. However, what Heidegger meant to convey with it is that technology has an inherently 'pulling' quality - that it pulls individuals and mankind at large into a specific direction, in thought and action. Therefore, when in existential risk theory the question of human destiny and the question of technology become in effect indistinguishable, this, from a Heideggerian perspective, is no surprise at all. Heidegger's philosophy of technology, as a result, provides us with a counterpoint for reflecting upon existential risk theory since both discuss technology as an existential question, but from very different angles.

Clearly, there are many other authors whose work also may have served as a starting point for this chapter's purpose – to begin carving out what might be the deeper philosophical and political significance of existential risk theory in relation to older discussions surrounding technology. One could for instance have drawn on several other early 'humanities philosophers of technology', such as Jacques Ellul (1912 – 1994), Lewis Mumford (1895 - 1990), or José Ortega y Gasset (1883 - 1955), who were of Heidegger's generation and are often included in the canon of founding fathers of philosophy of technology.[205] Their work, like Heidegger's, had enormous influence on later generations of philosophers of technology, such as Herbert Marcuse, Langdon Winner, Andrew Feenberg, Don Ihde, or Albert Borgman, who rose to prominence in the 1980s and 1990s and on whose work I will draw repeatedly throughout this thesis.

The works of Ellul, Mumford, or Ortega y Gasset, can be considered, like Heidegger's, as critiques of a specific type of Enlightenment optimism and the "idea that scientific and technological progress automatically contributes to the advancement of society by bringing about a unification of wealth and virtue".[206] Their main target of critique was a modernist spirit, often traced back to authors such as René Descartes and Francis Bacon, according to which humankind should strive to become the 'master and possessor of nature'. Part of what Heidegger set out to do, like the other authors, indeed was to show that this undertaking not only cannot be successful but, in dialectical fashion, is bound to undermine the very values it purports to serve. However, apart from the fact that the scope of this thesis would not allow for a comprehensive and systematic inclusion of all of these

[204] Ibid.
[205] Cf. Franssen, M., Lokhorst, G., et al. (2018), Mitcham, C. (1994).
[206] C. Mitcham (1994), p. 40.

writers, there are interrelated historical and conceptual considerations which suggest that Heidegger's philosophy is a particularly suitable touchstone for this project.

First, Heidegger clearly stands out amongst the above-named group of authors in terms of historical importance both within the field of philosophy of technology and beyond. His 1954 essay 'The Question Concerning Technology', which forms the backbone of my discussion of his philosophy of technology, is often referred to as the single most important text in the history of philosophy of technology in general.[207] Furthermore, Heidegger's philosophy had a lasting influence not only on later developments in continental philosophy but on a wide variety of fields of inquiry, from history, to literature, the visual arts, architecture and ecology.[208] This clearly positions him as a particularly important thinker amongst the first generation of philosophers of technology.

The second, related though perhaps more important reason relates to his role as a teacher during his time as a professor at the universities of Marburg and Freiburg in the 1920s and early 1930s. In that period Heidegger taught a number of students who later were to become eminent philosophers in their own right, including for instance Hannah Arendt, Hans Georg Gadamer, Leo Strauss, Hans Jonas, Herbert Marcuse, Günther Anders, and Karl Löwith. Amongst them, even those who broke with him following his support of the NSDAP and his unequivocal, public endorsement of Nazi ideology in the so-called 'Rektoratsrede' and 'The Introduction to Metaphysics',[209] understood him to be amongst the deepest thinkers of his time,[210] and the influence of his teachings on their work remains, on many dimensions, undeniable, if complicated.[211] My focus in Chapter 3 will come to rest on Hannah Arendt and Günther Anders, whose work is particularly illuminating in the context of existential risk. As has been repeatedly demonstrated, it is undeniable that Arendt's and Anders' thought was strongly influenced by Heidegger's philosophy.[212] Most importantly for the purposes of this thesis, their thought about modern science and technology and its pivotal role in modern human existence, from politics in Arendt to psychology in Anders, gains clarity and resonance when read against the background of Heidegger's phenomenologically rooted critique of technological modernity.[213] If one is interested in Arendt's and Anders' thought about science and technology, a rudimentary understanding of Heidegger's philosophy is therefore essential.

[207] See for instance Ihde, D. (2010), p. 12, Blitz, M. (2014).

[208] See Dawsey, J. (2017), p. 7.

[209] See Heidegger, M. (1933), p. 117, and Heidegger, M. (2000b).

[210] Blitz, M. (2014), p. 64.

[211] For in-debth discussions of Heidegger's general role, influence and legacy as a teacher see e.g. Wolin, R. (2001), or Dries, C. (2012).

[212] The complicated relationship, intellectual and personal, between Hannah Arendt and Heidegger has been subject to sustained scholarly attention. For studies of the influence of Heidegger's thought on Hannah Arendt's work see for instance Villa, D. (1996), Yaqoob, W. (2014), Hinchman, L. & Hinchman, S. (1984), for a discussions Heidegger's influence on Günther Anders' thought see e.g. Liessmann, P. (2002), Dawsey, J. (2017), Dijk, P. van (2000).

[213] For a discussion of Heidegger's important influence on Arendt's thought concerning specifically technology and science see for instance Hinchmann, P. & Hichmann, K. (1984), Yaqoob, W. (2014), for a discussion of Heidegger's equally strong influence on Anders' thought in that regard see Dijk, P. (2000).

In sum, there are historical and conceptual reasons for beginning the historical and theoretical anchoring of existential risk theory with Heidegger. Heidegger is at the origin of a tradition of thinking about technology in thoroughly existential terms and his work provides us with a set of concepts which can be meaningfully related to existential risk theory. Furthermore, it provides us with a historical basis for weaving in the thought of scholars who studied related questions throughout the following decades. I am nevertheless fully aware that other routes could have been chosen and other connections established. However, since the recently emerging genre of existential risk theory, to the best of my knowledge, has not been systematically connected to philosophy of technology before (nor to any other strand in philosophy or political theory for that matter) my goal is modestly to begin this conversation and Heidegger, I believe, is one sensible starting point.

My subsequent connection of existential risk theory to Heidegger's philosophy of technology will lead me to make two main points, which in turn rest on a range of observations I will discuss as the chapter progresses. First, that existential risk theory can be seen as a rather old-fashioned response to an old fear – the fear of losing control over technological progress resulting in an even more pronounced quest for technological mastery – but in a new context, comprising of new technologies, and under ostensibly escalating conditions. Second, that existential risk theory resonates closely with several of Heidegger's deepest fears regarding non-physical effects of modern technology. However, since it rests on the previously discussed new approach to study human extinction scenarios in an integrated manner and thus presents us with a distinctive perspective on the future it complicates Heidegger's story and his criticism of 'technological behaviour', with implications that will be discussed in chapter 3, by drawing on 'Heidegger's children'.[214]

The chapter is organised in two parts. First, an overview of Heidegger's philosophy of technology is provided. Here, key terms of Heideggerian ontology and his philosophy of technology are introduced, most importantly the concept of 'Enframing'. With Enframing Heidegger seeks to circumscribe the essence of technology, which he understands as a 'destining' of Being, a disposing power that challenges humanity to think, be and relate to all that is in a technological fashion. Further, his criticism of the everyday conception of technology is discussed, as well as what Heidegger considers to be the greatest dangers associated with technology – that beings will eventually disclose themselves exclusively as 'standing reserve', with which he describes a nihilistic world in the making and which we can understand as Heidegger's vision of humanity's technological destiny. This overview provides the historical and conceptual backdrop for the second part of the chapter. In the second part Heidegger's thus sketched out philosophy of technology is brought to bear on existential risk theory, constituting the first step in the endeavour to embed and situate existential risk theory in philosophy of technology.

[214] The term 'Heidegger's children' is borrowed from Richard Wolin's eponymous book, see Wolin, R. (2001).

2.1 Technology in Heidegger's philosophy

In the following three aspects of Heidegger's philosophy of technology are introduced and discussed. First, his claim that the essence of technology is a form of 'revealing', which means that he understands it as Dreyfus (1993) calls it, as an 'ontological condition' that governs the very way we see, understand, and act in the world.[215] Second, his argument that, as a form of revealing, technology is a 'challenging revealing' because it has an inherently forward-leaping dynamic, which compels humanity to conceive of ever more aspects of life and nature in an instrumental and calculative manner. Heidegger describes this as 'a destining'. Third, the normative dimension of Heidegger's philosophy of technology and what he considers to be the 'greatest dangers' of technological destining.

Before beginning to sketch out this overview of Heidegger's philosophy of technology, however, we must introduce a range of key distinctions in his fundamental ontology because otherwise his philosophy of technology would likely remain unintelligible. Heidegger's fundamental ontology is structured around three concepts. He distinguishes 'Dasein', 'Sein' and 'das Seiende'. Depending on the translator, Dasein is either left untranslated or it is translated as 'Being' (with capital B). Following Lovitt (1977), in this thesis Dasein will be translated as Being. With Being Hedeigger denotes the different modes in which humans 'are in the world' and experience it. Oftentimes and controversially Being is loosely equated with consciousness. But Being denotes not only the simple experience of 'being there' in the sense of the Cartesian cogito.[216] Being defines the way in which we *are in* the world and that includes the way in which we habitually conceive of the relationship between ourselves and our surroundings. 'Cogito', the reduction of our sense of reality onto our own immediate experience of selfhood in any given moment and, by extension, the subject-object distinction, can be seen as expressive of one *mode* of Being. Indeed, for Heidegger it is a mode of Being, namely the modern, but the two are not to be equated. On the contrary, for Heidegger, the Cartesian dualisms between mind and body and between subject and object, were a dangerous and ultimately self-defeating misrepresentation of Being. 'Sein', usually translated as being (lower-case b), denotes both the sheer fact of existence as such, i.e. that something exists rather than nothing, as well as the totality of all that exists.[217] 'Das Seiende', usually translated as 'beings', refers to all things that exist, either individually or collectively, i.e. it is employed either to refer to all things that exist (but not to be confused with 'being', as defined above) or to individual things that exist.

[215] Dreyfus, H. (1993), p. 305.

[216] Cf. Wheeler, M. (2011).

[217] Heidegger's work is typically referred to as *fundamental* ontology because he, as he himself believed, was the first to ponder the question of 'being' again, i.e. 'why there *is* something rather than nothing'. For Heidegger, Western philosophy since Plato had focused purely on 'beings', i.e. on the ontic rather than the ontological, things that exist within being rather than being as such. Cf. Thomson, I. (2009).

Another preliminary remark should be made concerning the position of the topic of modern technology within Heidegger's work. Heidegger's work is often divided into an early phase, gravitating heavily around *Being and Time,* which was published in 1927, and a late phase, following what Heidegger himself christened '*die Kehre'* (typically translated as *The Turn* or *The Turning)* in an eponymous lecture he gave in 1949.[218] The scope of this chapter does not allow for an in-depth discussion of *The Turn* and the precise nature of the shift in thinking it involved, not least because its intricacies are a matter of ongoing scholarly debate.[219] However, there appears to be wide agreement that the topic of modern technology began to occupy a central position in Heidegger's thought only in his later works. [220] This, however, appears to be no coincidence. Generally speaking, it is argued, that the *The Turn* was marked by Heidegger's attempt to break with what he himself considered to be remnants of subjectivism in *Being and Time.*[221] Wheeler (2011), for instance, argues that Heidegger's later philosophy "shares the deep concerns of *Being and Time*, in that it is driven by the same preoccupation with Being and our relationship with it that propelled the earlier work. In a fundamental sense, then, the question of Being remains *the* question. However, *Being and Time* addresses the question of Being via an investigation of Dasein […] the later Heidegger does seem to think that his earlier focus on Dasein bears the stain of a subjectivity that ultimately blocks the path to an understanding of Being".[222] In other words, whilst in *Being and Time* the subjective experience of the individual human being still assumed centre-stage, in his later works Heidegger seeks to address the problem of being head-on. Technology, it turns out, plays a central role in this enterprise. As will be discussed in greater detail below, for the late Heidegger, technology, in its essence, is nothing humans make but a way in which being reveals itself. The theme of modern technology hence provides him with another route into the investigation of being, a route outside of Being. It is in that vein that Borgmann (2005) argues that "technology is the most important topic of Heidegger's thought" because it became the converging ground of Heidegger's, previously separate, efforts to understand reality "in its deepest and most crucial dimensions".[223] According to Borgmann these efforts were tripartite, consisting of a) the exploration of the nature of being, b) the exploration of ancient Greek philosophy as well as German philosophy and c) the analysis of the modern human condition. These efforts, Borgmann claims, proceeded unevenly and side-by-side until they converged on Heidegger's understanding of modern technology.[224]

[218] See *The Turning,* In: Heidegger, M. (1977), p. 36 ff.

[219] For an instructive, detailed overview of contemporary debates regarding *The Turn* see for instance Sheehan, T. (2012, 2013), who distinguishes between at least three possible interpretations of the *The Turn*.

[220] Cf. Blitz, M. (2014), Wheeler, M. (2011), Dreyfus, H. & Spinosa, C. (1997, 2003).

[221] Wheeler, M. (2011).

[222] See Wheeler, M. (2011). Please be aware that Wheeler translates Heidegger's core vocabulary differently than Lovitt. Whilst, as discussed before, Lovitt translates 'Dasein' as 'Being' (capital B), Wheeler leaves 'Dasein' untranslated. 'Being' (capital B) therefore here refers to 'Sein', i.e. the fact of existence as such, which is translated as 'being' (lower-case b) in Lovitt's translation.

[223] Borgmann, A. (2005), p. 420.

[224] Ibid.

It is not my aim in this chapter, however, to develop a detailed account of the evolution of Heidegger's thinking about technology, nor to provide an account of this topic's role and status in Heidegger's oeuvre. The previous paragraphs' main purpose was to qualify why *Being and Time* is bracketed in my subsequent discussion of Heidegger's philosophy of technology which instead draws on a range of texts that were published throughout the late 1940s, -50s and -60s, most importantly the *Question Concerning Technology.*[225] The fact that *Being and Time* is bracketed here, however, does not mean, that it is not relevant in the context of Heidegger's philosophy of technology. As Blitz (2014) points out, "the most important argument in *Being and Time* that is relevant for Heidegger's later thinking about technology is that theoretical activities such as the natural sciences depend on views of time and space that narrow the understanding implicit in how we deal with the ordinary world of action and concern", that "science flattens the richness of ordinary concern". [226] We will indeed see that the clash between scientific and technological knowledge on the one hand and our 'ordinary' understanding of reality on the other hand is a core theme in Heidegger's philosophy of technology and that therefore one of *Being and Time's* main themes assumes a central role in Heidegger's later philosophy. Nonetheless, this chapter focuses exclusively on Heidegger's philosophy of technology and therefore his later work.

The fact that substantial continuities between the early and the late Heidegger's philosophy of technology do exist, however, is important to remember for historical reasons. It explains the partially substantial parallels between his philosophy of technology and Hannah Arendt's and Günther Anders' perspectives on that topic. Both these thinkers were taught by the 'early Heidegger', the Heidegger of *Being and Time,* during their student years at the Universities of Marburg and Freiburg in the 1920s and both entertained life-long, if complicated, relationships to Heidegger.[227] As we will see in the next chapter, both adopted Heidegger's phenomenologically grounded critique of modern technology and science, which is the common ground of *Being and Time* and his later works, but they diverged significantly from the deep fatalism that is characteristic for the late Heidegger's work and its complete departure from subjectivism.

2.2 Enframing

'Gestell', often translated as 'Enframing', is the term Heidegger employs to denote what he considers to be the essence of technology. In asking for its essence, Heidegger is asking for technology's 'whatness', i.e. that through which something is what it is, or which makes something what it is and

[225] Many of these essays are compiled in Heidegger, M. (1977).

[226] Blitz, M (2014), p. 67.

[227] For an in-depth discussion of the decisive influence Heidegger had on Arendt's thought see for instance Wolin, R. (2015), or Villa, D. (1996), or Dries, C. (2012). Heidegger's role in Günther Anders' intellectual development has been thoroughly discussed in Dijk, P. v. (2000), Dries, C. (2009, 2012), Liessmann, K. (2002), as well as Dawsey, J. (2017).

lets it endure as such through time. Heidegger here places particular emphasis on the temporal dimension of ‘essence’. As William Lovitt, who translated and published the first English version of the *Question Concerning Technology*, which I am using here, argues “essence does not simply mean what something is, but it means, farther, the way in which something pursues its course, the way in which it remains through time as what it is”. [228] Technology in Heidegger's view, if looked at in that sense, is not simply a neutral tool, a means to an end, or a human activity, a conception which he refers to as the "instrumental and anthropological definition of technology". For him, although being that too, technology is something much more powerful and much more fundamental than that, namely the basic ontological condition of modernity.

Heidegger's initial concern and the starting point of his discussion is that the anthropological and instrumental definition of technology cannot adequately capture what modern technology really is. He does not deny that the instrumental and anthropological definition is *correct*. However, Heidegger employs the term 'correct' in a particular way, which needs to be understood in connection to his phenomenological conception of Being. Under this conception the correct "is not yet the true". A correct statement is not *untrue* but it merely uncovers a partial truth and "fixes upon something pertinent in whatever is under consideration".[229] It is for instance correct to say that the moon is a shiny object in the night's sky. This observation, however, does of course not reveal the ‘true’, full nature of the moon, which must be considered to be much more than that. Accordingly, Heidegger holds that the instrumental and anthropological definition of technology correctly uncover a "fundamental characteristic of technology",[230] but that they do not yet uncover its essential nature. Focusing on an understanding of technology along these lines would thus mean to remain oblivious to the real nature of technology and thus its power. It is only when we uncover the essence of technology, Heidegger argues, that we can gain a free relation to it.[231]

What then is the essence of technology? The instrumental and anthropological definition of technology does, according to Heidegger, at least tell us what the central quality of technology is, namely instrumentality, which means that it is a way of attaining one's ends. Heidegger therefore begins his investigation into what the essence of technology is by asking "within what such things as means and ends belong"[232] and he observes that "wherever ends are pursued and means are employed wherever instrumentality reigns, there reigns causality".[233] Heidegger infers from this that modern technology is a specific form of causation and embarks on an in-depth investigation of the notion of causation, trying to uncover the way in which the modern technological way of causing something might differ from other forms of causation. The main reason why Heidegger characterises technology as a form of causation seems to be that in his search for the essence of technology he seeks to

[228] Please compare to Lovitt, W. (1977) in Heidegger, M. (1977), p. 3, fn. 1.
[229] Heidegger, M. (1977), p. 6.
[230] Ibid, p. 5.
[231] Ibid, p. 3.
[232] Ibid, p. 6.
[233] Ibid.

understand where modern technology, i.e. machines and modern production processes, really 'come from'. His observation is that modern technology brings things into existence and encompasses a range of activities that are qualitatively different from what nature brings into existence but also from what traditional forms of craftsmanship bring into existence. Understanding modern technology simply as a means to an end, which would put a modern hydro-electric power plant into the same category as for instance an ancient rake, is, for Heidegger, deeply mistaken. Heidegger wants to uncover the ultimate "from whence" from which the things and actions that are characteristic for modern technology "take(s) and retain(s) their (its) first departure".[234]

Heidegger arrives at the conclusion that modern technology can only be understood as the outgrowth of a revolution in concepts, in our relation to being, whereby nature has come to be understood no longer in ways in which it ordinarily occurs to us, i.e. in form of things, such as trees, or tables, or human beings, that have their own immediate reality, but in form of calculable, causal processes and functions which lend themselves to productive exploitation. The essence of technology is the ontological force which has us conceive of nature in such a way. Heidegger calls this ontological force 'Gestell', often translated as 'Enframing'. Enframing is the technological understanding of being as such, a paradigmatic ontological condition which "sets upon" natural entities, providing a framework for their analysis which lies outside of them, breaking them down into their smallest constituent parts and processes. Heidegger calls the process by which this happens 'ordering revealing'. Ordering revealing of nature from the outset is driven by the quest for ever more efficiency and the “demand that it supply energy that can be extracted and stored as such".[235]

Nature in the process is reduced to what Heidegger calls "standing reserve", an assemblage of intrinsically meaningless knobs of matter and functions standing by for endless optimisation: "Everywhere everything is ordered to stand by, to be immediately at hand, indeed to stand there just so that it may be on call for a further ordering. Whatever is ordered about in this way has its own standing. We call it the standing-reserve".[236] The essence of modern technology thus translates into the quest to seek more and more flexibility and efficiency simply for its own sake.[237] “This setting-upon that challenges forth the energies of nature is an expediting […] yet that expediting is always itself directed from the beginning . . . towards driving on to the maximum yield at the minimum expense”.[238]

The Heideggerian ontological landscape of technology can thus be tentatively organised in four concepts, which I will use throughout this chapter. First, the concept of Enframing. Through the concept of Enframing, Heidegger leads us through three additional key terms in his understanding of technology: ‘Ordering revealing’, which is the active, "challenging", component of Enframing. It has

[234] Heidegger, M. (1977), p. 4.
[235] Ibid, p. 14.
[236] Ibid, p. 13.
[237] Dreyfus, H. (2009), p. 27.
[238] Heidegger, M. (1977), p. 5.

us conceive of nature in terms of cause-effect coherence. ‘Calculative thinking’, which is the cognitive faculty of human beings that allows us to conceive of nature in accordance with the technological understanding of being. We may understand it as the part of human nature which is called upon by Enframing. Finally, the term ‘standing reserve’. Standing reserve is the way in which beings come to be perceived by us once they have been "enframed", namely no longer even as an object but only as a set of functions within a frame of instrumental ends external to them.

For Heidegger, the roots of this revolution, whereby being has come to be understood along these lines, lie in modern philosophy: "What is the ground that enabled modern technology to discover and set free new energies in nature?", Heidegger asks, and responds: "This is due to a revolution in leading concepts which has been going on for the past several centuries, and by which man is placed in a different world … This radical revolution in outlook has come about in modern philosophy. From this arises a completely new relation of man to the world and his place in it … This relation of man to the world as such, in principle a technical one, developed in the seventeenth century first and only in Europe. It long remained unknown in other continents, and it was altogether alien to former ages and histories. The power concealed in modern technology determines the relation of man to that which exists".[239]

The essence of technology, it turns out, is the paradigmatic ontological condition of modernity which governs (Western) humanity's relation to existence (being) and everything that exists (beings) and it is only against that ontological background that, for Heidegger, the functioning of both modern science and modern technology can become intelligible. As I will discuss below, Heidegger is convinced that for as long as we remain oblivious to this metaphysical essence of technology we also remain oblivious to its dangers and misconceive of it as something that we can get under our control. If technology is understood as a means to an end that implies that we are in charge, that we are its masters. However, as we have seen above, we are by no means the masters of the *essence* of technology. Rather, in Heidegger’s terminology, we are the ones being spoken to, being challenged by Enframing. Enframing determines what we perceive as real and it thus controls us, not the other way around, at least for as long as we do not open up to this fact and actively challenge the way we relate to being.

The technological understanding of being may have come about in a philosophical revolution, but the very fact that being allows for the possibility to be revealed to humanity in the associated technological manner, Heidegger claims, can itself not be the result of a philosophical revolution in concepts: "Only to the extent that man for his part is already challenged to exploit the energies of nature can this ordering revealing happen".[240] In order to understand what he means by that we need to briefly return to the shift in Heidegger’s thinking, i.e. to ‘The Turn’ that took place throughout the 1930s and 1940s and which separates the early Heidegger of *Being and Time* from the

[239] Heidegger, M. (1966), p. 50.
[240] Heidegger, M. (1977), p. 8.

late Heidegger of the ‘Questioning Concerning Technology’. This shift in Heidegger’s thought is associated with a complete departure from subjectivism, that still characterised *Being and Time*. As Dreyfus and Spinosa (1997), Loscerbo (1981), Wolin (1990) and many others have pointed out, Heidegger’s understanding of technology has changed with this major shift in his metaphysics, or perhaps it was even his changed understanding of technology that led to the Turn’.[241] In *Being and Time*, in any case, Heidegger still considered modern technology to be the expression of the Cartesian subject’s will to mastery.[242] As Dreyfus and Spinosa (1997) point out, in as late as 1940 he wrote that “Man is what lies at the bottom of all beings; and that is, in modern terms, at the bottom of all objectification and representability”.[243] This suggests that, at this point, he still located the problem of technology in humanity, in the modern subjects’ desire to objectify, exploit and dominate all other beings for their own satisfaction.[244] Only in his later works, he came to think about technology along the lines sketched out above, namely as part of the ‘history of Being’, as an epoch in the understanding of being.[245] The notion of ‘Enframing’ is indicative of what has changed. No longer is it humanity that objectifies but it is being that objectifies itself through humanity (Being) and ultimately humanity itself. Humans are recipients of how being reveals itself to and through them: “Who accomplishes the challenging setting-upon through which what we call the real is revealed as standing-reserve? Obviously, man. To what extent is man capable of such a revealing? Man can indeed conceive, fashion, and carry through this or that in one way or another. But man does not have control over un-concealment itself, in which at any given time the real shows itself or withdraws“.[246] In other words, the later Heidegger’s perspective on technology as a stage in the history of being meant that he was concerned that humanity was relentlessly being dragged into an ever more technologically determined world without being able to resist or shape this course. Superficially, this position resonates with Bostrom’s technological completion conjecture. But where existential risk researchers retain a certain level optimism, Heidegger was deeply fatalistic, focusing almost exclusively on the dangers he associated with this process.

2.3 The dangers

Heidegger saw, broadly speaking, two categories of danger in modern technology. The first category entails conventional, physical technological threats to the environment and to people - i.e. potentially harmful environmental and social impacts of technologies such as ecological destruction, nuclear pollution, unemployment or the increasing destructiveness of modern weaponry. The second

[241] Borgmann (2005) argues that it was his thinking about technology that ushered in the shift in his wider philosophy.

[242] Cf. Wolin, R. (1990).

[243] Heidegger as cited in Dreyfus, H. & Spinosa, C. (1997), p. 162.

[244] Ibid, p. 160.

[245] See Rorty, R. (1999), pp. 68-69, as well as Dreyfus, H. & Spinosa, C. (1997), p. 163.

[246] Heidegger, M. (1977), p. 18.

category of danger has its roots in metaphysics. It encompasses dangers originating from the transformation process which technology forces upon our understanding of being and the consequences of this process for our relationship to the world which surrounds us and our actions within it. Heidegger was more concerned about dangers of this second type than about dangers of the first type arguing that "the threat to man does not come in the first instance from the potentially lethal machines and apparatus of technology […] The rule of Enframing threatens man with the possibility that it could be denied to him to enter into a more original revealing and hence to experience the call of a more primal truth".[247] And at a different point: "The danger consists in the threat that assaults man's nature in his relation to being itself, and not in accidental perils".[248] To put it Dreyfus' (1993) words, Heidegger appears to have been less concerned with the physical havoc technology can wreak, than with the "devastation that would result if technology solved all our problems".[249]

According to Heidegger the danger associated with modern technology "attests itself to us in two ways": [250]

> "As soon as what is unconcealed no longer concerns man even as an object but does so, rather, exclusively as standing reserve, and man in the midst of objectlessness is nothing but the orderer of the standing-reserve, then he comes to the very brink of a precipitous fall: that is, he comes to the point where he himself will have to be taken as standing-reserve".[251]

In the following, I will elaborate on both these stages of what Heidegger considers the dangerous tendency inherent to modern technology. Taken together they provide a promising entry point for a discussion of what it might mean to *be* at existential risk or under 'existential pressure'. Before I do so, however, two clarifying remarks need to be made.

First, it is important to understand that these two stages should be understood as successive steps within a 'dangerous' process, which is that all beings will eventually turn into standing reserve, resulting in what Heidegger calls 'the oblivion of Being'.[252] To begin with, humanity's relation to the world is lastingly impoverished and constricted because we begin to perceive of nature only in functional ways and thereby to mistake the merely correct for the true. [253] Further, in mistaking the correct for the true, we are misled to conceive of ourselves as masters of the universe, not noticing that the logic of instrumentality has no Archimedean point of mastery but ultimately has a nihilistic,

[247] Ibid, p. 28.

[248] Heidegger, M. (2009), p. 115.

[249] Dreyfus, H. (2009), p. 26.

[250] Heidegger, M. (1977), p. 13.

[251] Ibid, pp. 13-14.

[252] As mentioned earlier, Being denotes the mode of humanity's being-in-the-world. Since humanity is part of 'beings', the turning into standing reserve of all beings will eventually encompass humanity and therefore amount to the 'Oblivion of Being'.

[253] Heidegger is critical towards the subject-object distinction to begin with. However, he concedes that conceiving of things as objects does not necessarily mean that one abstracts from their essence. In ordering revealing, however, we ultimately reach a stage where, according to Heidegger, we do not even conceive of things as objects but rather reduce them to processes and functions thus rendering the very notion of essence anachronistic.

inherently forward leaping dynamic that is poised to turn everything into standing reserve, including ultimately ourselves.

The second preliminary remark pertains to the position of Heidegger's work within long-standing debates surrounding inherently self-contradictory characteristics of enlightenment philosophy in so far as it is understood as the philosophy of human emancipation. Heidegger's criticism of ordering revealing and 'calculative thought' certainly can be located in a continuum of romantic, anti-modernist, anti-enlightenment philosophy that has accompanied modern philosophy, and science and technology from their very beginnings. Heidegger's concept of the Gestell, i.e. Enframing, in both wording and meaning, for instance, is clearly reminiscent of Max Weber's ‘iron cage’ - the straightjacket of necessities, of standardisation, uniformisation, rationalisation and bureaucratisation Weber feared was being erected around private and public life under the ascetic spirit of efficiency.[254] Similarly, Heidegger's fears surrounding the oblivion of meaning, resulting from the loss of a deeper, non-instrumental relation to being clearly echoes old fears of the "disenchantment of the world" characteristic for romantic philosophy and literature since at least Rousseau.[255] It is no coincidence that Heidegger, particularly in his later works, frequently cites romantic German poetry, in particular Hölderlin and Rilke.[256] Finally, if one were to embark on the quest to extract some form of final take-away point from his philosophy of technology, it certainly would not be wholly mistaken to quote the arguably most influential publication of Frankfurt School critical theory for that purpose, i.e. Horkheimer and Adorno’s 1944 work the *Dialectic of the Enlightenment*. Enlightenment, they claim, “understood in the widest sense as the advance of thought, has always aimed at liberating human beings from fear and installing them as masters. Yet the wholly enlightened earth is radiant with triumphant calamity”.[257] The underlying idea that there is a tendency in the enlightenment project to undermine its own cause, clearly can be identified as a central motif in Heidegger’s philosophy of technology, although, of course, Heidegger’s critique goes beyond the Enlightenment, addressing the entire Western philosophical tradition since Plato.[258]

However, what makes Heidegger particularly interesting for the purposes of this thesis is that he puts modern technology, specifically the everyday perception and understanding of technology, at the centre of his reflections about the modern human condition. This is what makes his thought a particularly interesting starting point for a discussion of existential risk. As we will see below, existential risk theory hinges on exactly the kind of understanding of technology, both in the way in which it conceptualises technology itself, as well as in the way in which it understands technology’s relationship to humanity, which Heidegger criticises.

[254] For a comparison of Weber's concept of the iron cage of rationality and Heidegger's concept of Enframing see Bambach, C. (2003).
[255] Featherstone, J. (1978).
[256] See specifically Heidegger, M. (2009, 2012).
[257] Horkheimer, M. & Adorno, T. (2002), p. 1.
[258] Rorty, R. (2005), p. 275.

The first stage – ‘the levelling of every ordo’

As discussed above, for Heidegger the essence of modern technology lies in Enframing. Enframing is the defining ontological force of modernity, which challenges man to conceive of the world in an ordering way, as an "object open to the attacks of calculative thought".[259] For Heidegger modern technology, in form of machines and know-how, is merely an ‘outgrowth’ of this challenging essence of technology, and so is modern science.[260] Heidegger refers to the form of revealing, which is challenged forth by Enframing, as an 'ordering' revealing’. The first metaphysical danger Heidegger saw was that with accelerating technological and scientific progress, humanity might become increasingly unable to see the world in non-calculative ways and as something other than stockpiles of resources and exploitable processes: "The coming to presence of technology threatens revealing, threatens it with the possibility that all revealing will be consumed in ordering […] ".[261] The danger thus is that we mistake ordering revealing for the sole truth, so that everything "exhibits itself only in the light of a cause-effect coherence",[262] and to the effect that we will eventually not only forget the deeper essence of being but might even reach a state at which we have forgotten that we have forgotten it.[263] Heidegger's phenomenological distinction between the correct and the true implies that the correct, albeit not being false, captures only a partial version of the truth because it does not uncover a thing in its essence, which can never be fully grasped. In that phenomenological understanding, ordering revealing thus reveals correct but only partial information about the world. It reveals instrumental truths and the danger is that "through these successes [of technological and scientific progress] … in the midst of all that is correct the true will withdraw".[264] Heidegger thus was concerned that Enframing might eventually come to entirely supplant "original revealing and … the call of a more primal truth".[265]

For Heidegger this is not only a dangerous tendency because it threatens to permanently impoverish modern humanity's relation to being, rendering it purely superficial and functional. For Heidegger the technological understanding of being is above all dangerous because it threatens to supplant the very *call* of a more ‘primal truth’, which is to say that he is concerned that we might forget that there is an independent reality of things as things. It is Heidegger’s conviction that one can never uncover a thing in its essence, i.e. understand its ‘truth’ in full, because it always only reveals itself partially to the human mind. Heidegger calls this the ‘concealing unconcealing’ property of revealing.[266] Whenever we encounter beings, they both show themselves (they are

[259] Heidegger, M. (1966), p.50.
[260] Heidegger, M. (1977), p. 116.
[261] Ibid, p.18.
[262] Ibid, p.13.
[263] Heidegger, M. (2009), p. xv.
[264] Heidegger, M. (1977), p.13.
[265] Ibid.
[266] For an in-depth discussion of the ‘concealing unconcealing’ property of revealing, see also Heidegger’s 1935/36 essay on ‘The Origins of the Work of Art’. Viz. Heidegger, M. (2002).

unconcealed) and simultaneously other properties of them withdraw (remain concealed) because the properties that are 'unconcealed' to us depend on our own a priori faculties (e.g. culturally imbued models by which we seek to understand reality, as well as momentary intentions), thus leaving other properties of the beings in question concealed. Expressed in analytical terminology, we might understand this position as expressive of 'epistemological humility' and a variant of 'metaphysical pluralism'.[267] When Heidegger speaks of 'things as things', he thus seeks to invoke a sense of wholeness, which can be posited when we speak of 'hammer' or 'tree', a wholeness that can be intuitively grasped or sensed and that has its own phenomenological reality, but that can never be fully explained or uncovered in scientific terms. Heidegger calls the awareness of this multifaceted nature of truth 'openness to the mystery'. The mystery denotes all the different fields of intelligibility which necessarily remain concealed to us in any given instance of revealing because we can always only occupy one such field at a time.[268] 'The call of a more primal truth' is hence the call for an awareness of the mystery - a recognition of the independent reality of things as things and the fact that aspects of their reality necessarily remain concealed due to our inherently conditioned ways of sense-making. Dreyfus (1991) illustrates this idea based on a comparison between how the Greeks, the Christians and the moderns encounter objects and people, demonstrating how the understanding of being has changed over time:

> "The Greeks encountered things in their beauty and power, and people as poets, statesmen and heroes; the Christians encountered creatures to be catalogued and used appropriately and people as saints and sinners; and we moderns encounter objects to be controlled and organized by subjects in order to satisfy their desires. Or, most recently as we enter the final stage of technology, we experience everything including ourselves as resources to be enhanced, transformed, and ordered simply for the sake of greater and greater efficiency".[269]

As part of the standing reserve entities are no longer represented as things, not even as objects, but are defined solely with reference to their place and function within a wider system of functional dependencies.[270] Their presence becomes fixed and 'unfree', as Heidegger calls it. Seen that way, the technological understanding of being, for Heidegger, involves an act of metaphysical violence – violence against the ontological integrity of things as things - which poses a danger not only to our ability to lead meaningful lives but to the freedom and autonomy (the 'free essences') of all things, including us, because it ultimately translates into acts based on such a reductionist, constricted worldview. First, things are reduced into functions and constituent parts on a metaphysical level and then on a physical level. To Heidegger the worldly destruction which modern technology can cause thus is merely a confirmation of the intrinsically violent nature of the essence of technology he

[267] See Cooper, E. (1997) for a discussion of Heidegger's epistemological humility and Dupré, J. (2001) for a discussion of metaphysical pluralism.
[268] Wheeler, M. (2011).
[269] Dreyfus, H. (1991), p. 338.
[270] Viz. Dreyfus, H. (1993).

believed to have uncovered. The nuclear bomb with its physical capacity to obliterate all things is merely the logical outgrowth of the prior metaphysical obliteration of all things as things: "Science's knowledge, which is compelling within its own sphere, the sphere of objects, already had annihilated things as things long before the atom bomb exploded. The bomb's explosion is only the grossest of all gross confirmations of the long since accomplished annihilation of the thing: the confirmation that the thing as a thing remains nil".[271] Heidegger also refers to factory farming, the use of the term 'human resources', and modern power plants as illustrations for how the metaphysical violence associated with ordering revealing comes to be reflected in real life. In a particularly callous example he even states that modern agricultural techniques, the Shoah and nuclear weapons are *wesensgleich*, i.e. essentially the same thing: "Agriculture is now the mechanized food industry, essentially the same as the manufacturing of corpses in gas chambers and extermination camps, the same as the blockade and starvation of nations, the same as the production of hydrogen bombs".[272] In Enframing things are metaphysically deprived of their own intrinsic reality and transformed into objectless heaps of functions, opening the door wide for their abuse along these lines. For Heidegger this danger is particularly great since the destructive essence of modern technology and science is kept concealed behind the impression that ordering revealing, scientific and technological knowledge, actually bring order and structure into the world, when in fact it is destroying the conditions upon which any form of normativity could arise: "What threatens man in his very nature is the view that technological production puts the world in order, while in fact this ordering is precisely what levels every *ordo*, every rank, down to the uniformity of production, and thus from the outset destroys the realm from which any rank and recognition could possibly arise".[273]

Obviously, Heidegger here implicitly distinguishes between two different levels of order, a distinction which again rests on his phenomenological distinction between the correct and the true. On the one hand science *correctly* reveals an order in nature, the order of cause and effect, natural laws, cosmic forces, of matter, energy, time, space, etc. On the other hand, however, it thereby overwrites an order, namely the order of how beings 'naturally' reveal themselves to our consciousness, i.e. as things and objects in their own right, which in Heidegger's view also speaks to the truth, a "primordial truth". For Heidegger this clash of modes of revealing, and thus the ontological effect of technology, is of enormous normative relevance because, he suggests, we ultimately rely on our 'natural' understanding of beings in order to be able to 'rank' things, to think normatively, at all.

This, however, and ironically perhaps, appears to be true for ordering revealing and calculative thinking too. If nothing were to be simply taken as a given, arbitrarily posited to be intrinsically valuable in and of itself, as it appears to our consciousness, calculative thinking would

[271] See Heidegger, M. (2009), p.168.
[272] Heidegger, M. (1949), as cited in Petrina, S. (2017), p. 16.
[273] Heidegger, M. (2009), p. 114.

itself lack any orientation for its inquiry. “Calculative thinking”, Heidegger argues, “always reckon(s) with conditions that are given".[274] Yet, calculative thinking transcends all conditions that are given and, as ordering revealing, turns them into processes that can be controlled at will - it turns givens into matters of choice, that is its very purpose. Thought to its end, this suggests that calculative thinking must lead into a world where the conditions from which ‘any order and recognition’ could arise, even for calculative purposes, have been transcended. Heidegger describes this vividly, in a strikingly prescient passage where he muses about the ultimate possibilities of global technological and economical interconnectedness:

> “When the farthest corner of the globe has been conquered technologically and can be exploited economically; when any incident you like, in any place you like, at any time you like, becomes accessible as fast as you like; when you can simultaneously 'experience' an assassination attempt against a king in France and a symphony concert in Tokyo; when time is nothing but speed, instantaneity, and simultaneity, and time as history has vanished from all Dasein [...] then there still looms like a spectre over all this uproar the question: what for? - where to? and what then?” [275]

In other words, if technological mastery were to unfold in full, Being would lose its place in space and time, Being would lose its spatio-temporal structure. But how could one begin to make sense of an existence without this structure? Enframing, according to Heidegger, in turning the most basic conditions of our existence into negotiable state of affairs, is supplanting our natural way of Being, of being-in-the-world, and thus threatens to undermine the conditions under which we can think normatively at all.

Heidegger thereby connects two fields of knowledge that are often considered separate: normative knowledge and scientific or descriptive knowledge. He is suggesting that technology, or rather the ontological effect of the technological understanding of being, which he calls ‘uprooting’, is not distinct from normative knowledge but transcends that distinction because it undermines our natural, ordinary understanding of beings on which we tacitly rely in order to bring order into the world, to orientate us in it, including instrumentally, and attribute value and meaning to anything at all. When Wittgenstein describes in his 1929 *Lecture on Ethics* how, seen scientifically, a murder is just another fact, that “the murder will be on exactly the same level as any other event, for instance the falling of a stone”,[276] Heidegger would presumably respond that, therefore, if the technological understanding of being were to dominate our thinking to the fullest extent, murder would indeed at some point no longer necessarily be regarded as something wrong, granted it would serve some, no matter how arbitrary, instrumental purpose. This indeed is implicit to his above quoted remark about the ‘manufacturing of corpses’ as an illustration for the material effects of Enframing.

[274] Heidegger, M. (1966), p. 46.
[275] Heidegger, M. (2000b), p. 40.
[276] Wittgenstein, L. (1965), p. 6.

Heidegger's philosophy would thus suggest that the distinction between scientific and normative knowledge is not as straightforward as intuition might suggest and that in reducing our perception of reality from things to manipulatable processes and functions, modern science and technology ultimately undermine the very possibility of normativity. Progressing technical order on the first, the ontological level, destroys the possibility of order on the second, the normative level. The fact-value dichotomy obviously belongs to the most controversial and most widely discussed topics in ethics and philosophy of science and I therefore cannot hope to provide a comprehensive discussion of Heidegger's position within it. For him, however, the attempt to draw that distinction itself, would have to be considered as expressive of the (Cartesian) technological understanding of being in its own right.[277]

The second stage - humanity as standing-reserve

As indicated above, the second stage of the technological danger is that the challenging, ordering force of Enframing will eventually turn onto humankind itself, transforming it into standing reserve as well: "everything, including man himself, becomes material for a process of self-assertive production, self-assertive imposition of human will on things regardless of their own essential natures”. As it stands, the human being is the last and ultimate given, from which ordering revealing takes its course. However, Heidegger believes that Enframing "threatens to sweep man away into ordering as the supposed single way of revealing, and so thrusts man into the danger of the surrender of his free essence".[278] If that were to happen, Heidegger argues "man would have denied and thrown away his own special nature—that he is a meditative being. Therefore, the issue is the saving of man's essential nature".[279]

In Heidegger's view, the technological understanding of being is inherently devious because it lures humankind into conceiving of itself as master of the universe when in fact humankind is *challenged* and *ordered* to conceive of nature that way and thus has no leverage over technological unconcealment as such: "If man is challenged, ordered, to do this, then does not man himself belong even more originally than nature within the standing-reserve? The current talk about human resources, about the supply of patients for a clinic, gives evidence of this".[280] Technological and scientific progress in Heidegger’s view is an inherently enslaving process that does not come to a halt for the sake of humanity. The logic of calculation and instrumentality itself is something that is not only conceptually independent from truly human purposes but runs counter to them, because of its metaphysically nihilistic nature but also because its essence is the will to attain control over nature as such, to order it, to make it calculable and predictable, and since humanity is itself still part of nature

[277] Hall, H. (1993), p. 129.
[278] Heidegger, M. (2009), p. xv.
[279] Ibid, p. 55.
[280] Ibid, p. 8.

there is no reason to assume that it will come to a halt for the sake of humanity's free essence. To put it in a nutshell, if humanity keeps believing in the idea of mastery and the associated idea that it will eventually attain mastery over technological mastery, it keeps acting as the 'orderer of the standing reserve', and is destined to turn itself into an instrument eventually because the logic of mastery itself has no Archimedian point: "He [‘man’] himself and his things are thereby exposed to the growing danger of turning into mere material and into a function of objectification. The design of self-assertion itself extends the realm of the danger that man will lose his selfhood to unconditional production".[281]

The irony is that the idea of technological mastery depends on the willing of the 'self', thus presupposing some essential ‘whatness’ of the individual, though its logic annihilates the idea of selfhood from the outset by reducing reality to causal processes, having no way to accommodate the very idea of ‘whatness’. The devious aspect of the process of ordering revealing thus is that humankind, as orderers of the standing-reserve, considers itself master of an undertaking that does not know any masters.

C. S. Lewis, a contemporary of Heidegger, arrived at almost the exact same conclusion about the idea of scientific and technological progress:

> "Each new power won by man is a power over man as well. Each advance leaves him weaker as well as stronger. In every victory, besides being the general who triumphs, he is also the prisoner who follows the triumphal car. I am not yet considering whether the total result of such ambivalent victories is a good thing or a bad. I am only making clear what Man's conquest of Nature really means and especially that final stage in the conquest, which, perhaps, is not far off. The final stage is come when Man […] has obtained full control over himself. Human nature will be the last part of Nature to surrender to Man. The battle will then be won […] But who, precisely, will have won it"?[282]

Heidegger would presumably respond that 'Enframing' has ‘won it’ - i.e. efficiency and instrumentality for their own sake. Humanity, or 'Man', will not have won, because ‘he’ can only 'win' by changing who or what he is.

Thomson (2015) argues that one can distinguish between three phases of the technological understanding of being in Heidegger’s ontological historicity, the early-modern, the late-modern and the post-modern. [283] According to Thomson, the early-modern technological understanding of being was epitomised by thinkers such as Bacon and Descartes, who understood nature in terms of objects of mastery, that could be controlled for human purposes by means of scientific inquiry, or ‘natural philosophy’ as it was called at the time. This phase hence was firmly rooted in an anthropocentric worldview, where nature, or, in Heideggerian language, *beings* were objectified. However, beings

[281] Heidegger, M. (2009), p. 113.

[282] Lewis, C.S. (1943/2013), pp. 41–42.

[283] See also Loscerbo, J. (1981), p. 138; or Pöggeler, O. (1994), pp. 135 - 144, for highly instructive discussions of Heidegger’s understanding of the different epochs of the technological understanding of being.

still retained their object-character and therefore some form of independent, albeit perhaps oversimplified, *given* reality as things. The late-modern phase then saw beings disintegrate, represented as meaningless stuff, standing-reserve as it were, where they were stripped even of their object-character, transformed into processes and particles that could be manipulated at will for human needs. We might visualise this as the world of Darwinism, of industrialisation, accompanied as it was by the compartmentalisation of production processes, the introduction of disposable consumer products and the advent of elementary physics. Here the worldview was still anthropocentric but non-human beings, humanmade or natural, were beginning to be represented no longer even as objects but purely in terms of objectless processes and as heaps of functions. The post-modern stage, finally, is characterised by a complete departure from both the anthropocentric and the object-character of the early-modern phase of the technological understanding of being. In the post-modern phase, to phrase it along C.S. Lewis' analogous lines of reasoning, human nature, too, is beginning to yield and to surrender to 'Man's' conquest of nature, which means that even "the ultimate springs of human action are no longer […] something given", [284] but are beginning to be understood in terms of manipulatable processes as well. At this point one can no longer meaningfully speak of anthropocentrism because that which used to function as the 'centre' of the 'conquest of nature' has itself been revealed as a matter of technological manipulability. Hence Lewis' question who, once that point is reached, 'will have won'? In Heidegger's view Enframing would have won - a world of literally 'no-thingness', of directionless forces, processes and particles that are organised and reorganised in an endless process of challenging-forth, i.e. of reconstitution and becoming for no clear purpose at all.[285] Such is Heidegger's vision of 'technology as destiny'. In the following chapters I will come back to it as it serves as an instructive counterpoint for thinking about existential risk theory and its aim to reach 'technological maturity'.

2.4 Modern technology and ordering revealing

According to Heidegger, within this process of ordering revealing, modern technology in form of physical machinery, technical systems, gadgedry etc., plays an important role. On the one hand it is an embodiment of ordering revealing, it gives physical form to the instrumental understanding of being and allows us to dominate and exploit nature. It is the physical embodiment of Enframing. However, it also by itself contributes to the establishment of a reductionist, instrumentalist understanding of being. Heidegger illustrates that property of technology by describing the effect which the presence of a hydroelectric power plant in the Rhine River has on his perception of the river:

[284] Lewis, C.S. (1943/2013), p. 41.

[285] Thomson, I. (2015), p. 9.

> "In the context of the interlocking processes pertaining to the orderly disposition of electrical energy, even the Rhine itself appears as something at our command. The hydroelectric plant is not built into the Rhine River as was the old wooden bridge that joined bank with bank for hundreds of years. Rather the river is dammed up into the power plant. What the river is now, namely a water power supplier, derives from the essence of the power station".[286]

This example also serves well as an illustration of Heidegger's understanding of the notion of standing reserve. As discussed above, as part of the standing reserve a thing is no longer perceived in terms of its own 'essential nature', or even as an object ("Whatever stands by in the sense of standing-reserve no longer stands over against us as object" [287]) but only in terms of its instrumental value for productive purposes. The river, for example, in Heidegger's view, once dammed up in the plant does no longer reveal itself to the human eye on its own terms. Rather the free essence of the river withdraws and remains hidden behind its function as part of something that lies outside of it, namely the wider system of energy generation. This is the 'setting upon', which Heidegger describes as the fundamental characteristic of the technological understanding of being, whereby nature is reduced to a standing reserve. At a different point Heidegger again illustrates this mechanism by discussing whether an airliner should be regarded as an object:

> "an airliner that stands on the runway surely is an object. Certainly. We can represent the machine so. But then it conceals itself as to what and how it is. Revealed, it stands on the taxi strip only as standing-reserve, inasmuch as it is ordered to ensure the possibility of transportation. For this it must be in its whole structure and in every one of its constituent parts, on call for duty, i.e. ready for takeoff […] Seen in terms of the standing-reserve, the machine is completely unautonomous, for it has its standing only from the ordering of the orderable".[288]

The machine, in Heidegger's view, derives its presence solely from within the essence of technology, i.e. the technological understanding of being, it has no presence of itself. The airliner's presence is defined through its function within the wider system of transportation and has no reality independent from that. Nature for Heidegger is fundamentally different because, originally, it has an essence and thus a presence independent of the functional value attributed to it, and yet, he argues, ordering revealing and its expression through technology is on its way to reduce our understanding of nature to a technical understanding as well, i.e. we are beginning to conceive of the river in a similar manner as of the airliner. As part of the standing-reserve the river is turned, metaphysically, into a machine.

To some extent this ontological effect might come across as an old-fashioned, conservative romantic concern about the disenchantment of reality. However, for Heidegger it is inherently violent. What he wants to point out, is that ordering revealing has an effect on what we conceive of as real and thus ultimately also on how we come to interact with our surroundings. Technology shapes our "understanding of what counts as things, what counts as human beings, and ultimately

[286] Heidegger, M. (1977), p. 8.
[287] Ibid.
[288] Ibid.

what counts as real, on the basis of which we can direct our actions toward particular things and people" as Dreyfus (2009) puts it.[289] The more we are surrounded by machines such as hydroelectric dams, by airplanes, by cars, etc., any type of object which can only be understood as part of and in itself made up of a wider web of functional relationships that lie outside of ourselves, the more we are inclined to think in such terms and the less we are able to meet things or beings on their own grounds.

Nuclear bombs and extermination camps, for Heidegger, were the most gruesome embodiments of this background understanding. However, modern technology in its material form not only embodies and enforces the technological world view, it also by itself continuously creates the conditions for the further extension of standing reserve and this, for the purposes of this thesis, is perhaps the most crucial aspect of Heidegger's philosophy of technology. Technology creates the conditions for a continuous extension of the standing reserve in a multitude of ways. First, new technological inventions typically open up routes for ever more technological inventions. Second, new technologies often open up novel path-ways and opportunities for theoretical scientific enquiry into nature (consider for instance the microscope, X-rays, the telescope, satellites etc., all of which facilitated scientific inquiry in uncountable ways), which in turn pave the way for new technological inventions and so forth. Third, and arguably most importantly for the purposes of this thesis, technology ushers in the ordering of ever more aspects of life precisely by potentially and actually leading to negative consequences, i.e. by entailing conventional physical dangers.

These negative, unintended or potential consequences, in Heidegger's view, lead to a further extension of standing reserve for as long as we do not challenge the essence of technology, which is the technological understanding of being. The reason is that for as long as we do not challenge that understanding we remain within the same metaphysical mindset, committed to the idea of attaining mastery over nature and thus, whenever technology leads to bad results, we are not led to question the undertaking as such but rather if and how we can master it better. We thus try to include and control ever more variables. Precisely by threatening us in a conventional sense, technology creates the needs for the ever further extension of its own logic.

2.5 The instrumental conception of technology

For Heidegger, the common understanding of technology plays a particularly important role in that self-reinforcing dynamic. As argued in Section 1 of this chapter, the starting point of Heidegger's discussion of the issue of technology is the "instrumental and anthropological definition" of technology, whereby technology is understood as a means to an end and a human activity. In philosophy of technology this conception of technology has come to be known as the 'neutrality

[289] Dreyfus, H. (2009), p. 27.

thesis', which holds that "technology is a neutral instrument that can be put to good or bad use by its users".[290]

For Heidegger the "instrumental and anthropological conception of technology", at least when applied to modern technology as opposed to ancient tekné, is not only misleading but dangerous. It is misleading because, as we have seen, modern technology is more than simple tool-use. It is based on and reinforces a specific understanding of being. The neutrality thesis thus misleads us in that it abstracts from the, in Heidegger's view, most important property of modern technology, its inherently violent metaphysical essence. However, for him the neutral conception of technology is even more problematic than that because it "delivers us over" to technology, where 'delivering us over' conceivably needs to be understood as 'clearing the path’ for the turning of humanity into standing-reserve:

> "Everywhere we remain unfree and chained to technology, whether we passionately affirm or deny it. But we are delivered over to it in the worst possible way when we regard it as something neutral; for this conception of it, to which today we particularly like to do homage, makes us utterly blind to the essence of technology."[291]

Now, there appear to be at least two reasons for that supposed effect of the neutrality thesis. First, as said above, by abstracting from the metaphysical essence of technology the neutrality thesis makes us blind to the nihilistic nature of technology and leads us to uncritically accept the world it creates. Secondly, the neutrality thesis delivers us over to technology because it implicitly reinforces the idea of mastery, i.e. the belief that we, as humanity, are in control of technology when in reality, as discussed, the logic of technology has no Archimedean point of mastery: "So long as we represent technology as an instrument, we remain held fast in the will to master it. We press on past the essence of technology".[292] Conceiving of technology as a mere tool implies that humanity is in control of it. The neutrality thesis hence fosters the idea that whenever we encounter a technological problem, we merely need to get better at using and designing the tool, to use technology in more efficient and more ingenious ways: "The instrumental conception of technology conditions every attempt to bring man into the right relation with technology. Everything depends on our manipulating technology in a proper manner as a means. We will, as we say, […] master it. The will to mastery will become all the more urgent the more technology threatens to slip human control".[293] In thinking about technology as a neutral tool we therefore pay heed to the will to mastery and inadvertently contribute to the ever further extension of standing reserve. Heidegger argues that all behaviour that is based on a neutral understanding of technology and the associated idea that we need to master the problems we are facing by becoming better at using tools leaves the underlying logic of mastery

[290] Franssen, M., Lokhorst, G., et al. (2018).
[291] Heidegger, M. (1977), p. 35.
[292] Ibid, p. 17.
[293] Ibid, p. 5.

untouched and thus is technological behaviour. Heidegger's fear hence was that, in times of growing conventional dangers resulting from accelerating technological progress this uncritical conception of technology could, rather than leading to ontological humbleness and a criticism of the will to mastery, translate into an ever more pronounced will for mastery, to get the world under technological control, and that we might become entirely forgetful about what we are doing to ourselves in the process:

> "The decisive question of science and technology today is no longer: Where do we find sufficient quantities of fuel? The decisive question now runs: In what way can we tame and direct the unimaginably vast amounts of atomic energies, and so secure mankind against the danger that these gigantic energies suddenly — even without military actions—break out somewhere, "run away" and destroy everything"?[294]

In other words, he was afraid that the conventional dangers associated with technology would become a matter of such urgency that they would lead humanity to focus all energies on futile attempts to get these dangers under control, to master the technologies it has unleashed on the world, whilst wholly ignoring the deeper but no less real dangers that go along with that undertaking. Against the background of what we are now calling existential risk, Heidegger thus saw the real possibility that our quest for mastery would lead us into a situation where, under the imperative of control, calculative thinking would finally oust all other modes of thinking:

> "the approaching tide of technological revolution in the atomic age could so captivate, bewitch, dazzle, and beguile man that calculative thinking may someday come to be accepted and practiced as the only way of thinking […] Then there might go hand in hand with the greatest ingenuity in calculative planning and inventing […] total thoughtlessness […] then man would have denied and thrown away his own special nature".[295]

We can thus distinguish between three concepts with which Heidegger denotes what he saw as the three interlinked features of the dangerous process he saw unfolding under the reign of Enframing: standing reserve (the ordering of things for purposeless production), objectlessness (the clash between scientific knowledge and the manner in which things reveal themselves naturally to our consciousness) and thoughtlessness (the emptiness of calculative thinking).

Thoughtlessness, understood along Heidegger's lines, is the narrowing of our thinking to calculative thinking, i.e. the total absence of meditative thinking. The 'greatest danger' of technological progress, Heidegger says, is that humanity could entirely forget about meditative thinking, and therewith lost its openness to the mystery, which would mean that being forgets itself. [296] ”Objectlessness is the ontological effect of the process, which I have characterised previously as the 'metaphysical violence' inherent to calculative thinking: the disintegration of things as things as

[294] Heidegger, M. (1966), p. 51.

[295] Heidegger, M. (1966), p. 56.

[296] Ibid, pp. 56-57.

the result of the scientific abstraction from their own particular reality as phenomena appearing naturally to our consciousness, from their 'grantedness' as it were, and their dissolution into their constituent elements, processes and functions. [297] Standing reserve is the new order in which the henceforth objectlessness world comes to 'presence', i.e. the mode in which things come reveal themselves under conditions of thoughtlessness and objectlessness, once they have been "challenged forth" by "Enframing", namely as sets of manipulable processes available to humanity's command and defined solely in terms of their instrumental value to a purpose that is independent from them: "Everywhere everything is ordered to stand by, to be immediately at hand, indeed to stand there just so that it may be on call for a further ordering." The motif of standing reserve, is not only intended to circumscribe a distinctly modern mode of mental representation, but seeks to explain actual material effects of thoughtlessness and objectlessness – i.e. how reality including ourselves, i.e. Being, is transformed at the hands of technology. Furthermore the concept of standing-reserve is intimately linked to the idea of mastery and therewith an illusion which Heidegger identifies at the heart of the technological project – the illusion that we can rearrange nature in a manner that suits our needs to the fullest extent, that we can become the masters of nature.

2.6 Existential risk from a Heideggerian perspective

Heidegger's philosophy of technology provides us with a rich historical and conceptual basis for embedding existential risk theory in a wider context of debates appropriate to its earlier distilled substance matter, i.e. in an environment of debates discussing technology in existential terms. Heidegger's account suggests that existential risk research is a rather old-fashioned response to an old fear – the fear of losing control over the consequences of technological progress – by seeking to perfect technological mastery. The reason is, of course, that, from a Heideggerian perspective, existential risk research is firmly rooted in the technological mindset and thus what he calls 'technological behaviour'. All the criteria Heidegger introduces for identifying this mindset are satisfied – most importantly that it thinks about technology in terms of the instrumental and anthropological definition. But, more than that, the fact that the non-domain specific treatment of potential human extinction scenarios elevates technology into the position of arbitrator over life and death and thereby frames the perfection of technological mastery as an undertaking of existential importance ultimately fuels Heidegger's deepest fear - that Enframing might turn onto humanity itself, reducing it to mere 'standing-reserve'. Provocatively put, existential risk theory, can hence be seen as an embodiment of the technological spirit that, according to Heidegger, had taken hold of modern humanity.

Technology in existential risk theory

[297] Heidegger, M. (1977), pp. 13-14.

In the first chapter I have argued that, initially, one might be inclined to think of existential risk theory as presenting a new, critical perspective on technology. Existential risk theory calls for greater vigilance, carefulness and, ultimately, an overhauled approach to regulating and organising technological developments, from a reactive approach to a proactive, precautionary approach, at least in research areas that are deemed particularly sensitive.[298] Olle Häggström of the FHF argues that "the attitude that dominates research and development today, is tantamount to running blindfolded and at full speed into a minefield".[299] Views such as these could certainly be taken as a criticism of what I have referred to as 'Enlightenment optimism' in the introduction , i.e. of the "idea that scientific and technological progress automatically contributes to the advancement of society by bringing about a unification of wealth and virtue".[300] Existential risk research could hence be seen as an awakening of parts of the scientific and technological community to the kind of criticisms that have been voiced by philosophers of technology, critical theorists, environmentalists and authors from many other backgrounds for decades. However, criticism of a particular modus of technological progress and criticism of technology are not the same. Heidegger's philosophy of technology allows us to see the difference between the two more clearly.

When seen against the background of Heidegger's philosophy of technology (or any other strand of philosophy of technology for that matter), the first thing one notices is that existential risk theory actually lacks a theory of technology - it never asks what technology *is*. This is both surprising and unsurprising. It is surprising in so far as technology assumes centre stage in existential risk theory and it would seem only natural to ask what it actually is that one attributes such an important role to. If technology is our destiny should we not 'question concerning technology'? At the same time, it does not come as a surprise because there is a default conception of technology which existential risk theory implicitly builds on. I have argued before that Bostrom for instance understands technology along the lines of making-and-using, as Langdon Winner puts it, i.e. as mere tools, and Bostrom in that regard clearly is representative for the wider field. This becomes apparent for instance in a joint report of the FHI and CSER on 'Unprecedented Technological Risks'. Here, it is argued that "over the next few decades, the continued development of dual-use technologies will provide major benefits to society. They will also pose significant and unprecedented global risks, including risks of new weapons of mass destruction, arms races, or the accidental deaths of billions of people".[301] 'Dual-use technology' is a term which is typically employed in military contexts to indicate that a technology can be used both for peaceful as well as for violent purposes. The term is frequently used in the literature on existential risk, for instance also by Séan ó Heigeartáigh of CSER,

[298] Wiener, J. (2016), Sunstein, C. (2005, 2009), Bostrom, N. (2013), Weitzman, M. (2011), Rees, M. (2004).
[299] Häggström, O. (2016), p. 0.
[300] Mitcham, C. (1994), p. 40.
[301] Beckstead, N. et al. (2014), p. 7.

who states that the “dual-use characteristic—that the underlying science and technology could be applied to both destructive purposes, and peaceful ones—is common to many of the emerging technologies that we are most interested in” and lists as examples “bioscience and bioengineering such as the manipulation and modification of certain viruses and bacteria […] Geoengineering: a suite of proposed large-scale technological interventions that would aim to “engineer” our climate in an effort to slow or even reverse the most severe impacts of climate change […]” and “Advances in artificial intelligence—in particular, those that relate to progress toward artificial general intelligence—AI systems capable of matching or surpassing human intellectual abilities across a broad range of domains and challenges”. In all of these cases, he argues, “progress on these sciences are driven in great part by a recognition of their potential for improving our quality of life, or the role they could play in aiding us to combat existing or emerging global challenges. However, in and of themselves they may also pose large risks”.[302]

The tag ‘dual-use’ appears to be entirely redundant here as it signals only how technology is understood anyway. As Iain Golding reminds us in a recent report to the Government Chief Scientific Advisor, *all* technologies are potentially ‘dual-use’: “Even when the economic incentives and technological breakthroughs allow advancement, they may be ill advised […] as we highlight in our discussion on systemic risks, the potential abuse of these technologies to create new biological pathogens reminds us that all technologies are potentially dual use”.[303] By adding the prefix ‘dual-use’ existential risk researchers hence merely highlight the property to which they reduce technology from the outset, namely that it is a tool that can be put to good or to ill use depending on the users’ (or developers’) know-how and intent, implying that technology’s consequences depend on how we handle it rather than the technology itself.

In other words, existential risk theory is rooted in a very traditional way of thinking about technology, referred to under varying labels in philosophy of technology, such as ‘instrumental theory’,[304] ‘instrumental vision’,[305] or the ‘make-and-use’ paradigm,[306] which, according to most authors in the field, is ‘the traditional liberal view of technology’,[307] and to this day the most widely held, or, as argued earlier, the default conception of technology.[308] Andrew Feenberg for instance argues that it is the default view in most social sciences, from economics, to policy studies or international relations.[309] As we have seen, this conception, or what Heidegger calls the ‘instrumental

[302] O’Heigeartaigh, S. (2017), p. 358; Further examples can be found for instance in Farquhar, S. et al. (2017), or Rees, M. (2008).
[303] Goulding, L. (2014), p. 28.
[304] Feenberg, A. (2002), p. 5.
[305] Franssen, M., Lokhorst, G., et al. (2018).
[306] Winner, L. (2001), p. 26.
[307] Thomson, I. (2009), p. 5.
[308] Böhme, G. (2012), p. 3.
[309] Feenberg, A. (2002), p. 6.

and anthropological definition' - the idea that technology is a means-to-an-end and a human activity - also is the very starting point of the 'Question Concerning Technology'.

We can hence infer that existential risk theory does not present a new take on what we might call the 'problem of technology as such'. That is, existential risk theory does not in principle present us with a new way of thinking about technology, despite the fact that it argues for a fundamentally overhauled approach to regulating technological change. Even if the particular technologies and dangers that are discussed in existential risk theory might be new, the manner in which they are discussed is not. This is also reflected in the fact that the fear underlying existential risk research, that technological developments might get out of control, lead to unpredictable and potentially catastrophic consequences, that "one species - ours - has Earth' s future in its hands and could jeopardise not only itself but also life's immense potential" [310] is not new. Heidegger in 1966 discusses this kind of fear as characteristic for the scientific and technological community of his time, asking on their behalf: "The decisive question now runs: In what way can we tame and direct the unimaginably vast amounts of atomic energies, and so secure mankind against the danger that these gigantic energies suddenly — even without military actions —break out somewhere, "run-away" and destroy everything?" [311] Existential risk research amounts to the realisation that we now face more kinds of these risks on more frontiers, asking exactly the same question in the context of new technologies, such as artificial intelligence, nano-technology, geo-engineering or any other of the unprecedented technological risks it considers. We thus see that existential risk theory treats technology in a rather old-fashioned way and also that its fears are not qualitatively new.

However, I have also discussed above why Heidegger regards the instrumental and anthropological definition of technology as inherently problematic and even dangerous. The problem of the instrumental vision of technology is that it translates directly into what philosophers of technology call 'the neutrality thesis'. Since technology is viewed as a mere means to an end, it is considered to be only contingently related to the substantive values it serves, which means, as Feenberg puts it, that it has no valuative content of its own and hence is normatively neutral.[312] Individual technologies, if seen that way, become in an ethical and political sense *invisible*. If a given technology leads to negative results, these are in consequence attributed to human failure rather than to the technology in question, i.e. either to a lack in skill or to malicious intent. As Heidegger argues, this conception of technology, therefore makes it impossible to think critically about technology.

However, the situation is even more problematic than that. As Franssen et al. (2018) argue, the neutrality thesis ultimately does *not* translate into a neutral perspective on technology (neutral in the sense that technology would be understood to be, on balance, neither good nor bad for humanity, or at least only as good or bad as the uses made of it). Rather, the instrumental view implies, prima

[310] Rees, M. (2008), p. xi.
[311] Heidegger, M. (1966), p. 51.
[312] Feenberg, A. (2002), p. 5.

facie, a positive ethical assessment of technology. [313] The defining quality of technology, when understood purely instrumentally, is that it increases the capabilities of humanity and provides it with new options. This, however, is prima facie considered to be something desirable. Furthermore, technology increases the efficiency with which resources (time, energy, materials, etc.) can be utilised, which is in fact how technology is defined in mainstream economics,[314] which is also generally considered to be desirable. Technology itself, under that conception, hence allows us to do more things, to utilise nature in more ways and with greater efficiency and wherever it does lead to bad outcomes these outcomes are attributed to human failure in design or application, rather than to technology itself. Negative outcomes therefore do not impair the generally positive view of technology as such. Under the assumption that humans utilise the new capabilities technology provides well, the default position hence is that more technology is better than less.

As a result, and we might understand this as the central concern of the *Question Concerning Technology*, Heidegger realised that the instrumental definition of technology, if it is the only way in which technology is understood, makes it not only effectively impossible to assume a critical position toward technology, but that it is emblematic of the technological understanding of being and therefore can only translate into an ever more pronounced quest for technological mastery wherever we are confronted with any type of problem, including technological problems: "So long as we represent technology as an instrument", we look 'past its essence', Heidegger reminds us, and therefore "remain held fast in the will to master it", which means that we end up in a circuit, forever chasing moving targets, "manipulating technology in the proper manner as a means", i.e. manipulating that with which we manipulate and so forth.[315] Heidegger thus calls all behaviour that is based on the instrumental conception of technology 'technological behaviour'.[316]

2.7 Existential risk research as technological behaviour

Against that background, it is abundantly clear that existential risk research, from a Heideggerian perspective, needs to be understood as technological behaviour. Rather than presenting us with a critique of technology, it presents us with a quest for technological mastery. It understands technology in the conventional sense as a tool and an instrument and considers every natural and technological problem from the vantage point of its potential for further technological optimisation. Existential risk research can, from that perspective, be seen as a rather old-fashioned response to an old problem: The fear of losing control over technology and a resultant quest to learn how to

[313] Franssen, M., Lokhorst, G., et al. (2018).
[314] Lawson, T. (2017), p. xi.
[315] Heidegger, M. (1977), p. 17.
[316] Ibid, p. 48.

manipulate technology in a ‘proper’ manner as a means, i.e. to eventually master it fully.[317] Existential risk theory demonstrates what such a quest would involve. It would involve micro-mastery, in form of ‘safely’ designed technological products, and macro-mastery, captured in notions such as ‘differential technological development’, ‘horizon-scanning’, ‘preferred order of arrival’, etc.

Furthermore, existential risk theory, resonates with Heidegger’s observation that the technological understanding of being turns, ontologically speaking, the entire world into a machine, where every natural process is seen from the perspective of making-and-using and, in principle, as something that can be brought under human command: “Meanwhile man, precisely as the one so threatened, exalts himself to the posture of lord of the earth. In this way the impression comes to prevail that everything man encounters exists only insofar as it is his construct. This illusion gives rise in turn to one final delusion: It seems as though man everywhere and always encounters only himself”.[318] This becomes clearly apparent in what I have described in the first chapter as the turning of every risk, natural and anthropogenic, into, in effect, a technological risk. Physicist David Deutsch perfectly encapsulates the underlying mindset when he argues that:

> “before our ancestors learned how to make fire artificially (and many times since then too), people must have died of exposure literally on top of the means of making the fires that would have saved their lives […] In a parochial sense, the weather killed them; but the deeper explanation is lack of knowledge. Many of the hundreds of millions of victims of cholera throughout history must have died within sight of the hearths that could have boiled their drinking water and saved their lives; but, again, they did not know that. Quite generally, the distinction between a “natural” disaster and one brought about by ignorance is parochial. Prior to every natural disaster that people once used to think of as ‘just happening’, or being ordained by gods, we now see many options that the people affected failed to take — or, rather, to create”.[319]

In existential risk theory the distinction between natural and man-made disaster has been entirely blurred. Both are rendered, in effect, as disasters of ignorance. In existential risk theory everywhere, as Heidegger argues, with reference to Heisenberg, we appear ‘to encounter only ourselves’. The purpose of existential risk theory is to prevent humanity from making the same mistake as those of our forefathers who froze to death whilst bedded on combustible material. Its goal is to leave behind ignorance for good, to anticipate potential disasters and technologically utilise the resources nature yields to prepare for such eventualities – from lighting a fire in winter to launching asteroid-deflecting missiles into outer-space, as it were. Heidegger has pre-empted this mindset when he argued that, prima facie, “the world now appears as an object open to the attacks of calculative thought, attacks that nothing is believed able any longer to resist”.[320] From this perspective, everything is possible, nothing is given, and what becomes of humanity depends entirely on how well we design and use the tools science and technology can, in principle, provide us with.

[317] Ibid, p. 5.
[318] Ibid, p. 27.
[319] Deutsch, D. (2011), p. 207.
[320] Heidegger, M. (1966), p. 40.

It is in that vein that Kateb (1997) understands Heidegger to be arguing that at the heart of the technological understanding of being, i.e. of Western humanity's enthusiastic compliance with the dictates of Enframing, its acting as the 'orderer of the orderable', is a 'rebellion' and a "war with given reality", the deepest root of which "is not scarcity, not the failure of nature to make better provision for a necessitous humanity, instead, a Western wilfulness, a will to power, to mastery, an overflow of energy that wants to shake the world to pieces and make it over. The craving is either to put the human stamp on reality or at least to rescue nature from the absence of any honestly detectable stamp, any detectable natural purpose and intention".[321]

Heidegger might have seen it that way, at least one can read him along such lines when he argues that what lies behind the technological will to mastery and hence 'endangers man', is "the view that man, by the peaceful release, transformation, storage, and channelling of the energies of physical nature, could render the human condition […] tolerable for everybody and happy in all respects".[322] This indeed suggests that Heidegger considered technological progress to be, at least on some level, driven or sustained by a utopian belief in the perfectibility of the human condition.

However that may be, existential risk theory certainly complicates a critique directed against technology and technological thinking on such grounds because it does not in the first place speak of an 'overflow of energy', of an, at bottom, irrational rage against the given, of a technology driven utopian visions of the future as we find them for instance in Soviet Cosmism.[323] It does not even necessarily have the goal to make the world a *better* place, but bases its call for technological mastery simply on the observation that without technological progress humanity could not be expected to survive for an extended period of time on cosmic timescales. It demonstrates, in other words, that the desire to bring order into the world, make it calculable and escape contingency, must not take its departure from utopian hopes, or a craving to compensate for the death of god, but can be born out of the simple, and arguably rather common-sensical, desire for survival.[324] A rebellion against the given and against contingency it is of course nonetheless, albeit a rebellion against the naturally preordained finitude of 'the human enterprise', rather than the rather benign imperfections of everyday existence.

[321] Kateb, G. (1997), p. 1239.

[322] Heidegger, M. (2009), p. 114.

[323] For recent thorough discussions of Russian cosmism see for instance Groĭs, B. (ed.) (2018); or Gray, J. (2012). As both authors demonstrate Russian cosmism, which emerged before and during the Bolshevik revolution, sketches futures in which communist ideals of perfect harmony amongst equals are realised on a cosmic scale through technology. We here find for instance visions that eerily echo contemporary conceptions of the singularity, where individuality has become a thing of the past altogether and humanity has morphed into one single consciousness in an event of technological symbiosis.

[324] This relates to Kateb's above quoted claim, according to whom technology is about putting a stamp on reality in the absence of a trace of any higher detectable purpose. Strong, T. (2012) explains the longing for technological mastery along similar lines.

Conclusion

This leads us to the, for the purposes of this thesis, perhaps most consequential difference between Heidegger's philosophy and existential risk theory: a (lack of) concern for the survival of the species. As we have seen, Heidegger, like existential risk researchers, identifies in technology an existential threat to humanity. However, he identifies in technology an existential threat to humanity in both meanings of the word, to humanity as humankind, i.e. to human life on earth, as well as to humanity's 'essential nature', i.e. to the properties that make us human, which to him, in the first place, is a certain form of being-in-the-world, involving an awareness of things that goes beyond the merely ontic, beyond the calculative, technological understanding of things. As we have seen, and crucially, for this discussion, for Heidegger these two dangers to humanity were not only essentially the same (essentially in the sense of 'expressive of the same essence') but they were connected in a highly problematic fashion. He was concerned that the possibility of total annihilation might come to dominate our thinking to the extent that we embark on a mindless frenzy to get technology technologically under control whilst turning a blind-eye to the less immediate, less visible, but in his opinion no less dangerous metaphysically rooted dangers associated with that quest. His fear thus was that the possibility of nuclear apocalypse would 'deliver us over' to standing reserve by providing the grounds for the ever further extension of technological control over ever more aspects of life.

His focus on the dangers to Being, humanity's humanness, hence was accompanied by near total neglect, even a dismissive treatment of the spectre of human extinction. But how are we to make sense of this dismissiveness? After all, the end of humanity would imply an end of the very possibility of a positive, meaningful transformation of Being in his sense - a shutdown of the very possibility of meditative thinking, dwelling, rootedness, openness to mystery, etc. - and, surely, Heidegger must have been aware of the peculiar imminence of the nuclear threat to the human being's survival under cold war conditions. Hence, if Heidegger's aim was to secure the conditions under which a meaningful life was possible how could he neglect the imminent threat that nuclear war presented to them. How could he neglect the need for action in the face of such a threat and focus all his philosophical energies on pinpointing the dangers of trying to master it instead?

Answering this question would deserve a more comprehensive and nuanced discussion than can be provided here. However, the remaining paragraphs of this chapter are intended to approximate an answer because doing so will lead us to what is at once at the heart of the tension between Heidegger's thought and existential risk theory and its common ground – namely that in both, albeit on different ontological levels, technology is the benchmark, the sole reference point for thinking about the human condition, present and future. As a result, in both cases, a more intricate understanding of the complexities of human existence, specifically of political reality, is squeezed out, resulting in, as it were, the motif of technology as destiny. Either one embraces it, or one rejects it, there is no path in-between.

One potential reason for Heidegger's refusal to take much interest in the danger to humanity's survival has been touched upon earlier. For Heidegger the atomic threat to humanity's survival was 'essentially the same' as mechanized agriculture or as the threat to the Rhine river, manifest in the hydro-electric power plant, or even as Nazi extermination camps - an embodiment of the metaphysical violence inherent to the challenging revealing of the modern ontological condition. Seen from that perspective the atomic destruction of humanity would merely amount to more of the same. It is in that vein that he argues in his 1950 lecture *The Thing* that:

> *"*Man stares at what the explosion of the atom bomb could bring with it. He does not see that the atom bomb and its explosion are the mere final emission of what has long since taken place, has already happened. Not to mention the single hydrogen bomb, whose triggering, thought through to its utmost potential, might be enough to snuff out all life on earth. What is this helpless anxiety still waiting for, if the terrible has already happened?"[325]

This ostensible lack of sensitivity to difference in ethical weight between the events in question earned Heidegger the scorn of many a prominent philosopher. Habermas (1989) calls this tendency in Heidegger 'abstraction by essentialization': "Under the levelling gaze of the philosopher of Being even the extermination of the Jews seems merely an event equivalent to many others".[326] And Richard Rorty calls Heidegger a 'self-infatuated blowhard' because "all that nuclear annihilation meant" to Heidegger, Rorty claims, "was one more bit of evidence for his claim to have understood *Das Wesen des Dinges* [the essence of the thing] better than Plato and Aristotle".[327] But does Heidegger's 'abstraction by essentialization' necessarily imply an ethical judgment on his part? Does it truly mean that he believes the utilisation of modern agricultural machinery to be no different, ethically speaking no more or less problematic than the utilisation of nuclear weapons and vice versa? Of course, if that were the case, that would explain his lack of interest in the survival of the species, because it would mean that humanity's destruction would be no more or less problematic an event than the mechanisation of agriculture. But for above reason that does not make sense. The existence of humanity is the condition of possibility for a positive transformation of Being and Heidegger does care about humanity, about poetry, etc., otherwise he would not care about Being's destiny. The only way in which we can make sense of the counter-intuitive callousness of Heidegger's levelling gaze is to understand it as the expression of a meditation about the *conditio sine qua non* of such events as a nuclear war or industrial agriculture. Interpreted that way, his statement that the production of nuclear weapons is 'essentially the same' as mechanized agriculture, does not necessarily imply an ethical evaluation. What it first and foremost means is that they speak

[325] Heidegger, M. (2009), p. 164.
[326] Habermas, J. (1989), p. 453.
[327] Rorty, R. (2005), p. 275.

the same essential language and become intelligible only when placed against the background of modern technology.

This, however, leaves the crux of the matter untouched, which is the question why, if there is an ethical difference, he did not take a heightened interest in the possibility of nuclear catastrophe and rally against it. A second potential answer to this question might be that, he simply did not think of the possibility of total annihilation as particularly terrible as compared to a world that would be barred such an abrupt ending and instead be allowed to continue on its path towards total technicalisation. George Kateb argues in that vein that Heidegger seems to suggest that "it is less bad for the human status and stature and for the human relation to reality that there be nuclear destruction than that […] genetic engineering should go from success to success".[328] Hubert Dreyfus arrives at a similar conclusion when he argues that Heidegger was less concerned with the physical havoc technology can wreak, than with the "devastation that would result if technology solved all our problems".[329] And, again, there are passages in Heidegger's work that do indeed suggest that he thought along such lines: "we do not stop to consider that an attack with technological means is being prepared upon the life and nature of man compared with which the explosion of the hydrogen bomb means little. For precisely if the hydrogen bombs do not explode and human life on earth is preserved, an uncanny change in the world moves upon us".[330] But even if Heidegger did consider nuclear destruction to be a lesser evil than total technicalisation, he must, at a minimum, have considered *both* extraordinary evils and thus have hoped for neither to materialise, which then again leads back to the original question.

The last potential answer to this question lies in Heidegger's peculiar perspective on human agency under the reign of the technological paradigm. What we find in Heidegger's thought is, in effect, deep fatalism and, correspondingly, a deeply apolitical, even anti-political attitude. He simply did not believe that anything practical could be done to rescue humanity from the technological evils it faced, be it from nuclear apocalypse or from the mechanisation of agriculture, for as long as the technological understanding of being itself had not been transformed, for as long as humanity's mode of Being had not fundamentally changed. But since the latter, for him, was out of the reach of active, wilful interference, nothing could be done at all: "no single man, no group of men, no commission of prominent statesmen, scientists, and technicians, no conference of leaders of commerce and industry, can brake or direct the progress of history in the atomic age".[331]

The scope of this chapter does not allow for an in-depth historical analysis of the anti-political dimension of Heidegger's thought and its connection to his philosophy of technology.[332]

[328] Kateb, G. (1997), pp. 1244-1245.

[329] Dreyfus, H. (2009), p. 26.

[330] Heidegger, M. (1966), p. 52.

[331] Heidegger, M. (2009), p. 22.

[332] For illuminating discussions of Heidegger's political thinking and its connection to his philosophy of technology see for instance Schürmann, R. (1978), Strong, T. (2012), ch. 7, or Blitz, M. (2014).

However, it might be worthwhile to recall that Heidegger's thinking changed profoundly throughout the 1930s and 40s (the so called 'Kehre', or 'turn') and that what is generally referred to as his 'philosophy of technology' is mainly the work of 'the late' Heidegger.[333] A reason for his anti-political post-war fatalism might hence be found in the 'pro-political' phase that preceded it.[334] As Heidegger pointed out in an infamously unapologetic letter to his former student Herbert Marcuse, he supported the Nazi regime precisely because he had hoped for a "spiritual renewal of life in its entirety" and even a "redemption of occidental Dasein".[335] In other words, he had hoped for a politically induced transformation of Being and thus a political solution to Western civilization's technological predicament. In 1935 Heidegger considered occidental Being to be equally threatened, virtually throttled, by US capitalism and USSR communism, whom he denounced as "metaphysically … the same".[336] Just as nuclear weapons and mechanised agriculture to him were essentially (i.e. metaphysically) the same, capitalism and communism were expressive of the same problematic relation to being and Heidegger, at the time, had hoped that something could be *done* in the political realm to rescue occidental Being from that root of all evil. After the war, whether out of real-felt disillusion, or simply in a calculated attempt to shield his personal and professional reputation, he argued that Nazi politics, too, had been but another incarnation of technological thinking, as evidenced by his callous remarks about extermination camps as 'essentially the same' as mechanised agriculture. In other words, he appears to have thought that he had made a mistake. He had placed his hopes in political action and ended up supporting a political movement which was but another incarnation of technology. What remained, in any case, was a deeply anti-political attitude and the idea that nothing good could possibly result of political action for as long as the Being of Western humanity had not fundamentally changed. But since humans, as we have seen, cannot exert control over Being, since we are the ones being 'challenged', 'spoken to', and since Heidegger accordingly did not himself claim to know wherein a new, non-technological mode of Being could consist, the late Heidegger's fatalism culminates in his well-known exclamation that "only a god can save us now" (not to be taken literally).[337] Until the arrival of such a new god, all humans could do was to keep meditative thinking alive, which means resisting the dominance of calculative thinking on a personal level, to remain open to the mystery and practice 'releasement towards things'.[338] Part of that exercise, to him, appears to have been to resist taking nuclear catastrophe seriously. In fact, and perhaps in response to his own earlier political hopes, he argued that the very act of thinking in terms of catastrophe, destruction, decline and loss are mere historiographical representations of technological consciousness:

[333] Cf. Borgmann, A. (2005).

[334] An argument to that effect can be found for instance in Wolin, R. (1990), specifically ch. 5.

[335] See letter from Heidegger to Marcuse on January 20th, 1948. In: Wolin, R. (ed.) (1998), pp. 162–163.

[336] Heidegger, M. (2000b), p. 40.

[337] Heidegger, M. (1981).

[338] Heidegger, M. (1966), p. 12.

> "All mere chasing after the future so as to work out a picture of it through calculation in order to extend what is present and half-thought into what, now veiled, is yet to come, itself still moves within the prevailing attitude belonging to technological, calculating representation. All attempts to reckon existing reality morphologically, psychologically, in terms of decline and loss, in terms of fate, catastrophe, and destruction, are merely technological behaviour".[339]

In other words, another potential explanation for why Heidegger might not have taken the problem of human extinction seriously, is that he feared that the mere act of thinking about the future under that aspect would necessarily lead him down a technological path of thinking. From his perspective, the very act of thinking about nuclear catastrophe, in fact the very act of thinking in the category of catastrophe or doom, is inherently technological. Nuclear weapons by the sheer fact of their existence were threatening to hijack the agenda, forcing a binary perspective on the future onto us, challenging us to think about the future of humanity in technical terms, whether we are thoroughly opposed to them or not. As the hydroelectric powerplant reveals the river Rhine as standing reserve, the categories of demise and doom, Heidegger feared, would reveal humanity as a technological problem. Interestingly enough, existential risk theory appears to substantiate these concerns. Existential risk research's self-declared mission is to establish the category of existential catastrophe as our primary benchmark for thinking about the future of humanity. As a result of doing so (cf. chapter 1), our visions of the future are reduced to purely technical ones. By focusing on catastrophe and destruction, humanity is mentally transformed into an optimisation process to the effect that the notions 'future of humanity' and 'future of technology' become undistinguishable.

Arguably, all of the above listed reasons might have played a role in Heidegger's decision not to seriously ponder the problem of human extinction in its own right. What is clear in any case, is that, at no point in his writings, he indicates that he considered the preservation of human life a cause worth struggling for. Whether that was because of a thoroughly fatalistic perspective on political action in a technological age, out of fear that it would deliver him over to technological thinking, or because he simply did not really care - what matters for the purposes of this thesis is that Heidegger did not do so. His holistic treatment of technology, his abstraction by essentialisation, meant that he brushed over the particular problem of the threat to the continued existence of the species.

We thus find ourselves confronted with a curious situation. In both, existential risk research and Heidegger's thought, technology occupies a similar role and emerges as an, at heart, existential problem, as intrinsically and inexorably intertwined with humanity's destiny. But, on the one hand, existential risk theory does not take technology philosophically seriously, making it blind to the dilemmas and paradoxes of 'technology as destiny', whilst Heidegger, on the other hand, transforms the problem of technology into a matter that is entirely out of humanity's reach, and does not take the problem of human extinction seriously, treating it, if at all, merely as derivative of the wider problem

[339] Heidegger, M. (1977), p. 48; The phrase 'historiographical representation of technological consciousness' was borrowed from Kroker, A. (2002).

of technology.

Accordingly, we find on both sides a near-complete neglect of the other sides' concerns. Heidegger's conception of technology as a metaphysical force meant that he did not allow himself to acknowledge the need for purposeful, organised action and the positive role technology might play in the face of total catastrophe. Existential risk researchers, on the other hand, for whom technology is a tool that might allow humanity to escape its otherwise naturally preordained destruction, abstract from the kind of fears that Heidegger entertained, fears associated with problems such as thoughtlessness, objectlessness and standing reserve, not least because, from their point of view, all of these threats, even if one were to take them seriously, would have to be considered secondary as compared to the threat of existential catastrophe.

In the next chapter I will turn to philosophers who occupy a position between these two poles and might provide us with a basis for developing a more nuanced perspective on the wider context of existential risk and the positions of Heidegger and existential risk theory within it.

3. Existential pressure

Introduction

The first chapter argued that in existential risk theory, the future is framed in such a way that technology emerges as humanity's destiny. It is only through a continuous extension of humanity's technological capacities and a technological transformation of the human condition that humankind can be expected to survive in the long run. The motif of 'technology as destiny' suggests that the phrases 'future of humanity' and 'future of technology' become for all practical purposes interchangeable. The preceding, second chapter surveyed Heidegger's philosophy of technology. According to Heidegger's 'Question Concerning Technology', the essence of technology itself is nothing technological but, rather, a destining of being; Namely a way in which reality reveals itself to us. Understood this way technology has an inherently challenging and ultimately all-encompassing dynamic. Heidegger's philosophy of technology frames 'technology as destiny' and therefore can be said to correspond with existential risk theory in that regard.

In sum, Heidegger's account of technology enables a critical reflection of existential risk theory as part of a temporally extended, philosophical effort to come to terms with the ultimate possibilities of technology's ideational and material hold over modern humankind. Heidegger's philosophy provides us with, as it were, a conceptual toolkit to embed existential risk theory conceptually and historically, and a counterpoint to reflect about what 'technology as destiny', fleshed out in terms of standing-reserve, might actually mean: an objectless world of perpetual mobilisation for no conceivable purpose.

At the same time, as was explained in Chapter 2, a concern for the survival of the human species appears to be absent from Heidegger's thought, despite an evident awareness of the threat that nuclear weapons posed to humanity's survival. At no point did Heidegger call for political action *against* nuclear armament. He did not even single out nuclear weaponry as a technology that would warrant special philosophical attention. Such absences stem from his holistic treatment of technology, whereby thinking about the future of humanity in terms of survival, catastrophe and doom, were considered by Heidegger to be technological behaviour that could only result in 'delivering us over' to technology. Existential risk research, for which thinking in terms of doom, catastrophe, and survival is its very raison d'être, inadvertently allows us to see what Heidegger might have had in mind – technology emerges as the only path forwards. Confronted with an aporetic ending, with the choice between doom in the form of a threat to the human species, and doom in form of a threat to humanity's Being, Heidegger appeared to have succumbed to fatalism - a deeply apolitical, even anti-political, retreat to thought and, in effect, pure presentism.

Heidegger’s students Hannah Arendt and Günther Anders (on whose work I will focus in what follows), certainly did not follow Heidegger’s neglectful treatment of the possibility of a nuclear catastrophe. Arendt and Anders appropriated many features of Heidegger’s philosophy of technology. But they neither wholly abandoned their belief in human agency, nor did they accept that the precariousness faced by humanity - epitomised in the nuclear arms race between East and West - could be explained in terms of unfolding epochs of being, i.e. under total abstraction from political, social, psychological, or economic realities.

As Waseem Yaqoob (2014) notes, Arendt believed that Heidegger “could grasp world-historical processes, but not the political character of the ‘world’ that was in the process of being lost. Rather than treating science and technology in terms of unfolding essences, Arendt sought to stress their contingent development as part of a parable about the unpredictability of human action”.[340] Additionally, Arendt was, to put it in Seyla Benhabib’s (2003) influential formulation, a “reluctant modernist”, who did not wholly break with the humanism of the Enlightenment, as Heidegger had done, nor did she wholly forsake her belief in the potential healing powers of politics. [341] Anders differs from Heidegger along similar lines. Like Arendt, he observed that “interest in moral or political participation or action […] has become extinct in Heidegger's philosophy. The only thing, the ‘Dasein’ takes into its own hands, is the Dasein itself; each individual in his individual hands - in spite of the world”.[342]

The differences between Heidegger and his students with regards to human agency are particularly noticeable with regards to the nuclear bomb and the problem of human extinction. For Arendt and Anders the nuclear bomb highlighted the exponentially increased powers of political elites, and thus indicated a monstrous transformation of human agency under modern technological conditions. As Steven White puts it, Heidegger “failed to see that the threat of nuclear extermination of life shifts the terms of attachment to existence in a fundamental way: the inessentiality of things, their precariousness, now has a novel relation to human choice”.[343] I have argued in Chapter 2 that, for Heidegger, the nuclear bomb was for the most part no more than the materialisation of the essence of technology and thus essentially no different from other technological problems. For Arendt and Anders, on the other hand, the nuclear bomb in a sense highlighted that the problem of technology could not be reduced to an essentialist treatment along the lines of ‘unfolding essences’. In a twisted way, the nuclear bomb had *re-established* human choice and thus agency as a decisive factor in humanity’s destiny, both in theory and practice.

Faced with the negative omnipotence of human beings (or rather, of the politico-military elites) under nuclear conditions, Anders in particular rallied against both, uncritical, technocratic

[340] Yaqoob, W. (2014), p. 205.
[341] Benhabib, S. (2003).
[342] Anders, G. (1948), p. 350.
[343] White, S. (2000), p. 28.

interaction with technology, as well as against Heideggerian fatalism, denouncing, in effect, the one as irresponsibly thoughtless and the other as irresponsibly intellectualistic. In his personal notes on Heidegger, collected in *Über Heidegger,* Anders expresses his deep frustration with Heidegger's treatment of the nuclear question:

> "In his second phase, just as in Being and Time […], Heidegger locates a momentous guilt at the origin of everything. A guilt that is no guilt – a guilt which needs to be atoned for by reclaiming being, the remembrance of being, a remembrance on which the destiny of the occident depends, indeed, the destiny of the entire new era. Such momentous moral tasks he gives himself […] Such a task. In times of camps and of 'the bomb'. Where the real tasks are to change beings, not being, to secure human beings that *are*, not being; at this point, suddenly, for him, whatever happens […] turns into a destining of being. The gruesome things that happen are a destining and one does not oppose destining; rather, one focuses all one's energies on attacking an allegedly far greater evil, the guilt of the forgetfulness of being – the disgrace; to fight it with all one's might, a struggle which is exclusively concerned with keeping thinking about being alive, which otherwise threatens to withdraw itself, this – oh this, in his eyes, is a fight that is much more than a deed that merely changes the world, it is world-transformation *qua actus*. Sure, he does vaguely touch upon specific, moral defects of the time. But all of them are nothing but symptoms of the oblivion of being. That is, of the 'Fall of Man'".[344]

It is important to note that Anders here does not necessarily express disagreement with Heidegger's analysis, instead protesting Heidegger's intellectualistic passivism and his associated recipe for dealing with the modern human condition: to develop *Gelassenheit* ('releasement-towards-things'), namely to focus on one's own mental sanity in spite of all of the factual and potential tragedies that (nuclear) technology held in stock for humankind. Anders' own philosophy is the exact opposite, a decades-lasting, deeply political rallying cry against nuclear weapons and modern humanity's naïve and paradoxical treatment of technology, both in thought and action. In doing so, Anders highlighted the momentous importance of political activism (understood as an offshoot of human imagination and human agency) in the 20th century, so as to 'recapture technology' and re-emancipate humanity.

[344] See Anders, G. (2001), pp. 299-300, translated by the author. The original text reads as follows: "Wie immer man diese Fragen beantwortet, entscheidend ist, daß H. genau wie in ‚Sein und Zeit' (das er mit dem verächtlichen, an das Zuhandene ‚verfallene' Dasein begonnen hatte) auch nun, in seiner zweiten Phase, eine ungeheure moralische Schuld, die keine ist, an den Anfang stellt - eine Schuld, die angeblich auf ebenso ungeheure Weise getilgt werden müsse wie, und zwar durch Rückgewinnung des Seins, durch das Andenken an das Sein, durch ein Andenken von dem das weitere Geschick des Abendlandes, ja das neue Zeitalter abhänge. So ungeheure (freilich durchweg durch das Denken zu lösende) moralische Aufgaben stellt er sich diese Aufgabe. In einer Zeit der Lager und der Bombe. Denn wo die wirklichen Aufgaben liegen, die Aufgaben, die darin bestehen, Seiendes, nicht Sein, zu verändern; seiende Menschen, nicht Sein zu retten; da wird für ihn plötzlich alles, was geschieht - ob es nun, wie in der Rektoratsrede oder der in diesem Jahre veröffentlichten Metaphysischen Einführung, der Nationalsozialismus, oder im Humanismusbrief, ganz vorübergehend (man kann es nie wissen) der Marxismus -, ein Seinsgeschick. Die elenden Dinge, die passieren, sind Seinsgeschick, ihm oponiert man nicht; statt dessen hämmert man los auf eine angeblich viel gewaltigere Schuld, eben die Seinsvergessenheit - sie ist eine Schande; und gegen sie mit aller Verwegenheit den Kamp auszufechten, einen Kampf, der ausschließlich im Denken an das sich entziehende Sein besteht, das - oh, das ist in seinen Augen weit mehr als eine "Handlung", die nur die Welt verändert, das ist qua actus die Weltveränderung. Gewiß, auch gewissen moralischen, vage formulierten Defekten der Zeit gedenkt er. Aber sie alle sind nichts anderes als Folgen der Seinsvergessenheit. Also des Sündenfalls".

However, for the purposes of this chapter it is less the precise nature of Arendt's and Anders' politics that matters, than the ways in which it influenced their philosophical analysis of the nuclear threat. What makes them interesting for the purposes of this thesis is that they constitute a middle ground between Heidegger and existential risk theory. As explained in previous chapters, through existential risk theory the problem of technology becomes invisible in normative and theoretical terms due to its instrumental conception of technology. In turn, in Heidegger's philosophy of technology technology assumes a position of such overpowering significance that everything else becomes invisible, leaving no room for human agency or an appreciation of the unique philosophical implications of the problem of human extinction.

Arendt and Anders are between the two poles. They agreed with Heidegger that modern technology is much more than a mere collection of neutral tools that can be used for good or bad but needs to be understood as something that is deeply entangled with and affects the particular way in which humans exist, perceive and act in the world. Like Heidegger, they can be said to have wanted to make technology visible as a force that shapes modern human existence. Accordingly, many aspects of Heidegger's philosophy of technology, in particular his phenomenological approach, are reflected in their work. At the same time, their thought diverges from Heidegger's holistic treatment of the modern human condition on many levels. It is united with existential risk theorists in a deep and practical concern for the survival of humanity. Anders' language in particular is characterised by a comparably alarmist tone, stressing the imminent importance of *action* to avert the worst. But what is arguably most important for the purposes of this thesis, is that their non-holistic perspective on human existence on the one hand side, and their phenomenologically schooled, critical perspective on technology on the other hand side, allowed them to discuss the problem of (what we now call) existential risk as part of a complex transformation of the human condition in the present.

According to Arendt and Anders the new negative omnipotence of modern humans, the monstrously inflated levels of human responsibility under nuclear conditions had transforming implications for the human condition of their own, they directly affected what it *meant* to be human. Apart from embodying the nihilistic essence of the technological understanding of being, it confronted humanity with a new perspective on its existence and thus a new sense of being-in-the-world. Heidegger curiously missed that idiosyncratic phenomenological effect of the spectre of total annihilation. He did not investigate if and how it might affect what he calls Being, the human way of being-in-the-world. Anders' thought on the other hand, and we can also include Arendt's here (at least where it is concerned with nuclear weapons), can be read as a meditation on the transformation of Being under nuclear conditions, or, as Babich (2013) phrases it, "the phenomenological effects of the end-time".[345]

Above all, humanity's power to destroy itself meant to them that humanity had lost its

[345] Babich, B. (2013a), p. 152.

innocence. Our idea of humanity had been transformed from one in which we could safely assume to be part of an open-ended community of generations into one where, suddenly, we had to think of humanity as mortal and of the future as a space that no longer came of its own but had to be 'produced'.[346] The nuclear bomb meant that the existence of the or, rather, *any* future at all could no longer be taken for granted and was dependent on decisions made in the present. To them this implied that humanity was deprived of central conditions that used to structure human life and shaped our categories of thought. Arendt for instance highlights the monstrous transformation of politics, and thus a vital aspect of how she understood the human condition, once infused with, what she calls, 'the radical evil' or 'all-or-nothing' questions. What Arendt and Anders have in common is the idea that humanity had therewith entered a schizophrenic age. In spite of the multiple ways in which their thought diverts from Heidegger's, Heidegger's characterisation of the technological understanding of being – the idea that modern technology is rooted in, expressive of, and conducive to the expansion of, a particular ontological condition or attitude to being, which differs categorically from how phenomenal reality reveals itself naturally to our consciousness - is of enormous importance if we are to understand their analysis of the transformative implications of the nuclear age, i.e. of the emergence of existential risk.

Part of what I want to show in this chapter is that existential risk theory can be understood as indicating a further complication of the situation Hannah Arendt and Günther Anders had been writing about. Whilst these authors were confronted with one existential risk, i.e. the possibility of an all-out nuclear war, existential risk theory highlights that we are now facing several risks that are comparable in scope. I.e. when according to Anders the bomb was 'ontologically unique',[347] this has now changed. I argue, that therefore, with existential risk theory, our conception on the future changes again and becomes in some ways more complicated than the one that the above authors described. What is new about existential risk theory, as was discussed in Chapter 1, is not so much that it takes the possibility of human extinction seriously. Arendt and Anders already did that and arguably they took it much more seriously than existential risk researchers do today. What is new about existential risk theory is the way in which the concept frames our perspective on the future, namely as a technological optimisation problem, where we are encouraged to arbitrate different extinction risks against one another and to develop a perspective on the future akin to that of a minefield, a space which needs to be scanned and mapped, where we must carefully consider each and every step we make in order to survive for any considerable amount of time.[348]

As will be outlined in this chapter, this is a perspective that differs in various respects from

[346] Anders, G. (1956a), p. 43.

[347] See Dijk, P. v. (2000), p. 134.

[348] Häggström (2016) is emblematic of this perspective on the future. Häggström's book is entitled '*Here be Dragons*' because he likens our present situation to that of medieval mapmakers who decorated unchartered territories with mythological creatures, such as dragons, in order to warn seafarers about unknown dangers. See Häggström, O. (2016), p. 6.

that of Arendt and Anders. This, however, does not mean the latter's analyses of how the threat to humanity's future affects us in the present should be considered outdated. On the contrary, I intend to demonstrate that existential risk theory in many respects underscores their respective analyses' ongoing relevance. The fact, for instance, that all risks, anthropogenic and natural are turned technological optimisation problems echoes Arendt's observation that history and nature are in the process of becoming one owing to progress in modern science and technology.

In sum, Anders' and Arendt's works provide us with a suitable touchstone to further embed existential risk theory in traditions of thought about the problematic relationship between modern technology, human agency and human value. They are much closer to the concerns of existential risk researchers than Heidegger in so far as they did take the risk of human extinction (or 'the nuclear question') seriously not only as a topic of philosophical investigation but also in political terms. At the same time, they were heavily influenced by Heidegger in so far as they understood that risk as resulting from and embodying a peculiar Western technological mindset or, as Heidegger put it, 'a revolution in philosophical concepts', that was ripe with dangerous and self-defeating pretensions of the possibility of mastery. Like Heidegger, they saw the possibility that, for as long as the underlying mindset was not challenged, the spectre of extinction would turn out to foster rather than weaken these pretensions and translate into an even stronger hold of technological rule over modern humanity. However, since they did not think about history in terms of unfolding metaphysical essences, they did not wholly forsake their believe in human agency. They conceived of the technological mindset as part and parcel of the political predicament of the day and something that needed to be engaged with both philosophically and politically. Since they had not wholly forsaken their belief in human agency and reason, they made it their mission to pinpoint exactly how and *where* the technological mindset leaves its mark in everyday and political practice, with what consequences, and why it cannot deliver what it promises.

Since the limited scope of this chapter does not allow for a comprehensive discussion of the role of science and technology in the works of Arendt and Anders, in the following I proceed by connecting existential risk research to four individual key-motifs in Arendt's and Anders' thought. Accordingly, the chapter is divided in four sections. The title of each section refers to the respective motif in Arendt or Anders works which forms the pillar of the discussion. These are, in order of appearance, 'Promethean shame', 'technology as action', 'all-or-nothing or the radical evil', and finally 'collective schizophrenia'. In each section I begin by briefly introducing and elaborating on a key theme in existential risk theory that is either explicitly discussed in the literature or that I have identified as implicit to it in my previous analyses. I then relate the thus established theme to the above motifs in Arendt's and Anders' thought. Sometimes the order is reversed with Arendt's, or Anders' work providing the vantage point of the discussion.

The concept of Promethean shame is a centre piece of Anders' philosophy of technology. It brings Heidegger's argument that the technological danger to humanity as a species cannot be separated from the danger to humanity's humanness 'down to earth', as Babette Babich puts it, by demonstrating how in our every-day interaction with it, technology has replaced the human as the measure of things. The concept of 'technology as action' can be seen as the cornerstone of Arendt's criticism of the instrumental conception of technology. It firmly embeds her in what Langdon Winner calls the 'anti-mastery tradition' of philosophy of technology, both highlighting that she was already attune to the generalised concerns about technology familiar from existential risk theory and providing us with a starting point for critical reflection about the policy recommendations in existential risk theory.[349] 'The radical evil' can be seen as Arendt's term for what we nowadays call existential risk. Both Arendt and Anders were no less concerned about the possibility of human extinction than existential risk researchers are today. However, as the concept of the 'radical evil' allows us to establish, Arendt's and Anders' main concern was with how the emergence of the spectre of self-annihilation affected humanity in the present, rather than seeing it simply as a threat to the future, and what this tells us about the role of technology in our lives. These types of reflections are largely absent from existential risk research. The motif of 'the schizophrenic condition of modern existence', lastly, brings the discussion back onto the ontological level introduced in the preceding chapter. Both Arendt and Anders shared Heidegger's view that the challenges stemming from modern technology have their roots in an ontological clash between what Heidegger terms the 'technological understanding of being' and the natural way in which things show themselves to our consciousness. This clash is what ultimately sustains 'Promethean shame', 'technology as action' and the 'radical evil'.

I close the chapter with a concluding section where these discussions are brought together in order to discern some general take-away points on what the established connection between existential risk theory and aspects of Arendt's and Anders' work entails for our perspective on existential risk theory. Generally speaking, Arendt and Anders can be seen as united in an effort to 'think through' and interpret the relationship between technology and human affairs under conditions of what I have earlier called 'existential pressure'. As we have seen, Heidegger escaped from facing that pressure by placing the blame on the history of being itself, effectively converting it into a question of fate.[350] Existential risk researchers can also be said to escape the problems posed by 'existential pressure'. By embracing a purely technical perspective, such researchers transform the problem into an ostensibly manageable one. However, in doing so, this approach evades the difficult questions that existential risk confronts us with. In particular, the methodology of existential risk researchers detaches itself from a variety of ontologically rooted complications that need to be taken into account if we truly are to 'take existential risk seriously'. Arendt's and Anders' thought allows us to bring those out.

[349] Winner, L. (2001), pp. 95–97.
[350] Cf. Dupuy, J.-P. (2015), p. 13.

3.1 Promethean shame

Agential risk

Phil Torres, director of the FHF and research associate at the FLI, recently published two papers entitled respectively 'Agential risks and information hazards: An unavoidable but dangerous topic?' and 'Who Would Destroy the World? Omnicidal Agents and Related Phenomena'.[351] In both papers Torres discusses the same question: how can humanity survive the dissemination of access to "advanced dual use technologies" across society in the coming centuries? Torres observes two trends in contemporary technological development: First, technology is becoming increasingly powerful in many sensitive areas. Second, technology is at the same time becoming increasingly accessible for a growing user base. For Torres, these trends imply that increasingly destructive technological capabilities are bound to disseminate across society in the future. The logical conclusion for Torres is that access to weapons of mass destruction – including 'weapons of total destruction' (WTDs) – will no longer be limited to actors with the financial and organisational resources of nation states but will eventually be available to non-state actors such as terrorist groups or even single individuals.

The production of military grade nuclear bombs for instance requires the concerted efforts of hundreds, if not thousands in scientific personnel; large-scale infrastructure; the development of global supply chains; as well as the investment of hundreds of millions of US dollars. In contrast, Torres believes that such might not be the case for 'dual-use emerging technologies':

> "CRISPR/Cas-9, base editing, digital-to-biological converters, nanotechnology, drones (e.g., "slaughterbots"), SILEX (i.e., separation of uranium isotopes by laser excitation) […] are not only enabling humanity to manipulate and rearrange the physical world, for better or worse, in unprecedented ways, but placing this power in the hands of more and more states, groups, and even lone wolves".[352]

If this scenario were to materialise, Torres claims, it would become virtually *inevitable* that existential risk levels would be inflated beyond anything conceivable today since "the probability of any given individual pressing a 'doomsday button' does not need to be very high per century for an existential catastrophe to be more or less certain" if the number of individuals with access to that button is sufficiently large.[353]

Such fearful expectations, identifying an omnicidal tendency inherent to the democratization of technological power, are not exactly new.[354] However, there is presently a renewed interested in

[351] Torres, P. (2017a, 2017c).

[352] Torres, P. (2017a).

[353] Ibid, p. 1.

[354] A prominent earlier articulation of this concern can be found for instance in Bill Joy's widely read article 'Why the future doesn't need us' Joy argues that "it is no exaggeration to say we are on the cusp of the further perfection of extreme evil, an evil whose possibility spreads well beyond that which weapons of mass

the problem due to existential risk research. This trend is for instance encapsulated in CSER's mission statement which warns that future "technologies might provide direct and relatively short-term control over circumstances essential to our survival, and place that control in dangerously few human hands".[355] Another example is Wittes' and Blum's 2015 volume *The Future of Violence: Robots and Germs, Hackers and Drones: Confronting a New Age of Threat*, where the authors identify the underlying problem as that of technological 'mass empowerment':

> "in our modern age . . . new technologies are able to generate and channel mass empowerment, allowing small groups and individuals to challenge states and other institutions of traditional authority in ways that used to be the province only of other states. They are growing increasingly cheap and available. They defy distance and other physical obstacles. And, ultimately, they create the world of many-to-many threats, a world in which every individual, group, or state has to regard every other individual, group, or state as at least a potential security threat".[356]

Bruce Schneier of the FHF and head of (cyber-)security at IBM states in a similar vein that "sooner or later, the technology will exist for a hobbyist to explode a nuclear weapon, print a lethal virus from a bio-printer, or turn our electronic infrastructure into a vehicle for large-scale murder".[357] Andrew Snyder-Beattie of the FHI, discusses this problem in an article published in the Bulletin of the Atomic Scientists entitled 'Small Groups, Dangerous Technology: Can They be Controlled?'.[358] And, to give one last example, Eliezer Yudkowsky of MIRI summarises these concerns in what he calls 'Moore's Law of Mad Science': "every eighteen months, the minimum IQ necessary to destroy the world drops by one point".[359]

In the light of such looming technological 'mass empowerment' or 'universal unilateralism', Torres posits whether there is an "increasingly pressing question of who would destroy the world if only the means were available"? [360] In other words, Torres claims that it is increasingly important to focus not on technologies and how to make them safer, but on "the users who would exploit them for

destruction bequeathed to the nation-states, on to a surprising and terrible empowerment of extreme individuals", see Joy, B. (2000). See also Kurzweil, R. (2013), or Drexler, (1986, 2006), both of whom voiced comparable concerns in connection to advanced nano-technology.

[355] Viz. CSER (2018a). The above quoted passage might be misleading in so far as it speaks of the placement of technological power in 'dangerously few hands', rather than 'dangerously many'. This might suggest that CSER identifies the problem in the centralisation rather than the decentralisation of technological power, which, however, is the problem at hand here. However, if read this quote against the background of publications associated with CSER, such as Martin Rees's *Our Final Century* (2004), it is clear that what CSER wants to convey is that the risk stems from the fact that fewer hands will be required to wield the kind of technological powers that formerly required the cooperation of 'many hands'. i.e. large-scale organizational efforts, as in the case of nuclear technology.

[356] Wittes, B. & Blume, G. (2015), p. 20.

[357] Schneier, B. (2013).

[358] Snyder-Beattie, A. (2015).

[359] Yudkowsky, E., as quoted in: Sandberg, A. (2012). Yudkowsky is founding-director of the Machine Intelligence Research Institute (MIRI). As such, he is a central figure in the existential risk movement, specifically when it comes to debates surrounding artificial intelligence.

[360] Torres, P. (2017a), p. 8.

bad ends".[361] In order to capture this shift in perspectives Torres proposes to introduce a new category of existential risk, namely 'agential risks', defined as "risks posed by any agent who could initiate an existential catastrophe in the presence of sufficiently powerful dual-use technologies". [362]

'Agential risk research' would, according to Torres, focus not on horizon scanning, differential technological development, safety engineering, etc., but on identifying groups and individuals, that might express "omnicidal, mass genocidal, anti-civilizational, or apocalyptic beliefs/desires".[363] Torres thus sets out to identify and categorise beliefs, desires and mindsets, of which he believes that they might result in a willingness to "to exploit current and future technologies to bring about an existential catastrophe". [364] On balance he appears to be most concerned by what he calls 'apocalyptic terrorists' who entertain 'active-cataclysmic modes of believes' (individuals who consider themselves as active parts in a wider eschatological narrative) because such forms of apocalypticism have repeatedly resulted in large scale violence in the recent past.[365] Torres lists for instance the 1995 Tokyo subway sarin attacks by Japanese doomsday-cult Aum Shinrikyo, whose adherents hoped to trigger a third World War, or specific forms of Islamic apocalypticism that played a critical role in the eschatological narratives entertained by Daesh.[366] Torres argues that a "book's-worth of examples could be adduced to show just how common across space and time, geography and history, violent apocalyptic movements have been",[367] and concludes that "we will need to devise effective strategies to mitigate agential risks".[368]

Promethean shame

The motif of 'agential risk' and related concepts such as 'universal unilateralism' or 'technological mass-empowerment' resonate deeply with the centre piece of Günther Anders' philosophy of technology: 'Promethean shame'. These concepts therewith underscore, as will be demonstrated in the following paragraphs, the striking actuality and analytical force of Anders' theses on the 'obsolescence of the human being' in the context of existential risk theory.

Given the instrumental conception of technology in existential risk theory, 'agential risk' has the benefit of focusing our attention on what, necessarily emerges as the ultimate source of existential risk: malicious intent, good versus evil, or ignorance. In other words, the whole human

[361] Ibid, p. 7.
[362] Torres, P. (2017a), p. 2.
[363] Torres, P. (2017c), p. 1.
[364] Ibid.
[365] See also Torres, P. (2017b), specifically section 1.7.
[366] Ibid, pp. 3-4.
[367] Torres, P. (2017a), p. 4.
[368] Torres, P. (2017c), p. 16.

predicament.[369] Only rarely is it expressed as clearly as by Wittes and Blum in their approving invocation of Microsoft security chief Nathan Myhrvold, who argues that "most people now will use new biotechnologies to prevent disease; a few will use them to cause it […] technology contains no inherent moral directive -it empowers people, whatever their intent, good or evil." [370] As we have seen, if technology is *defined* neutrally, i.e. as mere instruments that can be used for good or for ill, the risk necessarily lies with the human using or designing the technology and thus comes down to ignorance or ill intend. Torres hence presents as findings what is built into the assumptions of existential risk theory. As discussed in chapter 2, the problem of the neutrality thesis is that it makes technology invisible on a normative level. The problem, in all cases, is not technology, but humans: human unpredictability, human faultiness, human propensity for violence and warfare, human lack in prescience, etc. Technology itself, from this perspective, remains benign.

The relationship between humanity and technology that results from this perspective is the subject of Günther Anders' main work *Die Antiquiertheit des Menschen* (the 'obsolescence of the human being').[371] The central component of Anders' philosophy of technology is the concept of 'Promethean Shame', captured perhaps most vividly in his re-phrasing of the tale of Icarus: "If only", Anders argues, "his wings could drop Icarus as ballast – they would be free to conquer the skies".[372]

Most generally understood, Promethean shame denotes a feeling of inferiority vis-à-vis the perceived perfection of technology in comparison with which humans begin to conceive of themselves as "faulty constructions".[373] The saying by Protagoras that 'man is the measure of all things', Anders argues, has lost its validity in the machine age where technology has become the point of reference for how we conceive of ourselves and the world in general.[374] Only by elevating the category of the machine as universally valid and exhaustive, Anders claims, the non-constructed can even begin to appear as 'faultily constructed'. Promethean shame thus indicates that humanity has acknowledged the superiority of the machine and begins to confuse the machine's needs for its own.[375] There are echoes here of Heidegger's great danger, that humanity might deliver itself over to technology. In a similar fashion to Heidegger's concept of Enframing, Anders' Promethean shame implies that the project of scientific and technological progress is inherently *un-anthropocentric,* even *anti-anthropocentric*. It is not the human that occupies centre stage but technology itself.

But how does that come about? How is it that modern technology has become the measure of all things? Anders lists a variety of potential sources of this feeling. One is "the shame of having

[369] To be more precise: "maniacs," "lunatics," "misanthropes," "sociopaths," "nefarious dictators," "belligerent tyrants," "agents of doom", "suicidal regimes or terrorists", "garage fanatics and psychopaths", "criminal groups, terrorists, and lone crazies", as cited in Torres, P. (2017a).
[370] Wittes, B. & Blume, G. (2015), p. 20.
[371] Anders, G. (1956a).
[372] Ibid, p. 20.
[373] Ibid, p. 34.
[374] Ibid, p. 30.
[375] Ibid.

been born instead of having been made".[376] Anders identifies a deep sense of forlornness at the heart of modern human existence and a resultant, almost pathological desire to remake the human condition, to turn it into something artificial in order to finally have an explanation, no matter how arbitrary, for why things are one way and not another. [377] Hannah Arendt, in the preface of *The Human Condition*, similarly claims that modern man appears to be possessed by the desire to escape "human existence as it has been given, a gift free from nowhere (secularly speaking)" and to exchange it "for something he has made himself".[378] However, apart from such existentialist analyses of the desire to ground human existence by technological means, Anders demonstrates authoritatively, how Promethean shame results from our practical, everyday interaction with technology, which I focus on in the following sections.

I have argued in chapter 2 that our conventional conception of technology renders it normatively invisible and therefore *effectively* as something inherently positive because what remains as its only defining quality is that it expands our option-set, which is generally considered to be something desirable. Anders goes further than that. He argues that our understanding of technology as a neutral means to an end, translates into a situation where technology has become our idea of perfection. Our conception of technology as a neutral means implies that we, at any given point in time, conceptualise technology as *potentially* perfect, whilst faultiness is reserved for the human being. When we project ourselves into the future, making technology our benchmark, we thus see a world of *potential* perfection, where all options technology can in principle provide us with have been realised and make that vision the benchmark, the point of reference, for thinking about future and present. The implication is that we ourselves are constantly rendered the *saboteurs* of our own products, for, if it were not for our faultiness as engineers or users, these *could* yield perfect results. Anders' allegory of Icarus illustrates the underlying idea rather well. If it had not been for Icarus's hubris (or for Daedalus's bad design choices) the wings, as an idea or, to put it in Bostrom's words, as a 'technological capability' (see Chapter 1, p. 40), namely the capability of flying, could *in principle* have conquered the skies.

Anders hence inverts the perspective on the neutrality thesis – his question is no longer, at least not in the main: what does the neutrality thesis imply for how we view technology and its risks? Rather, he asks: what does the neutrality thesis imply for how we see ourselves? Promethean shame is felt *in response* to the machine. The effect Anders achieves with this simple shift in perspectives is remarkable and I think it comes out very clearly in above discourses about agential risk. The problem is that when we portray machines as neutral tools we make their *potential,* which is prima facie

[376] Ibid, p. 24.
[377] Anders first important publication 'Pathologie de la liberté', was published in exile in Paris in 1936 and has only recently been translated into German [see Anders, G. (2018)], had a profound impact on French existentialism, specifically on Sartre's work. For a discussion of Anders' existentialism and the impact of his thought on French existentialism see for instance Dries, C. (2018), In: Anders, G. (2018).
[378] Arendt, H. (1998), pp. 2-3.

limitless, the measure of things and feel *ashamed* for not being able to realise it, blaming it on human faultiness.

In the context of existential risk, as Anders makes clear in his discussions of the nuclear bomb, Promethean shame assumes a wholly new quality. For Anders we have erected a world of machines around us, which is premised on the idea of perfection and therewith we have created a new reality in which we no longer only *feel* as imperfect and faulty constructions vis-à-vis the machine but truly have become imperfect, faulty constructions in view of the requirements of the times:

> "Man […] in the midst of all the products he is producing, with his body that was the same millennia ago, seems not only outdated but hopelessly left behind. Promethean shame, the referral to machines as the measure of judgement in self-reflection and one another, implies the total rejection of anything given including of ourselves as givens. We become smaller than ourselves, because we are no longer on an equal footing with our own products".[379]

The above discussion of 'agential risk' illustrates Anders' fears closely. Technological progress is seen as a given, technology as neutral ('dual-use') and therefore the need for adjustment comes to rest on humanity, to such an extent in fact that the faulty, unpredictable, nature of humankind is beginning to be rendered a liability too large to bear. Anders therefore identified in humanity the longing to 'flee into the camp of its machines' – a point strongly echoing what Heidegger considers as the great danger of course, the idea that ultimately humanity would turn into 'standing reserve', into a product and a technological optimisation problem itself – to become as faultless and predictable as machines in order to live up to the requirements of the artificial world it had created. Promethean shame thus ultimately amounts to the longing to become a fabricated thing, to be freed from the constraints and limitations of ones given, embodied human existence and become one with the dynamic world of machines so that its enormous potential can be realised without human nature sabotaging the project.[380]

We see exactly that dynamic playing out in contemporary discourses surrounding agential risk. Martin Rees for instance hypotheses about potential futures in which pre-crime interventions of the type depicted in the science fiction film Minority Report might become both necessary due to 'agential-risk' as well as feasible due to advances in genetics and physiology which might help us identify potential high-risk individuals: "If our propensities are indeed determined by genetics and physiology […] then identifying potential criminals may soon not require psychic powers. There will then be growing pressures to institute this kind of pre-emptive action in the real world, as a safeguard against the outrages - ever more calamitous with each technical advance - that could be wrought even by one delinquent individual".[381] Torres speaks about the potential necessity to employ technological

[379] Anders, G. (1956a), p. 45.
[380] Ibid, p. 36.
[381] Rees, M. (2004), p. 70.

bio- and moral enhancement measures in order to manage ‘agential risk’, i.e. in order to make humanity safe for the technological advances lying ahead.[382] Bostrom (2013) speaks about the necessity to introduce global surveillance systems and perhaps even ‘mind-control’ in order to reduce the potential risks associated with emerging technologies,[383] and Oxford ethicists Ingmar Persson and Julian Savulescu argue that, since “the expansion of scientific knowledge and technological prowess will put in the hands of an increasing number of people weapons of mass destruction”, there is an “urgent imperative to enhance the moral character of humanity”. One idea they entertain is that of artificial moral enhancement in form of specific drugs or genetical interventions. This, they argue, might for instance be necessary in order to mitigate climate change. [384]

There could be no clearer illustration of Promethean shame hence – what ‘agential risk’ amounts to is a complete inversion of means and ends. Whereas technology has always been considered to be a means in the service of humanity, now humanity must instrumentalise itself in order to keep the machine going, it turns into, as Anders claims, a gadget for gadgets.[385] Humanity has to adjust itself in accordance with technology in order to make sure that it can function as well as it *potentially* could. The machine *could,* as we have seen, potentially, function perfectly and it is required of the human to adapt physically, politically and in whatever way necessary so that this limitless potential of the machine can be realised. This boomerang effect of technology, as Anders allows us to see, is implicit to concepts such as ‘dual use’ or ‘neutrality’. We begin to look antiquated and unfit for the requirements of our time - for the realities we ourselves have created ostensibly for our own good. At first our imperfections might have seemed merely regrettable to the utopian techno-fetishist, in future, in times of ‘universal unilateralism’, according to existential risk reseachers, they may become *unacceptable*. Our needs have become the needs of technology.

Anders critique of the neutrality thesis thus is based on the insight that it makes us ask the wrong questions. For Anders the relevant question was: What does modern technology do to us simply by virtue of its existence, what does it imply for how we see ourselves and the world around us? What Anders saw was that we are erecting a world in which there is no place for the unmodified human being – no place for human imperfection because we implicitly make projected perfection in the form of technology the benchmark for our interaction with the natural world and ourselves.

This, in turn, opens up new perspectives on the old problem of technology as a rage against the given. Carl Schmitt (1991) argues that “the machine is the tool of utopianism, the weapon of plan realisation” and therefore inherently hostile to “the human”.[386] Heidegger, as we have seen, similarly argues that the technological danger is sustained by the “the willed view that man, by the peaceful release, transformation, storage, and channelling of the energies of physical nature, could render the

[382] Torres, P. (2017a, 2017b, 2017c).
[383] Bostrom, N. (2013), p. 25.
[384] Persson, I. & Savulescu, J. (2008), p. 166.
[385] Anders, G. (1956a), p. 32.
[386] Schmitt, C. (1991), p. 83, note from 01.16.1948. This section was translated by the author of the thesis.

human condition, man's being, tolerable for everybody and happy in all respects",[387] and that "science (and that is modern natural science) is a road to a happier human life".[388] Both authors believed that technological progress was sustained by a perfectionist "revolt against the given", as George Kateb (1997) and Jean-Pierre Dupuy (2009a) put it.[389] Typically, criticisms of techno-perfectionist thinking take the form of rendering it illusionary, i.e. 'utopian' in the everyday sense of the word. The common concern regarding such forms of utopianism is that the pursuit of perfection will not only result in unfulfilled hopes, but that its lack of epistemic humility could lead utopians to do more harm than good along the way.[390] John Gray (2012) for instance argues in that vein that "Instead of enabling humans to improve their lot, science degrades the natural environment in which humans must live. Instead of enabling death to be overcome, it produces ever more powerful technologies of mass destruction. None of this is the fault of science; what it shows is that science is not sorcery. The growth of knowledge enlarges what humans can do. It cannot reprieve them from what they are".[391]

Against the background of existential risk theory, however, the problematic interface between utopianism, science and technology gains a novel dimension and Anders' concept of Promethean shame allows us to see which. If utopianism, naively understood, denotes the desire to perfect the human condition and human nature, in times of existential risk we are confronted with a situation in which humans increasingly can no longer be *allowed* to be what they are, i.e. in which we *need* to be perfected. If existential risk researchers are correct, we are in the process of creating a reality in which we have to be technologically perfected, or at least significantly improved, if we want to survive as a species. The problem is no longer that we *wish* to be reprieved from what we are, but that we *need* to be reprieved from what we are. In Günther Anders' view, too, technological progress is sustained by the desire to overcome chance, contingency and perfect the given. However, he allows us to see the irony of this undertaking - that humanity, in chasing technological perfection, creates realities in which the quest for perfection and the quest for survival can no longer be meaningfully distinguished. This is an extortive form of utopianism, a utopianism where the goal to perfect the human condition may have its roots in melioristic hyperbole but ultimately translates into and is sustained by fear and real necessities which result from this hyperbole. This irony manifests itself in the necessity for humanity to perfect itself because powers have been summoned that are *premised* on the idea of perfection. This turns Heidegger's claim that the technological danger to humanity as a species cannot be separated from the danger to humanity's humanness, into a more

[387] Heidegger, M. (2009), p. 114.
[388] Heidegger, M. (1966), p. 50.
[389] See Kateb, G. (1997), Dupuy, J.-P. (2009a).
[390] The arguably most influential critique of utopianism along such lines was articulated by Karl Popper in his seminal text 'The Open Society and its Enemies' Vol. 1, cf. Popper, K. (2013).
[391] Gray, J. (2012), p. 213.

concrete, graspable problem. As Babich (2013) argues, Anders in these kinds of observations brings Heidegger's philosophy "down to earth".[392]

Against this background it becomes clear why, as Müller (2016) accurately points out, in Anders' view, the enlightenment is not a project that aims at emancipating the human but a project which "represents an active turn away from, and a faltering trust in, everything human".[393] It means that, "as modernity unfolds, we seem to expect ever 'more from technology and less from each other'.[394] We hope, for example to stop global warming by developing new, less carbon-intensive means of energy generation, or, as an ultima ratio, geo-engineering, whilst our expectation that humans will change their increasingly carbon-intensive life-styles, economic growth model, etc., appear to be low, to say the least. Or, as we have seen above, we hope to keep 'universal-unilateralism' in check by means of technological moral enhancement or far-reaching systems of surveillance. But Anders' also tells us why this logic is inherently paradoxical. Someone still needs to develop such miraculous new technologies, including those intended to perfect the human being. Every dream - no matter how far-fetched - about the future capacities of technology simultaneously still is a dream about the future capacities of human ingenuity. Anders therefore describes the attitude characteristic for modern humanity as 'hubristic humility' or 'arrogant self-degradation'.[395] We consider ourselves faulty constructions whilst at the same time self-deifying ourselves when we think that we can recreate and redesign ourselves and perhaps all of earthly nature to perfection.[396] The figure of Prometheus, which according to Anders used to be invoked allegorically by authors from Goethe to Shelley, Ibsen and Sartre in order to describe the hubristic modern human mindset, according to him has lost is allegorical significance: "Our contemporaries", Anders argues, "are certainly still Prometheans, but strangely perverted ones […] They also have presumptuously self-aggrandising ideas of entitlement - but these are so aggrandising that they begin to feel inadequate themselves. They also suffer lacerations – but not because Zeus punishes their high-flying ambitions, but because they chastise themselves on account of their own 'backwardness' and the shame of having been born".[397]

At the bottom of the modern human (technological) predicament Anders hence identified a paradoxical cocktail of dynamics and emotions. We want to liberate ourselves from ourselves, from our limitations and the givens of nature, not noticing that our retreat to artifice in crucial respects produces the opposite effect, exposing us to an ever more problematic extent to our own limitations, reinforcing the hopes we place in our capacities and simultaneously highlighting our faultiness.

[392] Babich (2013a), p. 150.
[393] Müller, C. (2016), p. 102.
[394] Ibid.
[395] Anders, G. (2016), p. 47.
[396] Ibid, p. 50.
[397] Ibid.

This, in turn, tells us something about the high prominence of debates surrounding artificial intelligence in the context of existential risk. As will be discussed in greater detail in chapter 4, artificial intelligence is not only seen as a source of existential risk. By many in the field it is also seen as the potential solution to our problems. The fact that existential risks are threatening to get out of hand and that we increasingly seem to be incapable to control the consequences of technological progress results in the hope that a better problem-solver could be technologically produced. This is Anders' Icarus in a nutshell. It is hoped that the technological project can somehow extricate itself from its faulty human ballast, which would then leave it free to literally conquer the skies, colonising the universe and spreading intelligence on a cosmic level.[398] The idea, again, being that technology as such, if only we get it exactly right, *could* yield perfect results, echoing Heidegger's dictum that "the instrumental conception of technology conditions every attempt to bring man into the right relation with technology. Everything depends on our manipulating technology in a proper manner as a means. We will, as we say, […] master it. The will to mastery will become all the more urgent the more technology threatens to slip human control".[399] The problem of superintelligence, as will be discussed at greater length in the following chapter, indeed appears to collapse the issue of Promethean shame into a single technological problem.

However, apart from teaching us something about the instrumental vision of technology and the ongoing relevance of Anders' thought regarding that vision in the context of existential risk, Anders' concept of Promethean shame gives substance to what I have touched upon in the introduction of Chapter 2, namely how, in existential risk theory, the *terms* of human existence are drawn into 'existential space' and how this begins to transform the human condition and human nature into a technological optimisation problem.

3.2 Technology as action

Ultimately, the reason why the terms of human existence are drawn into 'existential space' is the growing power of our technological tools, which are projected to place ever greater powers in ever more hands. The theme of 'universal unilateralism' can be seen as the defining theme of technological existential risk in general, converging in the fear that human activity, increasingly technologically amplified as it is, has become inherently tied up with existential risk.

Hannah Arendt's thinking turns out to be strikingly prescient in that regard. Arendt sought to show, as Yaqoob (2014) argues, that modern science and technology should best be understood as

[398] Visions like that have been voiced by such prominent figures as Demis Hassabis of Google DeepMind, Stephen Hawking, Elon Musk, Peter Thiel, or Ray Kurzweil. See also Rees, M. (2017) for an in-depth meditation about the cosmic prospects that might await superintelligent human offspring.
[399] Heidegger, M. (1977), p. 17.

part of a “parable about the unpredictability of human action”.[400] One of the things Arendt adopted from Heidegger was the view that modern technology was radically different from the type of human activity that is its etymological root, namely tekné, i.e. the arts and crafts by the means of which humans erect what Arendt calls ‘world’, a concept which will be discussed at greater length below. Contrary to tekné, the purpose of which for Arendt was the creation of tangible tools and objects, such as tables or jewellery, modern technology really was a new type of *action*, namely “action into nature” as well as, action in the form of automation. The emergence of these new forms of action, for her was emblematic of a deep and uncanny transformation of the human condition, a transformation which we find reflected in existential risk theory but the philosophical implications of which remain at large unacknowledged by authors in the field, namely a transformation in which the realms of nature and history were in the process of becoming one. For Arendt who was, as Benhabib (2005) points out, a humanist, this, as we will see below, was an inherently dangerous process because it meant that the spaces which used to provide a stable backdrop for human life were under threat, with humanity exposing itself to entirely new risks, of which manmade threats to the survival of the species were but one.

History and nature become one - or ‘the altered nature of human action’

For Arendt, humanity’s existence used to be bracketed rather comfortably between the two separate realms of nature and human artifice. Nature, for her, were “all processes that come into being without the help of man”, i.e. things that are “not made but grow by themselves into whatever they become”.[401] The natural thing's existence, she elaborates, is not separate but is “somehow identical with the process through which it comes into being.” Human artifice on the other hand must be realised step by step and “the fabrication process is entirely distinct from the existence of the fabricated thing itself”.[402] The hammer or the house is being made with an eye to the finished product and the production process itself is exogenous to that product rather than endogenous. The natural thing, on the other hand, exists *only* as a process, as a becoming thing, where process and thing are undistinguishable. A tree, for instance, is already contained in the seed, the process of growth is part of what ‘tree’ means and when the process stops the ‘tree stops’. Nature, Arendt argues, is the realm of never-ending, automatic processes, whereas the human artifice is the realm of willed beginnings and definite ends.[403]

[400] Yaqoob, W. (2014), p. 205.

[401] Arendt, H. (1998), p. 150.

[402] Ibid.

[403] Arendt’s characterisation of the difference between nature and artifice clearly echoes Heidegger’s distinction between human artifacts and natural objects by reference to different forms of causation as articulated in ‘The Question Concerning Technology’ in Heidegger, M. (1977).

That contradistinction between human artifice and nature is central to Arendt's conception of the human 'world' which was in the process of being lost at the hands of modern science and technology. As Canovan (1995) demonstrates, 'world', in Arendt's thought, is not the natural environment or the totality of everything that exists on and constitutes this planet. 'World' to Arendt meant, on the contrary, the world of human artifice and civilisation, which provides an environment where "instead of an ever-changing natural environment, the man-made world of houses, artifacts and institutions provides a stable background against which individual lives can show up and have significance".[404] When Arendt speaks of world, this can hence be understood in the sense of a 'home' that humanity builds for itself in order to escape from the never-ending, purposeless processes of nature. The most important task of traditional artifice, she argues, is to "offer mortals a dwelling place more permanent and more stable than themselves". [405]

As George Kateb (1997) argues, Arendt was no nature lover.[406] She understood the natural realm as a realm of free-roaming, metabolic processes and the significance of world, in her eyes, consisted precisely in its function to limit and exclude these processes. It provides a backdrop of lasting significance for human existence, within which mortals appear and disappear whilst their products remain. Homo faber, she argues, "the toolmaker, invented tools and implements in order to erect a world".[407]

In Arendt's view, these contrasts between world and nature, artifact and process, had formed a pillar of the human condition and thus of humanity's self-understanding since the very beginnings of civilisation. With the emergence of modern science and technology, however, they were in the process of being blurred by two intertwined trends. First, by a trend towards automation, which was turning the world of human artifice into a world increasingly resembling the world of 'automatic' natural processes: "We call automatic all courses of movement which are self-moving and therefore outside the range of wilful and purposeful interference." Superficially viewed, of course, humanly initiated and sustained automated processes, such as automated industrial production processes, are not 'outside the range of wilful and purposeful interference', as they can be stopped at will. However, as will be discussed in greater detail below, Arendt, like Anders, argued that, for all theoretically and practically relevant purposes, this difference is becoming meaningless. Second, by the growing ability of humanity to "act into nature", by which Arendt means that humanity has acquired the ability to start natural processes entirely on its own, the capacity to wilfully unleash processes into nature which would otherwise not have come into existence.

[404] Canovan, M. (1995), p. 107.

[405] Arendt, H. (1998), p. 152.

[406] Kateb, G. (1997), p. 1243.

[407] Homo faber is Arendt's designation for the modern, Cartesian human subject and thus of the human whose conditions of existence are in the process of disappearing as a result of his successes in exchanging these conditions for something "he himself has made". See Arendt, H. (1998), p. 151.

Arendt, like Heidegger, observed a categorical difference between the ‘ancient arts and crafts’ and modern technology. Arts and crafts were dedicated to erect the world of human artifice - hence a world of permanence, of tangible objects and things, which would shield humanity from the constant changes of nature. It is with tekné that the separation of nature and the human world has become possible in the first place. Homo faber, Arendt argues, “used material as nature yields” and “changed and denaturalized nature for our own worldly ends, so that the human world or artifice on one hand and nature on the other remained two distinctly separate entities”.[408]

Modern technology and science, on the other hand, in Arendt’s view could no longer be accurately described as a “gigantic enlargement and continuation of the old arts and crafts”. With modern technology and science we had begun to “unchain natural processes of our own which would never have happened without us, and instead of carefully surrounding the human artifice with defences against nature's elementary forces, keeping them as far as possible outside the man-made world, we have channelled these forces, along with their elementary power, into the world itself”.[409]

Whilst humanity, for most of its history, had merely imitated and interrupted naturally occurring processes, that is, repurposed what ‘nature yields’ for its own ends,[410] modern humanity, according to Arendt, had begun to unleash natural processes of its own. Even if humanity was not able to ‘make nature’ in the sense of creation, Arendt argues, “we are quite capable of starting new natural processes, and that in a sense therefore we ‘make nature’ to the extent, that is, that we ‘make history’”.[411] Nuclear technology clearly served as Arendt’s prime example here. With nuclear technology she argues, natural forces are let loose that “would never have existed without direct interference of human action”.[412] In the preface to *The Human Condition* Arendt distinguishes between the modern age and the modern world. The modern age, she argues, began in the seventeenth century and came to an end in the beginning of the twentieth century when the modern world was born. The modern world, “was born [politically] with the first atomic explosions”.[413] It is not immediately clear what exactly Arendt was referring to when she speaks of a ‘political birth date’ - if she had in mind simply the birthdate of the atomically supercharged politics of the cold war era or a wider and deeper transformation of politics as a result of the confluence of history and nature, whereby nature, formerly the realm of independent, automatic processes, was drawn into the realm of action through science and technology. Most likely Arendt had both in mind, but in the light of the previous discussion, it seems reasonable to suspect that she saw the former mainly as a (monstrous) symptom of the latter, bringing out the politically problematic implications of the new world, into which humanity had been released by modern science and technology.

[408] Arendt, H. (1998), p. 148.
[409] Ibid, pp. 148-150.
[410] Ibid, p. 148.
[411] Arendt, H. (1958), p. 586.
[412] Ibid, p. 587.
[413] Arendt, H. (1998), p. 6.

Hans Jonas, a fellow student of Heidegger and close friend of both Anders and Arendt, can be read in parallel when he argues in a section of his *Imperative of Responsibility* (1984), entitled ‘'The Universal City as Second Nature', that, owing to technological and scientific progress, all of nature is poised to be drawn into the realm of human responsibility: “The boundary between 'city' and 'nature' has been obliterated: the city of men, once an enclave in the nonhuman world, spreads over the whole of terrestrial nature and usurps its place. The difference between the artificial and the natural has vanished, the natural is swallowed up in the sphere of the artificial”.[414]

In 2016, the Anthropocene Working Group of the Subcommission on Quarternary Stratigraphy proposed to formally adopt the term ‘Anthropocene’ as a geological time unit within the Geological Time Scale, at the same hierarchical level as the Pleistocene of the Holocene, for the present geological interval of the Earth. According to the working group the Anthropocene designates a “period of Earth's history during which humans have a decisive influence on the state, dynamics and future of the Earth system”, including for instance on “the chemical composition of the atmosphere, oceans and soils, with significant anthropogenic perturbations of the cycles of elements such as carbon, nitrogen, phosphorus and various metals”.[415] A recent paper published by authors of the group further holds that humans are the most significant global geomorphological driving force of the 21st Century.[416] There could be no better illustration for the transformation Arendt and Jonas were observing, the blurring lines between the artificial and the natural and the idea of action into nature, than the Anthropocene concept. Against the background of this, it may not be a mere coincidence that the Working Group proposes, as Arendt did, to consider the beginning of the nuclear age as one potential demarcation point for the beginning of this new epoch.[417]

According to Arendt the blurring of boundaries between nature and history was preceded by the blurring of these boundaries in the ontological outlook of humanity: “The modern age […] has led to a situation where man, wherever he goes, encounters only himself. All the processes of the earth and the universe have revealed themselves either as man-made or as potentially man-made”.[418] We know this idea already from Heidegger’s philosophy and his claim that, under the reign of ‘Enframing’, i.e. the ‘technological understanding of being’, nothing is believed to be able to resist the onslaught of calculative thought any longer.

In existential risk theory we find this mindset reflected vividly. In chapter 2 I have argued that, in existential risk theory, every risk, anthropogenic or natural, turns into a technological problem and that therefore, to put it in David Deutsch’s words, the distinction between a natural disaster and one brought about by ignorance has become ‘parochial’.[419] From that perspective, every

[414] Jonas, H. (1984), p. 10.
[415] See Working Group on the ‘Anthropocene’ (2018).
[416] Cooper, A., Brown, T. et al. (2018).
[417] Cf. Steffen, W., Crutzen, P., et al. (2011), p. 843.
[418] Arendt, H. (1961), p. 89.
[419] See Ch. 2, p. 80.

risk, anthropogenic or natural, is first and foremost a risk of ignorance. Hence, when Arendt argues that “wherever we go, we encounter only ourselves”, existential risk theory can not only be seen as a pure version of this ontological shift in perspectives but, with it, we might even update Arendt’s claim to ‘wherever we go, we encounter only our ignorance’.

Combining existential risk theory with Arendt’s and Heidegger’s characterisation of the modern ontological condition leads us back to what Günther Anders considered to be the defining, paradoxical attitude of the modern human being– ‘hubristic humility’ or ‘arrogant self-degradation’.[420] The fact that all the processes of the earth and the universe have revealed themselves either as man-made or as potentially man-made, is inherently self-aggrandising. It places us at the centre of the universe and implies that, in principle, if humanity plays its cards right, we could not only survive for an indefinite amount of time but have virtually infinite amounts of resources - a ‘cosmic endowment’ - at its finger-tips.[421] In that view, nature, Heidegger argues, “becomes a gigantic gasoline station, an energy source for modern technology and industry”.[422] However, the flipside of this mindset is that, due to the fact that wherever we go, we encounter only ourselves, we are also responsible for everything that happens, which means that our ignorance, our faultiness as it were, emerges as the all-defining problem of our existence. We become the eternal saboteurs of our destiny, which, if it were not for our ignorance, *could* potentially be never-ending.

For Arendt, however, the problem was that ignorance is not something that can eventually be overcome, for instance by careful planning and horizon scanning programs. On the contrary, Arendt argued that one effect of our increasing technological powers is that our ignorance on several vital dimensions, for instance our prescience and our ability to ‘understand’ the new world we are living in, necessarily *increases*. To the extent that with the help of technology we have become increasingly successful in acting in accordance with the technological understanding of being, that is, in remaking and repurposing processes of the earth and the universe for human ends, this trend, for Arendt, could only result in less control and less prescience, rather than more.

The reason is that history according Arendt is made up of 'events' and not of the statistically predictable developments of collectives. Events, in Arendt's view, are the spontaneous moments, where history takes the kind of entirely unexpected, unpredictable turns which tend to be remembered and recorded. The occurrence of events are a direct consequence of humanity's plurality and its capability for spontaneous action and thus are a defining component of what it means to live in a human world. In opening up nature as a field for action too we are therefore poised to make it as unpredictable as history: “The reason why we are never able to foretell with certainty the outcome and end of any action is simply that action has no end. The process of a single deed can quite literally

[420] Anders, G. (2016), p. 47.
[421] Torres, P. (2017b), p. 316.
[422] Heidegger, M. (1966), p. 50.

endure throughout time until mankind itself has come to an end".[423] Our ability to act into nature thus means that we are now able to release never-ending processes with unpredictable consequences into nature, just as we used to do in the exclusively human realm.

The implication of the merging of nature and history is that nature becomes as unpredictable as history and therefore that one of the core convictions of the 'Cartesian' or 'Baconian programme', the conviction that science and technology will help us to make nature more controllable, needs to, if not be turned on its head, at least notably qualified. Science and technology do not only make nature more predictable and more controllable they also make it less predictable because our interference with natural processes means that it depends on our action what nature and natural processes will be like in 50, 100, or 1000 years' time. Climate change here can serve as the prime example. Whilst our understanding of climatic processes is arguably better than ever before, allowing us to predict climatic developments more accurately than previous generations could, fossil fuel combustion at the same means that we are acting into the climate system and thereby risk to alter it in unprecedented and unpredictable ways.

What Arendt hence adds to our understanding of existential risk, is that she, as Yaqoob (2014) puts it, embedded her analysis of science and technology within her theory of the "unpredictability of human action".[424] Action, in Arendt's vocabulary means starting things, setting off trains of events into the open, releasing them into time, without being able to foretell their ultimate consequences. With technology turning into a new type of action, the starting of new chains of events in nature, humanity is thus, by definition, exposing itself to a new quality of unpredictability.[425]

Jean Pierre Dupuy for this reason argues that Hannah Arendt "brought out the fundamental paradox of our age: whereas the power of mankind to alter its environment goes on increasing under the stimulus of technological progress, less and less do we find ourselves in a position to control the consequences of our actions".[426] Dupuy here highlights one limit of human understanding in the context of science and technology, which is central to Arendt's thought: the *limits of our prescience* and its dialectical relationship with technological and scientific progress. The relationship between prescience and technology and science is dialectical because science and technology in one sense can be said to increase our knowledge about the future – they extend our knowledge about natural causal processes and thereby allow us to predict their future course far better than previous generations ever could. It is that very ability to predict natural processes which provides the basis of our ability to control the natural environment. At the same time, the very fact that we are increasingly able to predict and therefore control natural processes, in Arendt's view, has propelled us into a position

[423] Arendt, H. (1998), p. 233.

[424] Yaqoob, W. (2014), p. 205.

[425] Canovan, M., In Arendt, H. (1998), p. xvi.

[426] Dupuy, J.-P. (2009a), p. xii.

where we begin to act into nature with the implication that nature is bound to become as unpredictable as history. The blending of history and nature through action thus means that the future, in vital respects, becomes less predictable and less controllable than ever before, even though scientific and technological progress itself, is based on our increased powers of prediction and control over nature. For Dupuy this paradox means that *"The sorcerer's apprentice myth must therefore be updated: it is neither by error nor terror that mankind will be dispossessed of its own creations, but by design— which henceforth is understood to signify not mastery, but non-mastery and out-of-controlness*".[427]

The important lesson for existential risk research is that scientific and technological progress from an Arendtian perspective do not lead to unintended consequences by "terror or error", which is the oft invoked mantra in existential risk theory, but by design – the loss of control and prescience, our lack of understanding of where we are headed, is not the result of a lack of effort, of a lack of reflection or due diligence on the side of engineers, researchers and scientists, it is a necessary consequence of the very project of scientific and technological progress itself: "The dangers of this acting into nature are obvious", Arendt argues, "if we assume that the above mentioned characteristics of human action are part and parcel of the human condition. Unpredictability is not lack of foresight, and no engineering management of human affairs will ever be able to eliminate it, just as no training in prudence can ever lead to the wisdom of knowing what one does".[428]

Understanding technology as action along Arendt's lines provides us with yet another angle on existential risk studies. On the one hand Arendt can be seen as an example for the long history of fears surrounding potentially unpredictable and uncontrollable consequences of technological progress. She was clear that the blending of history and nature through technologically amplified human actions implied an absurd inflation of humanity's powers, propelling us into entirely uncharted waters: "It is beyond doubt that the capacity to act is the most dangerous of all human abilities and possibilities, and it is also beyond doubt that the self-created risks mankind faces today have never been faced before".[429] Arendt hence shared and to an extent even pre-empted many of the anxieties regarding technology in general that permeate in the field of existential risk studies today. On the other hand, against the background of Arendt's characterisation of technological existential risk as part of unpredictability of human action, the macro-strategic ambitions of existential risk research, notions such as 'preferred order of arrival', or 'technological maturity' appear rather naïve and even dangerous.

They seem naïve because, to employ Arendt's terminology, they are expressive of an 'engineering management of human affairs' approach to the problem, rather than seeing it for what it really is, namely a problem of the human capacity for spontaneous action. From Arendt's perspective

[427] Ibid.
[428] Arendt, H. (1958), p. 588.
[429] Arendt, H. (1958), p. 589.

existential risk theory appears to cling on to an anachronistic conception of technology as *tekné*, i.e. along the lines of the old crafts and techniques, which were producing finished products for definite ends, resulting in stable physical objects such as houses, hammers, or tables. This fosters the belief that one day humanity might be able to reach a 'house-like', stable endpoint, a state of perfect technological mastery resulting in some form of equilibrium, the idea being that unpredictability and out-of-controlness are merely characteristics of a transitory phase that can eventually be left behind. This is reflected in some existential risk researchers' aim to reach 'technological maturity', defined as "the attainment of capabilities affording a level of economic productivity and control over nature close to the maximum that could feasibly be achieved".[430] This, from Arendt's perspective, speaks of the mistaken idea that more technological control over nature is somehow equivalent to more control over the future when, in her view, the inverse is the case.

Arendt's characterisation of modern technology as 'action into nature' refocuses our attention on seeing modern technology not as tools that can be put to good or bad use in a more or less stable environment, but as interventions with that very environment, i.e. the never-ending processes of nature, the ultimate outcomes of which can no longer be predicted precisely *because* of humanity's 'control' over them. A close to maximum level of control over nature, from this perspective, would more likely result in a close to maximum level of unpredictability rather than some form of stability. Arendt therefore urges the reader to understand technological risk not in terms of 'unintended consequences', as a problem that can eventually be overcome by careful management but as a feature of a dangerous transformation of the human condition; an essential characteristic of the blurring lines between history and nature.

If anything, choosing an 'engineering management of human affairs' approach, from Arendt's perspective, is *dangerous* because, in fostering the illusion that we might one day reach some sort of stable state with the help of technological mastery, it has us push ahead with the very project that endangers us, expanding our technological capabilities and deepening the trouble we are in. It is furthermore dangerous because in last consequence it can only translate into the kind of boomerang effect which Anders calls Promethean shame – a rage against the perceived faultiness of the human being and the capacity for spontaneous action as such, which are the seeds of all unpredictability. As Arendt argues: "Only total conditioning, that is, the total abolition of action, can ever hope to cope with unpredictability".[431] The above discussed concept of agential risk unequivocally demonstrates that, at least in theory, the struggle against human spontaneity is already underway. But Arendt is clear that the hope to cope with unpredictability through the abolition of action is futile because "even the predictability which comes about through terror can never be sure of its own future".[432]

[430] Bostrom, N. (2013), p. 19.
[431] Arendt, H. (1958), p. 588.
[432] Ibid.

For Arendt, the main take away from that discussion was twofold. First, she was, as existential risk theorists are today, concerned about the dangers associated with action into nature as such: “to act into nature, to carry human unpredictability into a realm where we are confronted with elemental forces which we shall perhaps never be able to control reliably is dangerous enough”.[433] However, what she considered to be even more dangerous was the possibility that the situation might be misinterpreted and that people might “ignore that for the first time in our history the human capacity for action has begun to dominate all others—the capacity for wonder and thought in contemplation no less than the capacities of homo faber and the human animal laborans”.[434] Arendt here touches upon what she, at a different point, calls the ‘fundamental problem of modernity’ - that "man can do, and successfully do, what he cannot comprehend".[435] Günther Anders calls this the ‘Promethean disjunction’ – “the growing gap between what man can do and what man can mentally realise”.[436] The danger both authors saw was that the risks modern humanity found itself confronted with would be fundamentally misinterpreted, that they would be seen not as a problem that is inherent to scientific and technological progress, an inexorable symptom of a fundamental and growing mismatch between different human faculties, but as a problem that could ultimately be solved by the implementation of prudent technocratic measures. What both called for was interpretation, to take a step back and try to understand what the new powers of modern technology and science really spoke of. This is why Anders opens the second volume of *The Outdatedness of the Human Being* with the claim that “it is not enough to change the world, we do this anyway, and it mostly happens without our efforts, regardless. What we have to do is *interpret* these changes so we in turn can change the changes”. [437]Arendt, in the preface to the *Human Condition,* states in almost exactly the same words that the task she set herself was to “to think what we are doing”.[438]

As argued in the introduction, Arendt and Anders, contrary to Heidegger, understood the technological understanding of being as part of the political predicament of the day. Rather than locating the problem outside of human affairs (social, political, economic, or otherwise) as Heidegger had, and succumbing to fatalism, they saw the technological understanding of being as a historically contingent phenomenon and sought to highlight its paradoxes and limitations in order to pinpoint its dangerous role in everyday life and political decision making. Obliviousness to the Promethean disjunction, and the associated, thoughtless, uncritical interaction with technology, for both was expressive of the technological understanding of being and, arguably, the greatest political problem of the day. It meant that people, specifically technologists and policy makers, literally did not know

[433] Arendt, H. (1998), p. 44.
[434] Ibid.
[435] Arendt, H. (1963), p. 4.
[436] Anders, G. (1962), p. 494.
[437] Anders, G. (1992), p. 5.
[438] Arendt, H. (1998), p. 5.

and understand what they were doing and what technology was doing to them, instead pushing ahead, considering technology a mere instrument at the service of humanity, and thus widening the Promethean disjunction to an ever-greater extent. The Promethean disjunction manifests itself in a variety of ways, which will be discussed throughout the following sections. One way in which it manifests itself, was discussed above – it is the fact that 'man' can set off new processes into nature without being able to predict the ultimate consequences of such actions, thus rendering nature less predictable and less controllable than ever before. Another way in which it manifests itself is that, according to Anders, we are incapable to understand and imagine, that is, to mentally realise the *scale and severity* of the potential consequences of our technologically amplified action, which brings us to the problem of human extinction. According to Anders modern humans are 'inverted utopians' : "The basic dilemma of our age is that 'We are smaller than ourselves,' incapable of mentally realizing the realities which we ourselves have produced. Therefore we might call ourselves "inverted Utopians": while ordinary Utopians are unable to actually produce what they are able to visualize, we are unable to visualize what we are actually producing".[439]

3.3 All-or-nothing or the radical evil

Existential risk researchers are united in their conviction that the problem of human extinction warrants far more attention than it currently gets and seek to awaken humanity to the magnitude and nature of the risks we are facing this century. At the time, similar things could have been said about Günther Anders and to an extent also about Hannah Arendt. Anders, in a variety of essays, letters, books, and 'commandments', diagnosed modern humanity with 'apocalypse blindness' and made it his mission to awaken Western citizenry to the full scale of the horrors that loomed in the nuclear arsenals of the US and the USSR.[440]

In what sense did Arendt and Anders take the problem of human extinction seriously? They did not do so in the sense in which existential risk researchers do today, that is in the general, non-domain specific sense characteristic for existential risk theory. But they did try to begin to develop an understanding of what that threat in itself and, by extension, *any* threat to the survival of the human species, amounts to in terms of moral and philosophical weight. That is, they sought to open up to the true severity of what this notion amounts to, which, as we have seen, also forms the basis of existential risk research. But in doing so they went much further than today's existential risk researchers in so far as they went much deeper in trying to pay full heed to the meaning and

[439] Anders, G. (1962), p. 496.
[440] See specifically Anders, G. & Eatherly, C. (1961).

implications of this possibility. They wanted to understand what the at the time new possibility of existential catastrophe meant, what it implies and how it affects us our thinking *today*.

As for existential risk researchers, for Anders and Arendt, the most pertinent conceptual effect of the introduction of nuclear weapons was that it put the entire future of humanity at risk. However, whilst existential risk researchers leave it there and simply posit that this amounts to the absolute bad and that existential risk reduction should become a global priority, Arendt and Anders set out to ponder about the wider philosophical implications of that threat. As Robert Jungk argues in the foreword to Anders' *Burning Conscience*: "Since 1945, millions of words have been written by eminent authorities on 'the effects of nuclear weapons'. Nevertheless, there is still a yawning gap in this comprehensive volume of literature on the subject. The experts, it is true, have subjected whole mountains of ruins and tens of thousands of survivors to most minute examination, but in their meticulous studies, they omitted one very important object-themselves; and by so doing they have disregarded an aspect of decisive importance, namely, that atom bombs strike back at those who use them and, indeed, at those who earnestly labour at making their use possible".[441] In other words, what is of interest here, is Anders' and Arendt's analysis of how 'the bomb strikes back', how it affects humanity here and now and what this can tell us about existential risk theory. This 'phenomenology of the end-time',[442] is an integral component of what Arendt and Anders considered thinking about 'what we are doing', about the Promethean realities we are creating and the inhuman worlds we are erecting around us.

For Anders and Arendt, the fact that nuclear weapons put the entire future of humanity at risk implied a major shift in our thinking about how we think about humanity - it changed our perspective on the future from one in which the future existence of humans was taken for granted, i.e. in which humanity was thought of as effectively immortal, to one in which humanity had to be thought of as mortal. According to Anders, apart "from a handful of natural philosophers, and Christian thinkers", it did not even occur to thinkers of previous centuries to seriously entertain thoughts about human extinction.[443] In his view, it therefore was only with the introduction of

[441] Jungk, R. (1961), p. xi, in Anders, G. (1961).

[442] Babich, B. (2013a), p. 152.

[443] See Anders, G. (1956a), p. 241. In *Ketzereien* (1982), Anders mentions Immanuel Kant as an example for an earlier thinker who reflected about the possibility of human extinction. In 1763, in a brief treatise entitled 'Beweisgrund zu einer Demonstration des Daseins Gottes' Kant discusses Isaac Newton's calculations about the possibility that the solar-system (and by implication humanity) will 'disappear' in the distant future as a result of the sun burning out. Anders describes the surprise he felt upon encountering Kant's unimpressed reaction to this prospect. Whereas Newton had held, according to Anders, that God would intervene and rescue humanity, Kant claimed that the disappearance of humanity "ought not to be seen as a regretful loss, since we do not know how unmeasurably rich the continuously evolving nature in other corners of the sky might be. Its vast fertility shall compensate for this loss with great abundance elsewhere." Kant here appears to suggest that humanity has no intrinsic value, otherwise he certainly would have considered its disappearance a 'regretful event'. Anders was baffled by this complete lack of "anthropo-, geo-, and even helio-centrism" in Kant's assertion, arguing that he had not encountered anything comparable in any other author before or after Kant: "To encounter this kind of renouncement in the writings of the greatest moralist of all times, a man of whom one would expect that his reverence for moral law should lead him to regard the existence of humanity as

nuclear weapons that the old dogmas, “all humans are mortal”, and “all humans can be killed”, needed to be updated to “humanity as a whole is mortal” and even to “humanity as a whole can be killed”.[444] This new fact, of having to consider humanity as *killable,* for both Arendt and Anders, was an enormous conceptual shock.

Before discussing this shock and its repercussions within the authors’ work in greater detail there is of course a question to be asked about the historical accuracy of their account of past perspectives on the future. It is clear that apocalyptic thinking, eschatology, is a defining feature of Judeo-Christian thought. The historian of time Reinhard Koselleck (2004) for instance argued that “until well into the sixteenth century, the history of Christianity is a history of expectations, or more exactly, the constant anticipation of the End of the World on the one hand and the continual deferment of the End on the other. While the materiality of such expectations varied from one situation to another, the basic figure of the End remained constant”.[445] Further, it has been demonstrated that eschatological thinking has not only survived the Enlightenment but profoundly shaped Enlightenment thinking itself and persisted deep into the modern Zeitgeist, leaving its mark in ostensibly secular concepts such as ‘progress’, ‘revolution’, and even the concept of ‘modernity’ itself.[446] According to Derrida (1984), apocalyptic thinking might even be *the* characteristic feature of the entire tradition of European thought.[447] It is therefore not immediately clear exactly whose past perspective on the future Arendt and Anders were referring to when they claimed that the future existence of humanity used to be taken for granted, nor how representative for the temporality of the wider public they believed their claim to be. My aim here, however, is not in the main to establish the historical accuracy of Arendt’s and Anders’ account of past futures.

For, there is a case to be made that even if we allow for the fact that religious forms of apocalypticism were and perhaps tacitly continue to be a defining feature of occidental thought, this does not necessarily sit problematically with the above authors’ main argument. The reason is that in apocalypticism as a religious concept the future of humanity *effectively*, i.e. in all practically relevant

indispensable, is curious to say the least”. See Anders, G. (1982), p. 157. All of above quotations have been translated by the author.

[444] Anders, G. (1956a). p. 243.

[445] Koselleck, R. (2004), p. 11.

[446] Gray, J. (2007).

[447] Derrida in a piece entitled ‘Of an Apocalyptic Tone Newly Adopted in Philosophy’ argues that European thinking is characterised by the constant pursuit of apocalypse, which he understands as the pursuit of the ultimate truth and the final word: “Haven't all the differences [author’s note: differences between truth-claims and schools of thought, etc.] taken the form of a going-one better in eschatological eloquence, each newcomer, more lucid than the other, more vigilant and more prodigal too, coming to add more to it: I tell you this in truth; this is not only the end of this here but also and first of that there, the end of history, the end of the class struggle, the end of philosophy, the death of God, the end of religions, the end of Christianity and morals (that was the most serious naivete), the end of the subject, the end of man, the end of the West, the end of Oedipus, the end of the earth, Apocalypse Now … And whoever would come to refine, to say the finally final (le fin du fin) namely the end of the end (la fin de la fin) the end of ends, that the end has always already begun, that we must still distinguish between closure and end, that person would, whether wanting to or not, participate in the concert. For it is also the end of metalanguage on the subject of eschatological language. With the result that we can wonder if eschatology is a tone, or even the voice itself”. See Derrida, J. (1984a).

respects, too, can be said to be taken for granted. As for instance Himmelfarb (2010) argues, the apocalypse in Judeo-Christian tradition is associated with "the Last Judgment and cataclysmic end of the world but also reward and punishment *after* death, the heavenly temple, the divine throne room, and astronomical phenomena and other secrets of nature".[448] In other words, the apocalypse is envisioned as a moment of absolute transcendence, a cataclysmic instance of revelation in which everything is falling into place and humanity enters a higher form of existence (this is also the original meaning of the Greek term *apokalypsis,* which translates into 'dis-covery' or 'exposing'[449]). From this perspective, the notion of apocalypse does not contain a theory of the mortality of humanity, but rather a theory of its transformation. It does not present an image of humanity as mortal but hinges, on the contrary, on the notion of a collective afterlife and thus the ongoing existence of humanity in one form or another. Apocalypse as a religious concept, to put it briefly, presents us with a continuation of the story of humanity, rather than with one about its end.

This brings us to what Arendt and Anders considered to be so profoundly shocking about the nuclear bomb. If enlightenment, or 'the death of God', implied the end of the immortality of humans on a spiritual or transcendental level, the nuclear bomb implied the end of the immortality of humans in this world too. This effect of the nuclear bomb, according to them, catapulted humanity into genuinely uncharted philosophical terrain. Of course, apocalypse in the traditional sense would also mean an end of humanity as we know it. However, the difference between apocalypse as a religious concept and apocalypse as a secular concept is that the one bestows activities with meaning (be it that their meaning resides in averting the wrath of God or in hastening the arrival of the final day of judgement, or simply to present oneself as worthy in the eyes of God upon the day of its arrival), whilst, according to Arendt and Anders the other has the opposite effect – the mere possibility of this-worldly total doom threatens to deprive everything we do of meaning, irrespective of whether or not it is actually going to happen anytime soon. The one presents the culmination point of history, something that will in effect be the final confirmation of the value and meaningfulness (or lack thereof) of our collective and individual endeavours, whilst the other undermines what is, according to Arendt and Anders an unacknowledged condition for our very ability to make sense of our lives, namely that we think of ourselves as part of the "enduring chronicle of mankind".[450] From this perspective the intuitively attractive likening of existential catastrophe to apocalypse and of existential risk theory to eschatology is fundamentally misleading.[451] In fact, it would make much more sense to consider existential catastrophe an *anti*-apocalypse. Rather than unveiling the ultimate truths and meaning of our existence, enlightening it in all respects, it means ultimate darkness - the

[448] Himmelfarb, M. (2010), p. 2, italics added by the author.

[449] See Groĭs, B. (2012), p. 72.

[450] Arendt, H. (1994), pp. 421-422.

[451] Munthe, C. (2015) for instance likens the logic of existential risk theory to that of Pascal's wager, asking why existential risk researchers do not all attend mass. Whilst the concerns might be similar in structure (both concern events with low probability but arguably infinite impact) they are nonetheless very different in nature.

end of the very possibility of finding truth and meaning in life. It is in that sense that Anders claims that "if the mankind of today is killed, then that which *has* been dies with it; and the mankind to come too. The door in front of us bears the inscription: 'Nothing will have been'; and from within: 'Time was an episode'. Not, however, as our ancestors had hoped, an episode between two eternities; but one between two nothingnesses; between the nothingness of that which, remembered by no one, will have been as though it had never been, and the nothingness of that which will never be. And as there will be no one to tell one nothingness from the other, they will melt into one single nothingness. This, then, is the completely new, the apocalyptic kind of temporality, our temporality, compared with which everything we had called 'temporal' has become a bagatelle".[452] Or, as Hans Jonas argued: "Now we shiver in the nakedness of a nihilism in which near-omnipotence is paired with near-emptiness".[453] Groys (2008) in that vein refers to the possibility of nuclear war as 'the apocalypse of apocalypse', arguing that it "destroys everything without uncovering any kind of truth and without leaving behind any kind of reality".[454] The introduction of nuclear weapons hence meant that even the last safe haven of the very idea of value and purpose was imperilled, if not obliterated outright.

The categorical shift in our thinking about the future of humanity, from one in which we could (albeit perhaps mistakenly as existential risk research points out by highlighting the inescapability of natural extinction events in the long run) posit its immortality to one were we suddenly had to think of it as mortal, according to Anders and Arendt challenged the moral and ethical foundations of our existence and opened up, at bottom, the problem of nihilism - the possibility of absolute nothingness, extending right into the present. Hans Jonas made a highly similar remark, arguing that "the presence of man in the world had been a first and unquestionable given, from which all idea of human obligation in human conduct started out. Now it has itself become an object of obligation: the obligation namely to ensure the very premise of all obligation, that is the foothold for a moral universe in the physical universe - the existence of mere candidates for a moral order".[455]

However, given that the possibility of human extinction had not featured as a matter of sustained philosophical reflection before the 20th century - simply because it did not need to – the philosophical implications of this shift in perspectives, specifically for our thinking about the meaningfulness of human conduct, too, were a rather underdeveloped topic. Of the authors covered here the arguably clearest account of how the assumption that human life is an open-ended continuum served as a tacit precondition for how we attribute meaning to our lives can be found in Arendt's work.

[452] Anders, G. (1961), p. 11.

[453] Ibid, p. 23.

[454] As argued above, for Derrida we are in a constant state of apocalypse, always thinking of ourselves as occupying a position at the end of revelation, possessing final truths. For an authoritative analysis of the theme of apocalypse in Derrida's thought see Groïs, B. (2012), p. 69 ff.

[455] See Jonas, H. (1984), p. 10.

As discussed in section 2 of this chapter, in Arendt's thought 'world' can be seen as a bridge between the mortality of the human individual and the immortality of the human species. This notion provides us with a basis for reflecting about how the tacit assumption that the human species is immortal used to condition human life on an individual and collective level. Referring to the ancient Greek conceptions of mortality and immortality, Arendt argues that mortality, had come to be understood as the quintessentially human property, 'the hallmark of human existence': "Men are 'the mortals', the only mortal things there are for animals exist only as members of their species and not as individuals. The mortality of man lies in the fact that individual life, a *zoe* with a recognizable life-story from birth to death, rises out of biological life, *bios*. This individual life is distinguished from all other things by the rectilinear line of its movement, which, so to speak, cuts through the circular movements of biological life. This is mortality: to move along a rectilinear line in a universe where everything, if it moves at all, moves in a cyclical order".[456]

In other words what makes humans mortals is their individuality, the fact that, contrary to animals, they lead lives that differ from one another in recognizable ways to the effect that the otherwise uniform circularity of natural life is interrupted. Because gods are immortal, and animals, in her view, exist only as members of their species, as part of bios, mortality emerged as the distinctive feature of the human individual: "embedded in a cosmos in which everything was immortal, it was mortality which became the hallmark of human existence: Men are 'the mortals,' the only mortal things there are".[457]

The human world on the other hand, Arendt claims, was understood to be immortal.[458] World is the realm created by human work, into which individuals are born and from which they disappear, whilst their works in form of arts, culture, institutions, buildings, etc., remain, forming a continuum and providing mortals with a stable backdrop for their endeavours. As Canovan (1998) points out, at the heart of Arendt's "analysis of the human condition is the vital importance for civilized existence of a durable human world, built upon the earth to shield us against natural processes and provide a stable setting for our mortal lives".[459] The mortality of the individual, on the other hand, in Arendt's view, ensures that the human world is one of constant beginnings, of change and fresh ideas; it ensures that new individuals can leave their mark in the world and obtain, through their works and deeds, a share of immortality. What Arendt calls 'world' can hence be seen as the material and ideational bridge between individual mortality and collective immortality. Individual mortality would be much harder to come to terms with without the presumed immortality of our world and the world would become stagnant, without new persons, unique in their individuality, leaving their mark in it and changing its course.

[456] Arendt, H. (2000), p. 279.
[457] Arendt, H. (1958), p. 571.
[458] Arendt, H. (2000), p. 278.
[459] Canovan, M. (1998), p. xiii.

Once we insert nuclear weapons into this complex it becomes immediately clear why Arendt regarded them with a horror that went beyond the vague horror that everyone is likely to feel who dares to think seriously about the spectre of nuclear conflict. Nuclear weapons put what Arendt called 'world' at risk, they not only loomed large over our heads as the ever-present possibility of unfathomable levels of suffering and loss of future utility, but their mere existence affected us by rendering outdated one of the conditions which made us human in the sense in which we, according to her, used to understand that notion, namely as mortal beings in an immortal human world. Nuclear weapons by virtue of their mere existence, irrespective of whether or not they would ever be utilised, uproot us in time and, by implication, affect our idea of ourselves. This is why Günther Anders argues that the mere existence of nuclear weapons is a form of action. He repeated this time and again, almost like a mantra: "the nuclear bomb is not a means to an end", it is an act qua existence.[460] The immortality of the human world became a thing of the past once the end of human life became mere possibility with the appearance of nuclear weapons.

If we now return to Hans Jonas' claim that the threat to the ongoing existence of humanity imperils the very idea of human obligation in human conduct, Arendt's conception of world provides this claim with some more substance. Obligation, in so far as it is deeply connected with what Arendt calls world, has an inherently temporal dimension, relying on a temporally extended sense of commitment. Recently, Samuel Scheffler, writing in the analytical tradition, made the exact same claim, arguing that humans rely on the existence of future generations for leading value laden lives and that present generations therefore have egoistic reasons to secure the continued existence of the species.[461] This is intended to highlight that one must not necessarily care about the existence of future generations in Parfit's or Singer's pan-generational utilitarian way in order to care about the survival of the species but that there are reasons to care for the future out of a concern for the *present*.

[460] See for instance Anders, G. (1956a), p. 247.

[461] Scheffler presents an intriguing argument based on two thought experiments in which he asks the readers to imagine that they were confronted with two different end-time scenarios: In the first case he asks us to imagine that we are confronted with the information that 30 years after our death an asteroid will collide with Earth and annihilate all life on the planet. In the second case he asks us to imagine a scenario in which humanity would become infertile so that, without anyone dying a premature and/or violent death, humanity would go extinct within a couple of decades. He arrives at the conclusion that this would lead to a widespread feeling of meaninglessness because in many ways our ability to lead meaningful lives and even our ability to attribute value to not straightforwardly instrumental activities, such as "reading 'The Catcher in the Rye' or trying to understand quantum mechanics", depends on our confidence that there will be future human generations, or, as he calls it, "a collective afterlife". Without this confidence, he argues, our conception of a "life as a whole" would break down. Scheffler even suggests that "we cannot assume that we know what the constituents of a good life would be in such a world, nor can we even be confident that there is something that we would be prepared to count as a good life". Crucially, he does not suggest that we need to have confidence that humanity will exist "forever". Rather, Scheffler argues, we need to have confidence that humanity will not cease to exist for a "considerable" amount of time after we die. He infers from that that our confidence as individuals in the value of many of our activities depends to a large extend on our confidence that there will be future generations. Scheffler concludes that it is part of what it means to be a human being to think of oneself as part of a continuum stretching far into the future: "Our values express our own understanding of ourselves as temporally extended creatures with commitments that endure through the flux of daily experience". See, Scheffler, S. (2016).

Economist Partha Dasgupta of CSER in a recent working paper made a point along comparable lines when he argues in relation to art and other cultural products, including ideas and institutions, that "future people add value to the creators' lives by making their creations durable. Here the fact of a general assumption that people desire to have children is significant. An artist may regard his work to be far more important than parenting, but he is helped by the presumption that there will be future generations to bestow durability to his work".[462] Dasgupta goes even one step further, arguing that one can understand procreation as a means of making one's values and practices survive, which means that not only values that are part of the public realm tacitly depend on our assumption that there will be future generations but that "many are private, even confined to the family, and it is important to us that they are passed down the generations. Procreation is a means of making one's values and practices durable. We imbue our children with values we cherish and teach them the practices we believe are right not merely because we think it is good for them, but also because we desire to see our values and practices survive".[463] We thus see that ideas which Arendt entertained about 60 years ago in the wake of the nuclear arms race are now beginning to be taken up in altered form and different vocabulary by analytical philosophers and economists in the context of existential risk.

As argued in chapter 1, what existential risk theorists originally want to raise attention to is the 1 per cent in Derek Parfit's thought experiment, arguing that the stakes symbolised by this one per cent are all too often underappreciated. Arendt and Anders clearly agree that the 1 per cent transform a threat into one of an entirely different category as compared to all other kinds of catastrophes and disasters we have experienced so far. Yet they went further than existential risk researchers do today - they argued that the stakes involved are so vast indeed that they changed the very nature of what it means to be human and of what it *can* mean to be human *in the present*. What is immediately threatened by existential risk is not only future generations, future utility, but ourselves, irrespective of whether an existential catastrophe is actually ever going to materialise or not.[464]

This leads us back to the problem of action in the context of existential risk. As argued before, for Arendt the modern world was born, *politically*, with the first atomic explosions. The fact that human action, by attaining the capability to set off new chains of events into nature had also attained the capability to annihilate life on Earth was one reason why Arendt saw the conditions of

[462] Dasgupta, P. (2017), pp. 38-39.

[463] Ibid, p. 39.

[464] A similar argument can be found in Dupuy, J.-P. (2012). Dupuy suggests with reference to Anders that for the above reasons the West's outlook on the future of humanity should be inverted, from one in which the future depends on us, and we are responsible for future generations to one in which we consider ourselves to be dependent on the future: "Whether or not the future has any need of us, we, for our part, need the future, for it is the future that gives meaning to everything we do". See Dupuy, J.-P. (2012), p. 11. Robert Jungk (1961) in his foreword to Anders' *Burning Conscience,* also makes that point when he argues "that atom bombs strike back at those who use them and, indeed, at those who earnestly labour at making their use possible… Under the weight of them, the very foundations of our moral and political existence are collapsing." See Anders, G. (1961), p. xi.

our existence transformed to such an extent that, to her, we had begun to occupy an entirely new world, of which she was not sure how to make sense. In the context of politics, it meant that humanity found itself confronted with what she calls the 'radical evil' and therewith with something that according to her should never be involved in politics as we used to understand it, namely all-or-nothing questions:

> "It is the appearance of some radical evil, previously unknown to us, that puts an end to the notion of developments and transformations of qualities. Here, there are neither political nor historical nor simply moral standards but, at the most, the realisation, that something seems to be involved in modern politics that actually should never be involved in politics as we used to understand it, namely all or nothing - All, and that is an undetermined infinity of forms of human living-together, or nothing, for a victory of the concentration camp system would mean the same inexorable doom for human beings as the use of the hydrogen bomb for the human race".[465]

In other words, atomic weapons, by turning the continued existence of the human species into a matter of political choice, had led to such a grotesque over-inflation of political power that, in Arendt's view, politics itself was deformed beyond recognition, rendering our habitual understanding of it anachronistic.

Unfortunately Arendt typically remains rather vague in her remarks about the radical evil and all-or-nothing questions, and how, in her view, it transformed politics, despite the fact that she repeatedly hints at how central that observation was to her thought, arguing that the Human Condition was written against the background of that circumstance, [466] and that "the whole political and moral vocabulary in which we are accustomed to discuss" matters such as violence, peace, war or courage had been rendered practically meaningless with the appearance of nuclear weapons.[467]

The phrases 'transformation of qualities' and 'infinity of forms of human living-together', however, provide us with an indication of why politics 'as we used to understand it', in her view, had come to an end once infused with the radical evil of all-or-nothing questions. Politics, just as any other aspect of human life, in Arendt's view hinged on the tacit assumption that the human species is immortal, it used to be seen as part and parcel of an open-ended transformation of qualities and reorganisation of forms of 'human living-together', negotiating and renegotiating human terms of coexistence, and *not* their arbitrator. This perspective on the future is reflected in categories based on which we habitually think about political developments, as 'progress', 'regress', or 'modernisation'. As Margaret Canovan (1995) points out, Arendt did not believe in the idea progress to begin with, she "did not share the barely conscious assumption of modern publics that with the growth of

[465] Arendt (1973), p. 443.

[466] Ibid.

[467] Arendt, H. (1994), p. 421. In *On Revolution* Arendt makes a similar point. She here discusses the transformation of warfare under modern conditions, i.e. under nuclear conditions, and argues that "seventeen years after Hiroshima, our technical mastery of the means of destruction is fast approaching the point where all non-technical factors in warfare, such as troop morale, strategy, general competence, and even sheer chance, are completely eliminated so that results can be calculated with perfect precision in advance." See Arendt, H. (1990), p. 17.

prosperity and enlightenment, and in spite of setbacks on the way ranging from concentration camps and nuclear weapons to mounting crime-rates and international terrorism, we are somehow moving toward a world in which violence will no longer exist".[468] Irrespective of whether or not Arendt believed that such assumptions had a basis at any point in time, it is clear that, in her view, the appearance of nuclear weapons rendered them obsolete once and for all:

> "the fearful imagination has the great advantage to dissolve the sophistic-dialectical interpretations of politics which are all based on the superstition that something good might result from evil. Such dialectical acrobatics had at least a semblance of justification so long as the worst that man could inflict upon man was murder. But, as we know today, murder is only a limited evil, the murderer who kills a man - a man who has to die anyway - still moves within the realm of life and death familiar to us; both have indeed a necessary connection on which the dialectic is founded, even if it is not always conscious of it. The murderer leaves a corpse behind and does not pretend that his victim has never existed; if he wipes out any traces, they are those of his own identity, and not the memory and grief of the persons who loved his victim; he destroys a life, but he does not destroy the fact of existence itself".[469]

This quote makes it very clear that, in Arendt's view, the infusion of politics with the absolute of all-or-nothing questions has catapulted us into an entirely new reality in which our inherited 'superstitions' and intuitions about politics were rendered obsolete and even profoundly misleading. Conceiving of nuclear weapons as 'set-backs', as in the above quoted passage, means thinking in terms of 'dialectic acrobatics' and thus to try and fit them into political categories which nuclear weapons, by mere virtue of their existence, render obsolete. Speaking of set-backs makes sense only if one can take an open-ended transformation of qualities for granted. Nuclear weapons, however, put an end to that very notion. The true horror of nuclear weapons is that their presence transcends the 'realm of life and death familiar to us' by threatening to 'destroy the fact of existence itself'. Their mere existence deprives humanity of the felt solidity and permanence of the world which used to give structure and meaning to our pursuits as mortals, political and otherwise. Whether or not they are ever utilised, nuclear weapons are the 'radical evil' because they have catapulted us into an *unhuman* reality – unhuman because the all-or-nothing questions they confront us with transcend the very parameters and conditions of our existence based on which we have made sense on what it *means* to be human.

One instance where Arendt's work is more concrete in pinpointing the implications arising from this shift in perspectives for political life is in a brief essay entitled '*Europe and the Atom Bomb*' from 1954. She here, amongst other things, discusses the mindset underlying the idea that "it is better to be dead than a slave",[470] encapsulated in the, at the time common, battle cry 'rather dead than red'. Arendt argues that this conviction implicitly appeals to the political virtue of courage and demonstrates that, with the appearance of atomic weapons, this appeal has become "all but

[468] Canovan, M. (1995), p. 185.
[469] Arendt, H. (1973), p. 443.
[470] Arendt, H. (1994), p. 421.

meaningless".[471] Courage, according to Arendt, used to be defined by two main pillars of the human condition, the mortality of the individual and the immortality of the species. In order to be courageous, Arendt argues, humans must be sure both of their own mortality as well as of the existence of a posterity, i.e. the immortality of the species. If humans were immortal, if life were not bound to be taken from us one day anyhow, she argues, we could never mount the courage to risk it, because the stakes would be so high that the kind of courage required would be literally 'inhuman'. Courage, Arendt thus claims, in the world of the ancients was the only virtue that was reserved exclusively for humans, the mortals, and was denied to the gods, the immortals. If humans were to become immortal, Arendt henceforth argues, life would no longer only be our highest good, as it is today, but become our central concern, overruling all other considerations.[472] Harrari (2017) makes a similar point when he suggests that, if the present Silicon Valley quest for immortality were to be successful, it would most likely turn its beneficiaries into "the most anxious people in history".[473] The resulting mode of human existence, Harrari argues, would not be that of immortality but that of a-mortality. Humans would not become immortal in the sense that death would become an impossibility, they would merely have a *potentially* infinite life-span. That is, humans could theoretically live forever under the condition that their existence is not ended by accident or choice. In Arendt's vocabulary, an individual with technologically attained a-mortality would not turn into a god, it would still move within the realm of life and death familiar to us in that its mode of existence would still be defined in relation to its opposite, i.e. non-existence. Rather than freeing us from our preoccupation with the problem of finitude, the quest for immortality hence might result in the opposite, our preoccupation with finitude could turn into a nightmare.

The second condition of courage is that humans are convinced of the existence of a posterity which will "understand, remember and respect" their sacrifice: "Man can be courageous only as long as he knows that he is survived by those who are like him, that he fulfils a role in something more permanent than himself, the 'enduring chronicle of mankind'."[474] What Arendt demontrates here, is that our traditional idea of courage is deeply entangled with tacit assumptions regarding the mortality of the individual on the one hand as well as the permanence of what she calls world on the other hand, and thus temporally extended commitments.

[471] Ibid.

[472] Arendt (1994), p. 421. 'Life' here needs to be understood in the specific sense in which Arendt uses the term, i.e. following the Aristotelian distinction between zoë and bios. 'Life', in Arendt's terminology, refers to bios, to 'bare life'. 'Concern for life' by implication comprises all activities concerned exclusively with 'staying alive', with keeping the metabolic processes that allow us to endure as living creatures on Earth, going. Human life for Arendt was different from animal life precisely in that it went beyond bios. Human life, for Arendt, was embedded in world and history and thus characterised with the activities associated with these two realms, i.e. work and action; activities, that is, that are not purely concerned with the necessities of 'staying alive'. Hence, when Arendt feared that 'life' could become humanity's only concern, what she fears is that our concern for bios could become so towering that the conditions under which make a truly human life in the sense of zoë, work and action, would be undermined.

[473] Harrari, Y. (2016), p. 29.

[474] Arendt, H. (1994), p. 422.

All or nothing questions thus sit uneasily with the political virtue of courage as understood by Arendt. On the one hand, she argues, the possibility of total annihilation introduced by modern warfare transforms the individual into a 'conscious member of the human race', for "whose survival he must care more than for anything else".[475] On the other hand, humans need to be sure of a posterity in order to act courageously at all. By undermining the conviction that there will be a posterity, the political virtue of courage is thus at risk to lose its basis at the very point in time, at which courageous action would arguably be needed the most.

This argument clearly is most compelling when placed against the backdrop of the cold war context of the time of Arendt's writing. However, the basic insight remains relevant in the context of existential risk theory. If we can no longer conceive of ourselves, of our actions and commitments innocently as part of an ongoing transformation of qualities, we end up in an unprecedented, circular situation where we must secure the conditions under which we habitually attribute meaning to our endeavours - a meta-ethically problematic situation to be in, as Scheffler demonstrates. Politically arguably more problematically still, it means that our intuitions, the concepts and categories by the means of which we habitually think about political matters might no longer map onto reality and therefore might in fact misguide us. What Arendt's discussion of courage authoritatively shows is that in our everyday interactions with modern technology we are prone to rely on categories and concepts which are rendered anachronistic by the very technologies we are trying to make sense of in such terms; a circumstance which is particularly pronounced and dangerous in the context of nuclear weapons.

This concern occupies an arguably even more important role in Anders' thought. For Anders, the advent of nuclear weapons had transformed the parameters of humanity's existence to such an extent that we could no longer be considered 'human' at all. The introduction of nuclear weapons meant, Anders claims, that all vicissitudes and changes of history have been reduced to the status of a prelude, of mere pre-history and that "we are not merely representatives of a new historical generation of humans but, […] because of our radically changed relation to the cosmos and ourselves, creatures of a new species".[476] The generation of his parents, Anders states, had been the "last humans" and everything that "had been valid for them", had "become invalid for us […] their

[475] Ibid.

[476] Anders, G. (1956a), pp. 239-240. This is an abbreviated version of Anders' original text, which was translated by the author. The original text reads as follows: "Da wir die Macht besitzen, einander ein Ende zu bereiten, sind wir die Herren der Apokalypse. Das Unendliche sind wir. – Das sagt sich leicht. Ist aber so ungeheuerlich, daß alle Wechselfälle der bisherigen Geschichte daneben beiläufig zu werden scheinen, und die bisherigen Epochen zur bloßen ‚Vorgeschichte' zusammenzuschrumpfen scheinen. Denn wir sind nun nicht einfach nur Vertreter einer neuen geschichtlichen Generation von Menschen, sondern, obwohl anatomisch natürlich völlig unverändert, durch unsere völlig veränderte Stellung im Kosmos und zu uns selbst, Wesen einer neuen Spezies; Wesen die sich vom bisherigen Typus ‚Mensch' nicht weniger unterscheiden, als sich etwa, in Nietzsches Augen, der Übermensch vom Menschen unterschieden hätte."

dearest emotions have become alien to us; and the juxtapositions by the means of which they understood themselves and articulated their Being, have become inapplicable".[477]

In other words, like Arendt, Anders claims that the appearance of the radical evil in form of nuclear weapons had created a rift, separating the living generations from all generations that preceded them. When Arendt argues that it put an end to politics as we used to understand it, Anders goes even further than that in claiming that it put an end to humanity as we used to understand it: "we are no longer what until today men have called 'men'".[478] The negative omnipotence that came with nuclear weapons had transformed the living generations into "titans", "lords of the apocalypse", who were in need of an entirely new conceptual repertoire to make sense of their situation in the cosmos and of themselves.

Just as mounting the courage to sacrifice one's own life would be inhuman if humans were a-mortal, the infusion of politics with the radical evil means that we are confronted with an inhuman problem. The fact that they threaten to annihilate existence itself means that we are confronted with a problem outside of the realm of life and death familiar to us, which used to be the very bracket for our decision-making, political and otherwise. That's why Anders claims that "it is misleading to say that atomic weapons exist in our political situation. This statement has to be turned upside down in order to become true. As the situation today is determined and defined exclusively by the existence of 'atomic weapons,' we have to state: political actions and developments are taking place within the atomic situation".[479]

The underlying problem both Anders and Arendt identified was that modern technology, epitomised by nuclear weapons technology, by design, not by virtue of the uses we make of it, transcends conditions on which we rely to make sense of our situation and ourselves. They were convinced that, if we do not systematically challenge ourselves to imagine, think, and interpret what we are doing, our interaction with technology is at risk to become inherently *thoughtless* in the sense that our thoughts would no longer correspond with the realities we are producing. Arendt and Anders set out to do just that, to interpret and reveal the new, technologically defined reality to their contemporaries and to awaken them to their titanic predicament. This leads us to what Arendt and Anders, in a very Heideggerian fashion, considered to be the most basic problem associated with scientific and technological progress.

[477] Cf. Anders, G. (1956a), p. 240. This passage has been translated by the author. The original text reads as follows: "Das Wichtigste, was von unseren Eltern, den ‚letzten Menschen', gegolten hatte, ist für uns Söhne, die ‚ersten Titanen', ungültig geworden; ihre liebsten Gefühle sind uns bereits fremd; und die Alternativen, mit deren Hilfe sie sich verstanden und ihr Dasein artikuliert hatten, schon außer Kurs".

[478] Anders, G. (1956b), p. 146.

[479] Anders, G. (1962), p. 494.

3.4 The schizophrenic condition of modern existence

According to Arendt and Anders the core problem of the Promethean disjunction is that, despite the fact that the above described transformative implications associated with modern technology are unfolding in plain sight, we appear to be not only oblivious to them but unable to grasp them. We cannot comprehend the uprooting and alienating dynamics of modern machine amplified action: "man can do, and successfully do, what he cannot comprehend and cannot express in everyday language".[480] In other words, we appear to be unable to understand that we have been transformed into titans and what that means - that politics as we used to understand it has come to an end, that nature and history were in the process of becoming one, and that our commitments and obligations are premised on conditions that have exploded under the weight of nuclear weapons.

According to Arendt and Anders, the underlying reason is that humanity had begun to occupy in effect two different realities, to live a schizophrenic life: One life in the reality of things as they 'appear naturally to our consciousness', i.e. the world of the particular, of sense perception and everyday language, and one that is based on an entirely different ontology, namely the scientific and technological realm, embodied in modern machinery, the underlying ontology of which, as we have seen in the previous chapter, is in fundamental respects at odds with our naïve, phenomenal ontology. Contrary to Heidegger, however, who was afraid that the technological understanding of being might one day become our only one, Arendt and Anders saw not one ontological condition replacing the other. For them the situation was politically far more problematic – they saw a world in which both had begun to coexist alongside the other to the effect that humans were in vital respects out of touch with the new reality they were erecting around them. This, Arendt argues, is true for the average citizen just as much as for scientists and engineers:

> "The fact is not merely that the scientist spends more than half of his life in the same world of sense perception, of common sense, and of everyday language as his fellow citizens, but that he has come in his own privileged field of activity to a point where the naïve questions and anxieties of the layman have made themselves felt very forcefully, albeit in a different manner. The scientist has not only left behind the layman with his limited understanding; he has left behind a part of himself and his own power of understanding, which is still human understanding".[481]

For Arendt and Anders this disjunction was the greatest political problem of the time because it meant that the scientific and technological problems that were beginning to define human existence could no longer be meaningfully translated into everyday language and thus into a language that corresponds with the categories of human understanding. This means that our interaction with modern technology had become, not only on several dimensions thoughtless but also that we could

[480] Arendt, H. (2007), p. 46.
[481] Arendt, H. (2007), p. 45.

not express and meaningfully speak with one another about technological questions and hence politically engage with them in the public realm.

In order to carve out more clearly what Arendt and Anders had in mind, when they spoke of humanity's modern condition as fundamentally thoughtless, I will contrast their position with another school of thought that was prominent at the time of their writing and that, if looked at superficially, could be taken to diagnose similar problems. In the 1950s and 1960s, the diagnosis of a 'lag' between humanity's ethical, moral and political development on the one hand and its scientific and technological development on the other hand was commonplace amongst social and political scientists. An early version of the underlying mindset can be found in Bertrand Russell's 1924 piece 'Icarus or the Future of Science'. Russell here claims that:

> "the sudden change produced by science has upset the balance between our instincts and our circumstances, but in directions not sufficiently noted. Over-eating is not a serious danger but over-fighting is. The human instincts of power and rivalry, like the dog's wolfish appetite will need to be artificially curbed, if industrialism is to succeed".[482]

Russell here entertains the idea that humanity had succeeded to create a new scientific and technological reality for which its social, cultural and biological make-up was ill-equipped. Perspectives on technology such as these were later formalised in the thought of sociologists such as William Ogburn, Bernard Brodie, or Hornell Hart, in a tradition of sociology called 'cultural lag theory'. The central idea of authors writing in this tradition was that material and technological developments were the main drivers of history and that social organisation and norms were lagging behind these developments to the effect that the latter had to 'catch up' with the former.[483] In more technical terms, a cultural lag was defined as "a condition of strain or maladjustment produced by the lagging of one of two correlated parts of culture behind the other".[484] As argued above, cultural lag theorists tended to put emphasis on a lagging of non-material culture behind material culture. According to Sylvest and Munster (2016), cultural lag theory formed the basis of nuclear strategy during the thermonuclear age in the 1950s and 1960s, both of deterrence-based approaches and internationalist approaches, and was sustained by the belief that "social man could catch up with scientific man" with the help of modern techniques of social and political science.[485] Existential risk theory clearly can be placed in this tradition of thinking about technology. Like cultural lag theorists, existential risk theorists demand a socio-political adjustment to the technological and scientific demands of the time. For Arendt, however, cultural lag theory was a "red herring":

[482] Today, of course, as the dangerous effects of global mass consumption are making themselves felt in environmental degradation and climate change, it becomes clear that 'over-eating' is a serious problem too. This of course does not affect Russell's argument, if anything it renders the situation he describes even more problematic. See Russell, B. (1924).

[483] Please compare to Munster, R.v. & Sylvest, C. (2016a), p. 10.

[484] Schneider, J. (1945), p. 786.

[485] Sylvest, C. & Van Münster, R. (2016), p. 10.

> “the often mentioned ‘lag’ of the social sciences with respect to the natural sciences or of man’s political development with respect to his technical and scientific know-how is no more than a red herring drawn into this debate […]”.[486]

Arendt did by no means deny the existence of a deep divide between what we might call the realm of science and technology and the ‘*world of sense perception, of common sense, and of everyday language*’. On the contrary, she considered this growing divide to be the perhaps most dangerous political problem of the time. Her question, though, was what story this divide really told and what lessons should be drawn from it. As we will see below, cultural lag theory, by jumping straight from the diagnosis of a lag to the diagnosis of a need for adaptation on the social and political side and, by implication, presupposing the possibility thereof, in her view exhibited a peculiar kind of thoughtlessness she considered typical for her times. It failed to take a step back, to inquire into the nature and the origin of the problem at hand and therewith to grasp the profundity of the implications of the divide between ‘social man’ and ‘scientific man’ that was unfolding.

For Arendt lag theory was a ‘red herring’ because it lacked an understanding of what the project of modern science and technology actually was about and hence diverted attention from what she considered to be the root cause of the divide in question. In noticeably Heideggerian fashion Arendt argued that the goal of modern science is “no longer to ‘augment and order human experience’ […]; it is much rather to discover what lies *behind* natural phenomena as they reveal themselves to the senses and the human mind”.[487] Modern science, in her view, was the search for ‘true reality’ marked by a loss of confidence in “appearances, in the phenomena as they reveal themselves of their own accord to human sense and reason”.[488] The crucial moment, the turning point in that epistemological revolution was the introduction of the telescope and the following realisation that, contrary to what the senses had suggested for ages, the Earth revolves around the sun. This technologically assisted realisation, Arendt argued, told man “that his senses are not fitted for the universe, that his everyday experience, far from being able to constitute the model for the reception of truth and the acquisition of knowledge, was a constant source of error and delusion“.[489]

The trouble was, in her view, that a) “the categories and ideas of human reason have their ultimate source in human sense experience […] all terms describing our mental abilities as well as a good deal of our conceptual language derive from the world of the senses and are used metaphorically” [490] and b) that “what defies description in terms of the ‘prejudices’ of the human mind defies description in *every conceivable* way of human language; it can no longer be described

[486] Arendt, H. (2007), p. 46.

[487] Arendt, H. (2007), p. 44.

[488] Ibid, p. 48.

[489] Arendt, H. (1958), pp. 582-583.

[490] Arendt, H. (2007), p. 47.

at all, and it is being expressed but not described, in mathematical processes".[491] She approvingly cites Erwin Schrödinger's dictum that the universe we are trying to "conquer is not only practically inaccessible, but not even thinkable", perhaps not as meaningless to our minds, including those of the scientist, as a " 'triangular circle' but much more so than 'a winged lion' ".[492]

Cultural lag theory, from this perspective, oversimplifies the problem at hand. The problem is not that one part of culture is lagging behind another, correlated one, but that one part of us is lagging behind another. A rift has occurred between our understanding, which corresponds with the world of appearances and phenomena as they reveal themselves naturally to our consciousness, and the realm of technical, scientific knowledge, which is based on systematic abstraction from phenomenal reality.

As a result of this alienating process, Arendt argues, another, practically more problematic, rift has occurred, namely the rift between the capacity for acting and the capacity for understanding – a rift which Anders calls the Promethean disjunction. The problem Arendt and Anders identified was that, even though we (as we have seen that 'we' includes scientists and engineers) cannot fully understand the findings and methods of modern scientific inquiry, we can still apply and utilise them in modern technology. To use a perhaps exceedingly simple example, we can easily represent infinity mathematically as ∞ and we can apply this representation in for instance computer programs, even though we cannot grasp its meaning and therefore are unable to translate it into the categories and ideas of human reason.

Arendt therefore argues that "the lost contact between the word of the senses and appearances and the physical world view has been re-established not by the scientist but by the 'plumber'. The technicians, who account today for the overwhelming majority of all 'researchers', have brought the results of scientists down to earth".[493] The result of this process is that we now have the ability to not only think 'from the point of the universe' but even to "handle nature from a point of the universe outside the earth" and, one might add, outside of ourselves.[494]

Furthermore, modern scientific inquiry itself, according to Arendt, is more accurately described in terms of 'doing', i.e. practice, than in terms of theory and contemplation. The fact, for instance, that modern science is increasingly reliant on mathematics and statistics, meant, for Arendt, that its findings cannot be meaningfully translated into the categories of human language and thought. Another reason is that its core method is that of experimentation. Homo faber, Arendt argues, creates knowledge by 'making nature', by forcing nature into specific conditions that do *not* naturally occur and thus do not ordinarily reveal themselves to thought and observation. The

[491] Ibid, pp. 47-48.
[492] Ibid, p. 46.
[493] Ibid, p. 49.
[494] Ibid, p. 54.

situations we study and based on which we acquire knowledge about the causalities reigning in nature are artificial, they are produced.[495]

For Arendt, advances in modern scientific inquiry hence cannot be separated from the technological will to work on nature to begin with. Not only does 'the plumber', i.e. technology, bring "the results of science to earth" and hence present material evidence for the validity of scientific theory, not only does scientific knowledge require active intervention with natural processes in order to advance, technology also plays a pivotal role in providing scientific inquiry with the means necessary for its advancement. As discussed above, at the very beginning of the modern age and thus of the loss of trust in the truth-telling capacities of the human senses and unaided reason, Arendt locates the invention of the telescope and hence a technology. The telescope "pierced the distance between earth and sky and delivered the secrets of the stars to human cognition,“ revealing worlds behind those that appear to the senses.[496]

In Arendt's view, the technological world and the world of ordinary human experience, in particular the different types of knowledge associated with them, thus were in many ways irreconcilable. There was no gap to bridge, no lag to close. The lag identified by cultural lag theorists, if anything, gave evidence of the fundamental irreconcilability between the two worlds modern humanity had begun to occupy. This is what Arendt and Anders sought to convey in almost identical terms when they argued that "man can do, and successfully do, what he cannot comprehend and cannot express in everyday language"[497] and that there is a "growing gap between what man can do and what man can mentally realise".[498] The Promethean disjunction speaks of a new and permanent facet of the human condition rather than a lag that can eventually be closed.

From Anders' point of view, in fact, one could see in cultural lag theory an instance of Promethean shame. It translates into the demand that human life on a political and social level adjust to the demands of technology, rather than the other way around. It thus highlights what Anders and Arendt considered to be the deeper struggle at the heart of the modern human condition, namely a struggle for the 'stature of man', i.e. for the status of what Arendt and Anders call the *human qua human* - the human as it is, limited, imperfect and faulty.[499]

[495] Arendt here again closely echoes Heidegger, who had made in effect the exact same claim in 'The Age of the World Picture', cf. Heidegger, M. (1977), p. 121.

[496] Arendt, H. (2007), p. 49.

[497] Arendt (2007), p. 46. Italics added by the author.

[498] Anders, G. (1962), p. 494.

[499] The term *human qua human*, to the authors knowledge, is never really specified by Arendt or Anders. It would be a worthy subject for a separate piece. Arendt uses the phrase in a letter to Karl Jaspers, where she discusses her understanding of the 'radical evil'. See Arendt, H. & Jaspers, K. (1987), p. 202, No. 109; For Anders' invocation of the phrase see Anders, G. (1956a), p. 48.

Conclusion - a problem of benchmarks

As this discussion of the topic of nuclear weapons in the thought of Hannah Arendt and Günther Anders shows, both authors shared many of the concerns we can identify in existential risk scholarship today. They did take the problem of human extinction seriously as a problem of unique and unprecedented moral, ethical and political importance and argued, as existential risk researchers do today, that its emergence has catapulted humanity into a new epoch.

Yet, their discussions of the threat to human life on Earth is embedded in a discussion of what they considered to be a much deeper and wider transformation of the human condition through modern technology. Arguably, they took the threat to human life on Earth much more seriously than existential risk researchers do today in trying to make sense of it in such terms and in trying to uncover its deeper philosophical implications. If one wanted to try and boil the two authors' shared concerns regarding modern technology down into one sentence, one might argue that they both were concerned that humanity was in the process of building around it a world of technological systems and apparatuses of which it could make ever less sense and which, in turn, could not 'make sense' of humanity.[500] That is, they feared that humanity was beginning to occupy an 'unhuman world', a world increasingly alien to it.

The fact that, due to the appearance of nuclear weapons, the human species suddenly needed to consider itself as mortal, for both authors, was arguably but the most momentous and monstrous instance of this alienating dynamic. From the moment of the arrival of nuclear weapons onwards humanity's "mode of being", as Anders argues, had been transformed into "not yet being non-existing".[501] Humanity had begun to live in the "Age of Respite" and this age, according to Anders, had to be considered humanity's "Last Age" because, no matter for how long it would last, its "differentia specifica, the possibility of self-extinction can never end but by the end itself". Humanity thus saw itself confronted with an open-ended end-time, in which its only aim could be to delay the end for as long as possible - to make "the time of the end endless".[502] Anders succinctly summarises this shift in perspectives in his claim that the future "no longer 'comes'; we can no longer understand it as 'coming'; instead we are 'making' it. And we are making it in such a way that it always contains the possibility of its abrupt ending in itself".[503]

[500] Saying that apparatuses cannot 'make sense' of humanity may seem inept. However, what this is intended to convey is an ontological problem. - the Heideggerian idea that modern technology embodies an ontology that inherently abstracts from phenomenal reality.

[501] Anders, G. (1962), p. 493.

[502] Ibid, p. 494.

[503] Anders, G. (1956a), p. 282. The text was translated by the author. The original German version reads as follows: "Denn die Zukunft „kommt" nicht mehr; wir verstehen sie nicht mehr als „kommende"; wir machen sie. Und zwar machen wir sie eben so, daß sie ihre eigene Alternative: die Möglichkeit ihres Abbruchs, die mögliche Zukunftslosigkeit, in sich enthalt. Auch wenn dieser Abbruch nicht morgen schon eintritt — durch

This new temporality, however, according to both Arendt and Anders is profoundly alien to us. As Arendt's concept of world shows, taking the future for granted, considering ourselves as part of an ongoing transformation of qualities, thinking in terms of temporally extended commitments, etc., is what, according to Arendt, it *means* to be human. It is how we make sense of what a human life *is*. In all our commitments, in everything we aspire to, and in everything we value the future is present and implicitly taken for granted. This is also why Anders argued that we have been transformed into a new species, into titans, by becoming 'Lords of the Apocalypse' and suddenly responsible for 'making' the future – a species for which the emotions and intuitions of past generations have become alien.

Existential risk research can be seen as an attempt to come to terms with this new perspective on the future; its mission could literally be summarised as that of trying to 'make the time of the end endless'. But existential risk theory complicates Anders' and Arendt's concerns because it highlights that, whilst these two authors were concerned with just one threat to the survival of the species, a technological one, there are in fact many and that, on cosmic time-scales, the open-ended end time really is an open-ended obstacle course, where humanity needs to deploy multiple strategies to survive for any considerable amount of time. What the generalised perspective on existential risk shows is that humanity has always been living in the 'Age of Respite', albeit perhaps without realising it. In other words, when Anders claimed that nuclear weapons transformed the temporality of humanity from one in which the future could be taken for granted into one of a permanent respite, from an existential risk perspective this merely means that humanity was forced to give up on a beautiful illusion.

The generalised perspective on existential risk therefore can be said to rehabilitate technology demonstrating its ambivalent role when it comes to the survival of the species and that it is not solely technology's 'fault' that humanity now finds itself confronted with the problem of finitude on a collective level too. On the contrary, it stresses that even though technology may be our greatest threat in the short term, in the long run it also is our only hope for survival. To put it in the words of Tegmark: "If we don't keep improving our technology, the question isn't whether humanity will go extinct, but how. What will get us first—an asteroid, a supervolcano, the burning heat of the aging Sun, or some other calamity".[504]

Against that background, Günther Anders' and Hannah Arendt's critique of technology by reference to nuclear weapons may begin to seem dated. If technology is our only means for survival, then thinking about it in categories such as 'the radical evil' might appear short-sighted and metaphysically grounded concerns could appear as of secondary importance. It is important to note at this point, however, that neither Anders, nor Arendt, nor even Heidegger, would have thought of

dasjenige, was wir heute tun, kann er übermorgen eintreten oder in der Generation unserer Urenkel oder im ‚siebten Geschlecht'."

[504] Tegmark, M. (2017), p. 317.

themselves as Luddites, i.e. as wholly opposed to technological progress. Arendt conceded for instance that "it is only the rise of technology, and not the rise of modern political ideas as such, which has refuted the old and terrible truth that only violence and rule over others could make some men free".[505] Anders rejected criticism labelling his philosophy of technology as reactionary by claiming that he was not "metaphysically conservative" and did not insist "on an alleged (metaphysical) status of the world as it is", framing "human morality along the lines of 'things are the way they are and should be'".[506] Even Heidegger argued that "it would be foolish to attack technology blindly. It would be short-sighted to condemn it as the work of the devil. We depend on technical devices; they even challenge us to ever greater advances".[507] That is, none of them would have denied that technology plays a critical role in sustaining human life, nor even that the desire to push the boundaries of the possible by technological means is deeply connected to what it means to be human, that it is expressive of our freedom, our curiosity, our capacity to wonder and ponder and the associated desire to make sense of our 'thrownness into being'.

What Heidegger, Arendt and Anders warned against is *uncritical* interaction with modern technology by conceiving of it as a mere means to an end, as no more than a neutral tool in the service of humanity. The above discussions of Promethean shame, technology as action, the radical evil, and the schizophrenic condition of modern human existence, were intended to convey an idea of the pathologies and paradoxes which Arendt and Anders saw as resulting from such uncritical interaction with technology. The concept of Promethean shame shows that, by conceiving of technology as a neutral tool, humanity implicitly makes an unspecified idea of technological perfection the measure of things and begins to conceive of itself as outdated, inadequate and expendable, resulting in a complete inversion of the means-ends relationship. The notion of technology as action shows that conceiving of technology as a means to an end not only fundamentally misrepresents the categorically altered nature of modern technology as compared to ancient tekné but nurtures the anachronistic belief that more technological control over natural processes somehow translates into more control over the future (human destiny) whilst in reality it translates into the exact opposite in several respects. The notion of 'the radical evil' highlights that modern technology confronts humanity with problems that exceed what humans can imagine and understand, both because of the scale of the horrors in question, and because the philosophical implications of these horrors undermine categories of thought based on which we are accustomed to make sense of ourselves and our situation. The notion of the 'schizophrenic condition of modern existence', lastly, refers to what Arendt and Anders saw as the deeper reason underlying these pathologies and paradoxes – the fact that the ontological condition underlying technological progress (i.e. the technological understanding of being), is fundamentally at odds with our phenomenal

[505] Arendt, H. (1990), p. 114.
[506] Anders, G. (1956a), p. 328, fn. 45.
[507] Heidegger, M. (1966), p. 54.

consciousness and thus the reference units of ordinary human understanding, allowing us to do what we cannot understand.

These observations provide us with different angles on the same problem: that scientific and technological development appear to have a deeply paradoxical effect on us and our image of ourselves. We simultaneously conceive of ourselves as in principle omnipotent masters of the universe and at the same time as hopelessly outdated, unacceptably faulty creatures. On the one hand, science and technology increase our knowledge about the future of the universe to an extent unimaginable for past generations - we are able to predict that the sun will turn into a black hole hundreds of millions of years from now. On the other hand, our technologically amplified powers imply that the future states of our own earthly environment are arguably less predictable for us than they were for past generations. We at once consider ourselves more knowledgeable than ever before and find ourselves mistrusting our unamplified, natural judgment to an increasing extent. In brief, the above observations explain why Anders identified "hubristic humility" or "arrogant self-degradation" as modern humankind's defining attitude.[508]

Existential risk theory arguably contains the purest possible version of this paradoxical attitude. Taking existential risk seriously in the macro-strategic sense means thinking in cosmic timescales, reflecting about earth and humanity from a detached, exterior perspective, and using this perspective to "weigh ethical dilemmas, and evaluate global priorities" in order "to clarify the choices that will shape humanity's long-term future".[509] In Anders' view this is inherently hubristic. In the context of existential risk, scholars do not assume this abstract viewpoint for purely theoretical purposes, as theoretical physicists and cosmologists do, but to guide our practical considerations and policy making in the here and now. What both Arendt and Anders stressed was that making the point of the universe one's benchmark for thinking about human affairs does *not* mean that one assumes a neutral or objective viewpoint. On the contrary, it means that one takes a highly normatively charged viewpoint from which everything seems arbitrable, negotiable and perfectible.

Existential risk research hinges on the assumption that humanity actually could, one day, occupy this detached, abstract point in time and space, from which all the things that might otherwise result in extinction are negotiable and manageable. Cotton-Barratt and Ord (2015) concede that this factor is already reflected in the standard definition of existential risk ("an existential risk is one that threatens the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development"[510]). Under this definition, Cotton-Barratt and Ord argue, we are "comparing ourselves to the most optimistic potential we could reach".[511]

[508] Anders, G. (2016), p. 49.
[509] See FHI (2018a).
[510] Bostrom, N. (2013), p. 15.
[511] Cotton-Barratt, O. & Ord, T. (2015), p. 4.

Anders' concept of Promethean shame allows us to see that this comparison is inherent to our customary idea of technology. As argued in section 3.1, for Anders our conception of technology as a neutral means *implies* that we are constantly 'comparing ourselves to the most optimistic potential we could reach'.

This brings us to what both Anders and Arendt saw as profoundly dangerous about uncritical interaction with technology. They were concerned that when we think of technology as a mere means to an end we see a world of potential perfection, where all options technology can in principle provide us with have been realised. This idea of technological perfection, however, in their view, is nothing but an empty projection surface and making it our benchmark for thinking about present and future is not only bound to fail on multiple dimensions but profoundly dangerous.

First, because, as has been established, in our pursuit of perfection we risk creating a world which is actually more dangerous, less controllable and predictable and for which we, as imperfect beings, are ill-suited, ending up being the eternal saboteurs of our products. Second, because even if technological maturity, "the attainment of capabilities affording a level of economic productivity and control over nature close to the maximum that could feasibly be achieved",[512] was possible, it would raise Heidegger's question: "what for? - where to? and what then?" [513] Arendt and Anders agreed with Heidegger that modern technology has its roots in a normatively problematic understanding of being, an ontology that abstracts from phenomenal reality which is the basis of the human way of being-in-the-world and thus of the reference units of our values and understanding. The basic property of modern technology is that, ontologically and practically, it transcends givens and transforms them into manipulable processes and negotiable states of affairs.

Like Heidegger, they were concerned that the pursuit of technological perfection could only result in a profoundly unhuman if not outright anti-human world. Both sought to tease out what they considered to be, at bottom, anti-humanist tendencies inherent to progress in modern science and technology, fuelled by a self-contradictory cocktail of emotions, passions and philosophical principles as it is. According to Arendt and Anders the passionate pursuit for perfection, driven by the belief that 'everything is possibly', is inherently self-contradictory and anti-human, because it is only within certain, from a technological standpoint arbitrary, conditions that the term 'human' has any content at all - conditions in the form of world (intergenerational permanence, stability, durability), plurality, and spontaneity, conditions in the form of mortality and natality, and conditions constituted by limits to what we can know, speak and think meaningfully about, connected to understanding and thus sense-experience and emotion as these notions are. Even the desire to overcome these conditions and limits itself can only be understood from within them. Without limits, from the point of the universe, if truly everything was possible and all parameters of existence negotiable and adjustable, there would be nothing at all to strive for, nothing to desire at all, because

[512] Bostrom, N. (2013), p. 19.

[513] Heidegger, M. (2000b), p. 40.

there would be no reference point to such desires. There would be, as Heidegger puts it, nothing from which any *ordo* could arise. As argued in chapter 2, even calculative thinking takes its course from something given and, therefore, if it were to become the only way of thinking, it would render itself directionless.

C.S. Lewis (1948), in *The Abolition of Man*, arrives at the exact same conclusion about the prospects of what he calls 'analytical understanding', which he sums up in a powerful metaphor that illustrates Heidegger's, Arendt's and Anders' concerns pointedly:

> "You cannot go on 'seeing through' things forever. The whole point of seeing through something is to see something through it. It is good that the window should be transparent, because the street or garden beyond is opaque. How if you saw through the garden too? […] If you see through everything, then everything is transparent. But a wholly transparent world is an invisible world. To 'see through' all things is the same as not to see".[514]

For Arendt and Anders, the problem of technology hence came down to a problem of benchmarks. Neither took issue with technology per se, with making use of artifice to make our lives better. What they warned against was making technology, i.e. the logic of total negotiability and perfectibility, the measure of things. It is this criticism of the technological mindset which shows why Arendt's and Anders' critique of technology is not rendered outdated by the generalised perspective on existential risk. The Promethean disjunction – "that man can do what he cannot comprehend" is, in effect, an ontological credo about the essential role of givens and limits in human life which existential risk theorists either disregard, reject or are unaware of.

So far, the debates therefore largely run parallel to one another, observing, arguably, the same phenomenon, the same transformational process, but analysing it in an entirely different light, the one from a technical, the other from a phenomenologically informed perspective. At the heart of this divide are two fundamentally different conceptions of technology. On the one hand side we have the mainstream conception of technology, where technology is understood as having no valuative content of its own. On the other hand, we have a critical school of thought which argues that technology cannot be discussed in isolation from ontological questions - that it in fact is based on and perpetuates a problematic attitude to being which not only undermines the possibility of normativity but also makes it impossible for us to mentally realise what we are doing.

These two positions, however, appear to converge in debates about risks associated with 'superintelligence'. The fear in existential risk circles, to put it in Heideggerian terminology, appears to be that superintelligent machines could turn everything that exists into standing reserve. Interestingly, as will be discussed in the following chapter, the underlying reason for that fear can be shown to be an ontological one. The debates surrounding AI demonstrate precisely what Heidegger, Arendt, and Anders claimed all along, namely that one cannot use modern technology as if it were a

[514] Lewis, C.S. (1943), p. 51.

mere means to an end and abstract from the deeper ontologically rooted dangers it poses indefinitely. In the fears surrounding AI we encounter many of the ontological concerns regarding technology, which Heidegger, Arendt and Anders discussed using categories such as ‘thoughtlessness’, ‘objectlessness’, ‘Promethean shame’, or ‘standing reserve’, couched in a technical language under labels such as ‘value alignment’, ‘orthogonality thesis’, or ‘Realistic World Models’. To this point of convergence I am turning in the following chapter.

4. Technology awakes

Introduction

Selmer Bringsjord (2015) summarises Nick Bostrom's influential 2014 book 'Superintelligence' in a rather parsimonious one-liner: "We should be deeply concerned about the possible future arrival of super-intelligent, malicious computing machines (since we might well be targets of their malice)".[515] This characterisation certainly is not flat-out wrong but it misrepresents the concerns of Bostrom and many other artificial intelligence (AI) researchers, who voice concerns about AI from an existential risk perspective. My aim in this chapter is to unpack these concerns because, irrespective of how likely or unlikely, judged by scientific standards, the emergence of superintelligent machines might be, the present debates surrounding the topic, featuring a colourful amalgam of existential fears, existential hopes, ridicule and serious scholarly concern, happen to reflect the whole spectrum of complications and puzzles I have sought to raise attention for throughout the past chapters.

That is, these concerns provide us with a magnifying glass through which to examine the puzzles at the heart of existential risk research, highlighting, once again, the prescience and ongoing actuality of Heidegger's, Arendt's and Ander's thinking about technology. The debates about the long-term consequences of progress in AI can be seen as a technical meditation about what 'technology as destiny' might actually mean. It is interesting to see, therefore, that in these debates, existential risk researchers and those who share their concerns regarding progress in artificial intelligence, appear to articulate similar concerns regarding technological progress as those we have encountered in Heidegger's, Arendt's and Anders' critical analyses of scientific and technological progress. They revivify what is perhaps Heidegger's, Arendt's and Anders' core concern with respect to technology, namely that at the heart of the modern condition is an ontological struggle, a clash between two conflicting ontologies, or relations to reality and thus two different worlds which modern humanity had begun to occupy - worlds both in an ideational sense as well as in a physical sense – phenomenal reality, the world of sense experience, of the particular, common sense, of ordinary language on the one hand side and the technological, the world of abstraction and generalisation and its products on the other hand side, in which our ordinary categories of thought and action are making less and less sense. AI can be seen as a pure version of this dilemma.

Artificial intelligence, or machine intelligence, without doubt is the most widely and most persistently debated single technological issue within the existential risk movement and, as I intend to demonstrate, against the background of the discussion of the previous chapters, this should not come as a surprise. The reason lies in the fact that, to put it in Stuart Russell's words, contrary to

[515] Bringsjord, S. (2015).

nuclear weapons, bio-technology, geo-engineering, or other previously touched upon technologies, AI is not really *a* technology at all, nor is it a specific class of technical approaches, but a *problem*, namely the "general problem of creating intelligence in machines".[516] By implication the often anticipated existential risk from AI, at least at the present point in time, also cannot really be seen as a risk that is associated with a specific technology, i.e. as a risk comparable to those from nuclear weapons or geo-engineering, but needs to be seen as a risk that is associated with a largely *theoretical* problem. This problem, however, appears to be the same as the by now familiar problem of how to reconcile the technological understanding of being with the world of human sense experience and ordinary understanding. When Blitz (2014) claims that Heidegger sought to draw "attention to technology's place in bringing about our decline by constricting our experience of things as they are" and to the fact "that we now view nature, and increasingly human beings too, only technologically […] as raw material for technical operations", this assertion, curiously, sounds uncannily similar to the kind of concerns that are being voiced by existential risk researchers in the context of AI.[517] The fear regarding AI is, to put it in the words of Jaan Tallinn and Huw Price, "that by creating artificially intelligent machines we risk yielding control over the planet to intelligences that are simply indifferent to us and to the things we consider valuable".[518] Now, of course this is precisely what, according to Heidegger, Arendt and Anders, we were at risk of doing all along, by yielding to the relentless logic of scientific and technological progress both in mind and in practice. As we have seen in chapter 2, Heidegger was convinced that the challenging revealing underlying modern technology is inherently oblivious 'to us and to the things we consider valuable' because it cannot make sense of such notions as 'us' and 'things' to begin with. In the discussions surrounding AI, as we will see, this ontological clash between the technological understanding of being and how things appear naturally to our consciousness, which Heidegger had uncovered and which shines through in Arendt's and Anders' diagnoses of the 'schizophrenic condition of the modern age', re-emerge as engineering problems.

One might be inclined to think, in the light of this ostensible common ground, that in the case of AI the two branches of discussion, with Heidegger, Anders, and Arendt on the one hand side and existential risk theory on the other hand side begin to converge. However, as we will see, in last consequence such appears not to be the case. The reason is that their respective starting points of reflection are completely different ones – the latter think 'from the point of the universe' the others from the point of the 'the human qua human'. My aim in this chapter is to pinpoint exactly where, how and ultimately why the concerns of the two camps might begin to converge in the case of AI in order to then show why they nonetheless remain unreconcilable.

[516] Stuart Russell is one of the most widely-used text book on AI. For the above quoted cited definition of AI as a problem rather than a technology, see Russell, S. (2018).
[517] Blitz, M. (2014), p. 63.
[518] Price, H. & Tallinn, J. (2012).

The chapter is organised in the following way. First, a brief summary of the standard argument that artificial intelligence potentially harbours existential risks is provided. In a second step I will then frame the thus presented argument and relate it to some core distinctions in the field. This will serve to bring out what I consider to be the most interesting characteristic of the existential fears surrounding AI, namely the fact that they alternate between highly vague and highly specific conceptions of AI in their visions of AI's future. In the remaining sections I will then proceed to discuss the literature on superintelligence from the perspective of Heidegger's, Arendt's and Anders' philosophy of technology and in the light of the preceding chapters.

4.1 The existential risk from AI

On the surface, the argument that the prospect of superintelligent machines should leave us deeply concerned is rather straightforward. Nick Bostrom opens his book '*Superintelligence'*, with a fable – "The unfinished fable of the sparrows" – which tells the story of a flock of sparrows that discusses the potential advantages and disadvantages of taming an owl. Some sparrows praise the potential benefits that might come with taming an owl and how the owl might help them to build their nests and protect them from the neighbourhood cat. Only one old, half-blind sparrow seeks to caution such hopes and warns that the attempt to domesticate an owl seems like an inherently dangerous thing to do, in particular since the sparrows have no experience in 'the art of owl domestication'. However, the optimistic sparrows prevail: "It will be difficult enough to find an owl egg" the flock leader proclaims, "so let us start there. After we have succeeded in raising an owl, then we can think about taking on this other challenge".[519] Obviously, this fable is mainly intended to function as a teaser and it is not to be taken too seriously as an analogy for the concerns surrounding artificial intelligence. It does nevertheless capture the spirit of the debate rather accurately and, perhaps unwittingly, reveals several of the argumentatively problematic features of the take on the problem of AI in existential risk theory.

The 'existential risk from AI argument' (in the following abbreviated as 'AI-risk argument') proceeds as follows. First, it holds that we might be approaching a moment at which AI becomes broadly comparable to human intelligence in its general applicability.[520] In the literature this hypothetical AI is generally referred to as artificial general intelligence (AGI). Once machine intelligence reaches such a threshold, it is further argued, it might soon become better than humans at the specific task of designing intelligent machines. What might then follow is often referred to as an 'intelligence explosion' - a run-away process or feedback cycle of exponential, self-augmenting and -optimising machine intelligence which would propel machine intelligence onto levels beyond

[519] Bostrom, N. (2014), p. 0.

[520] See for instance Shanahan, M. (2015), or Bostrom, N. (2014).

anything humanly imaginable, resulting in so called superintelligence.[521] Superintelligence is typically defined as a system that supersedes human intelligence in practically all domains. Soares and Fallenstein (2014) for instance define superintelligence as an AI that is "smarter than the best human brains in practically *every* field",[522] Bostrom et al. (2016) similarly define superintelligence as a system that is "more cognitively capable than humans in *all* practically relevant domains",[523] and CSER defines superintelligence as being "superior to human performance in many or nearly all domains".[524] The significance of such an event, i.e. of an intelligence explosion, it is further argued, would be hard to overstate, because, as for instance Russell (2017) holds: "everything our civilization offers is a consequence of our intelligence; thus, access to substantially greater intelligence would constitute a discontinuity in human history".[525]

We can distinguish between two basic hypotheses, the 'tipping-point hypothesis' and the 'discontinuity hypothesis', that jointly form the basis of the AI-risk argument. The tipping point hypothesis holds that if AI capacities reach a certain threshold, namely human level intelligence, it is likely to supersede this level across all domains soon after.[526] The discontinuity thesis holds that such an event would constitute a rupture in human history because whatever follows would be radically different from our present point of view.[527]

The existential risk is associated with the possibility that this cataclysmic event might turn out to have catastrophic consequences from a human point of view. Price (2013) for instance argues that "we humans are nearing one of the most significant moments in our entire history: the point at which intelligence escapes the constraints of biology" and adds that he sees "no compelling grounds for confidence that if that does happen, we will survive the transition in reasonable shape".[528]

Price bases his concerns on what he calls a 'pragmatist conception' of intelligence: "Don't think about what intelligence is, think about what it does […] we tend to be much better at controlling our environment than other species […] the question is then whether machines might at some point do an even better job".[529] In other words, intelligence is understood as the property which manifests itself in our ability to control our natural environment. This perspective is representative for the literature on existential risk. Soares and Fallenstein (2014), of the Machine Intelligence Research Institute (MIRI), for instance argue that "the property that has given humans a dominant

[521] See for instance Ó hÉigeartaigh, S., et al. (2018); or Shanahan, M. (2015).
[522] Soares, N. & Fallenstein, B. (2014), p. 1.
[523] Bostrom, N., et al. (2016), p. 2.
[524] Compare to CSER (2018b).
[525] Russell, S. (2017), p. 179.
[526] The designation 'tipping-point' hypothesis, in this particular context, needs to be understood as corresponding with Huw Price's (2013) characterization of the problem.
[527] The designation 'discontinuity hypothesis' follows Stuart Russel's previously referenced account of the potential consequences of an intelligence explosion.
[528] Price, H. (2013).
[529] Price, H. & K. Vold (2018).

advantage over other species is not strength or speed, but intelligence",[530] and Bostrom (2014) claims that "the human brain, has some capabilities that the brains of other animals lack. It is to these distinctive capabilities that we owe our dominant position on the planet. Other animals have stronger muscles and sharper claws, but we have cleverer brains".[531] By implication, the existential risk associated with superintelligence is (since a superintelligence would have to be assumed to be much better, perhaps unimaginably better, at controlling the environment than us) that humanity could find itself pushed to the brink of extinction as a mere *side-effect* of an intelligence explosion, by superintelligent machines that pursue their goals without taking into account human interests.[532] Just as many species find themselves pushed to the brink of extinction due to the advance of human civilisation, we might find ourselves wholly dependent on the will of these hypothetical future agents: "As the fate of the gorillas now depends more on us humans than on the gorillas themselves, so the fate of our species would depend on the actions of an alien intelligence".[533] Existential risk researchers, however, are careful to avoid the impression that they are concerned about the emergence of malevolent, superintelligent robots that try to enslave humanity, as depicted in science fiction movies from 'The Terminator' to the 'The Matrix'.[534]

The FLI states that such representations "succinctly summarise the scenario that AI researchers *don't* worry about" and that, in fact, it "combines as many as three separate misconceptions: concern about *consciousness*, *evil,* and *robots"*.[535] As we have seen above, the risk is understood in pragmatic terms. Accordingly, consciousness, it is argued, is irrelevant to the AI risk because what matters is what AI does, not if, how, or what it might be feeling or thinking whilst doing it. By implication, in so far as consciousness is considered to be a precondition for the presence of motivations and emotional states, concerns about *evil* are not part of the argument. The real worry, the FLI states, is not malevolence but *competence*, where competence is understood as the ability to efficiently attain goals. Superintelligences need not be hostile to humanity, mere mis-alignment between their goals and our values would suffice to pose an existential threat if the AI is sufficiently competent to realise its goals irrespective of our preferences: "Humans don't generally hate ants, but

[530] Soares, N. & Fallenstein, B. (2014), p. 2.

[531] Bostrom, N. (2014), p. i.

[532] Muehlhauser, L. & Salamon, A. (2012), p. 28 ff.

[533] Bostrom, N. (2014), p. i.

[534] When CSER was founded in 2012, *The Sun* for instance reported about the event with an article entitled 'Terminator Centre to be opened at the University of Cambridge' a phrase that was soon picked up by other media outlets, such as the Daily Mail, the Express or Fox News. *Wired Magazine,* in a similar fashion, referred to the members of the FHI and CSER as "Earth's Guardians, the real-world X-men and women saving us from existential threats", referring amongst other potential threats to 'rogue AI'. For existential risk researchers articles such as these are a nuisance as they feel that their cause is not only ridiculed and belittled by being relegated to the realm of science-fiction but that the real substance of their concerns is entirely misrepresented. Martin Rees therefore felt compelled to counter this kind of coverage in an article, published in *The Guardian,* entitled 'Cambridge University's 'Terminator studies' department – do we really need it?' And Huw Price published an article in *The New York Times* where he clarified his reasons for joining CSER and explicitly concerned himself with science-fiction analogies such as the above. See Price, H. (2013).

[535] Compare to FLI (2018b).

we are more intelligent than they are – so if we want to build a hydroelectric dam and there is an anthill there, too bad for the ants".[536] As we will see below, in AI safety research this problem has come to be known as the 'value-alignment problem'. This of course raises the question how a hypothetical AI could pose a threat to humanity without a body, i.e. without a connection to the physical world that would allow it to exert control over and manipulate the environment, including us, to such an extent that it could conceivably pose a threat to humanity's survival. To this AI researchers tend to respond that an internet connection would suffice for a misaligned superintelligence to cause significant, perhaps catastrophic harm. An internet connection "may enable outsmarting financial markets, out-inventing human researchers, out-manipulating human leaders, and developing weapons we cannot even understand. Even if building robots were physically impossible, a super-intelligent and super-wealthy AI could easily pay or manipulate many humans to unwittingly do its bidding".[537] In the literature this problem is referred to as the 'control' or 'containment problem'.

In brief, intelligence is understood as competence, and competence is understood as the ability to maximise one's expected utility by means of exerting control over a given physical and/or digital environment. As Russell (2018) argues "a computer is intelligent to the extent that it does the right thing rather than the wrong thing. The right thing is whatever action is most likely to achieve the goal, or, in more technical terms, the action that maximises expected utility".[538] Tegmark (2017), of the FHI, similarly defines intelligence as the "ability to accomplish complex goals".[539] Hence, intelligence in the discussions surrounding the existential risk from AI is broadly reduced to categories of instrumental rationality, i.e. means-ends thinking.[540]

Now, this definition of intelligence is not simply a theoretical model of intelligence which is used as a basis for conjectures about potential future superintelligent behaviour, it is the model of intelligence based on which AIs are designed today. Presently existing AIs *are* expected utility optimisers.[541] However, at the moment AIs tend to be limited to individual tasks and domains. For this reason, they are typically referred to as 'narrow AIs'. Narrow AIs provide domain-dependent and problem-specific solutions. They are single-purpose programmes that perform well, often on superhuman levels, at clearly specified tasks within neatly compartmentalised and controlled environments such as board or computer game environments, factories, public transportation infrastructure, or social media websites.[542] On the other hand, narrow AIs are incapable of executing

[536] Ibid.

[537] Ibid.

[538] Russell, S. (2018).

[539] Tegmark, M. (2017), p. 71.

[540] According to the Stanford Encyclopedia of Philosophy's entry on instrumental rationality, for instance, "someone displays instrumentally rationality insofar as she adopts suitable means to her ends". See Kolodny, N, et al. (2016).

[541] Häggström, O. (2016), ch. 4.

[542] Etzioni, O. (2016).

any task apart from the one they have been explicitly designed for. Deep Blue for instance defeated former world champion Gary Kasparov in chess but was incapable to decide on a single move in checkers.[543] In other words, narrow AIs are inevitably rendered dysfunctional once they are asked to optimise outcomes in a different domain than the one they have been explicitly designed for, or once the environment within which they are supposed to execute the task changes in unprovided-for ways.[544] Descartes' assessment of the promises of machine intelligence, against that background, apparently still holds: "although machines can perform certain things as well as or perhaps better than any of us can do, they infallibly fall short in others, by the which means we may discover that they did not act from knowledge, but only from the disposition of their organs".[545]

On the other hand, progress in narrow AI has been enormous over the past approximately 15 years. Driven largely by breakthroughs in machine learning, self-steering cars are becoming part of the fabric of everyday life,[546] computers have beaten the best human players in games from DoTA2, to Jeopardy! and Go,[547] workers in call centres give advice that is provided by AI expert systems,[548] speech-recognition and -processing algorithms that allow for complex phone conversations are now promising to make the call centre workers themselves redundant,[549] AIs supply us with individually tailored news and shopping suggestions, algorithmic trading is revolutionising financial markets, the 'internet of things' points the way to the 'smart home', where the seamless supply with fresh groceries, heating, illumination, etc., can be fully customised and automated, and 'robo cops' that are equipped with facial recognition software have begun patrolling the streets of Dubai.[550] Some authors even suggest that, in some economic sectors, algorithmic resource allocation might soon replace markets as central resource allocation mediators because the growing abundance of data allows for centralised decision making where before decentral approaches were more effective.[551] The success of algorithm-coordinated resource allocation platforms such as Uber can serve as a case in point.

In the eyes of many researchers in the field the present dynamic is so impressive indeed that they consider it a real possibility that in the foreseeable future every single work-related task could be fully automated and executed by a machine. A widely cited study from 2013 suggests that more

[543] Bostrom, N. & Yudkowsky, E. (2014), p. 318.

[544] Sotala, K. & Yampolinsky, R. (2015), p. 2. The underlying technology or range of technologies can of course be used for several different applications (the algorithms used in self-driving Tesla cars could plausibly be repurposed for navigating boats) but the AI itself cannot recognise that and it typically requires extensive human intervention and modification to make any, no matter how miniscule, adaptation work.

[545] R. Descartes (1911), p. 116.

[546] Lawrence, N. (2016).

[547] Silver, D. et al. (2017).

[548] Boden, M. (2016), p. 101.

[549] Google recently launched 'Google Assistant', a digital assistant that can execute phone calls on behalf of its user in order to, for instance, make appointments, reservations, etc. It has been reported that Google Assistant's conversation skills are so highly developed that humans at times cannot distinguish it from a human caller. See BBC (2018).

[550] Crawford, K. et al. (2017).

[551] Brown, A. (2015).

than half of US jobs are vulnerable to automation in the near future,[552] and in a recent survey conducted with 352 AI experts the aggregate forecast gave a 50 per cent chance that "unaided machines can accomplish every task better and more cheaply than human workers" within 45 years and a 10% chance of this happening within the next 9 years.[553]

In effect, the existential fears surrounding AI pertain to the possibility that artificial expected utility optimisers will one day no longer be narrow in their applicability but general, performing on, or above human levels across most or all domains. Such hypothetical AIs are generally referred to as artificial general intelligences (AGIs). AGIs are expected to be capable of "efficient cross-domain optimization", where 'optimisation' denotes the ability to 'steer the future into regions of possibility ranked high in a preference ordering', 'cross-domain' denotes the ability to optimise in many domains rather than just one, and 'efficient' refers to speed, computational efficiency and resource efficiency.[554] Authors in the field hence abstract from notoriously thorny questions regarding the potential mental states, motivations, phenomenal experience, consciousness, etc., of AGIs, focusing on intelligence as instrumental problem-solving capacity.[555] A superintelligent AGI, accordingly, is pictured to be a superhumanly powerful cross-domain optimisation process, outperforming humans in attaining any given set of goals by controlling its environment across all practically relevant domains.[556]

However, it is precisely this understanding of intelligence which leads some authors, following Nick Bostrom,[557] to argue that the *default* outcome of an intelligence explosion would be an existential catastrophe.[558] In order to develop a better idea of the precise nature of the fears surrounding AGI it is central to be acquainted with two regularly invoked theses based on which Bostrom (2012, 2014) and Omohundro (2008, 2012) seek to approximate how 'sufficiently rational' future AGIs might behave.

First, the so-called orthogonality thesis, according to which "intelligence and final goals are orthogonal axes along which possible agents can freely vary", which means that "more or less any level of intelligence could in principle be combined with more or less any final goal".[559] Another way of expressing the same idea is, to follow Omohundro (2012), that "rational agents […] keep their goals separate from their model of the world. Their goals are represented by a real-valued utility function U which measures the desirability of each possible outcome".[560] In other words, if intelligence is associated with identifying rational action based on specific models of the world, these

[552] Frey, C. & Osborne, M. (2013).
[553] Grace, K. Dafoe, A. et al. (2018).
[554] Compare this summary to Häggström, O. (2016), p. 104; see also Yudkowsky, E. (2008) for a comparable account.
[555] For an instructive overview of definition of intelligence see Legg, S. & Hutter, M. (2007).
[556] See for instance Häggström, O. (2016), p. 104; or Shanahan, M. (2015), p. 204.
[557] Bostrom, N. (2002; 2014, specifically ch. 8).
[558] See for instance Yudkowsky, E. (2008); Price, H. (2013).
[559] Bostrom, N. (2012), p. 3.
[560] Omohundro, S. (2012), p. 163–164.

models form one side of the equation whilst goals, i.e. the basis on which the quality of the action is to be evaluated, form the other. Hence, intelligence is understood to be value-neutral, no more than a means to an end and, as such, an AI can theoretically be built with any final goal.[561] The orthogonality thesis is generally invoked in order to discourage thinking about superintelligences along anthropomorphic lines, i.e. to warn against the idea that superintelligent AIs will somehow naturally be imbued with human values.[562] The orthogonality thesis implies that AIs can have any goal, no matter how bizarre or trivial, and that they will pursue it as effectively and efficiently as possible, *irrespective* of our interests, *unless* they are explicitly designed in ways that prevent them from acting in ways that conflict with human values.[563]

The second central thesis used to approximate the space of potential 'superintelligent motivations' is referred to as the 'instrumental convergence thesis'. It holds that sufficiently rational agents are likely to exhibit a range of sub-goals regardless of what their respective final goals might be because these sub-goals are instrumentally valuable for the achievement of *any* final goal or set of goals. Such sub-goals are henceforth argued to be likely to emerge in most, if not all,[564] sufficiently intelligent agents unless explicitly counteracted.[565] Bostrom (2012) and Omohundro (2012) argue that such agents would need to be expected to converge on four sub-goals: cognitive enhancement, technological perfection, resource acquisition, and goal-content integrity.[566] That is, any sufficiently rational agent is expected to exhibit the following drives: 1) A drive to preserve its own existence since an agent cannot achieve its objectives if it is destroyed or discontinued before its task is completed. By default, any sufficiently rational agent would therefore have to be expected to take precautions against events that might result in its premature termination. 2) A drive to increase either its own cognitive and physical capacities or its access to such capacities, since that would improve its decision-making capacities and allow it to pursue its objectives more effectively. A rational agent is therefore likely to strive for cognitive enhancement (i.e. potentially resulting, for instance, in the above-mentioned intelligence explosion) as well as to perfect its technological capacities. 3) Since any objective can be better met with more resources (given that these can be utilised either for the satisfaction of final or of instrumental goals), every rational agent should be expected to exhibit a drive to maximise its access to resources. 4) An instrumental goal to prevent alteration of its final goal structure because an alteration of its goal structure would prevent it from achieving its original

[561] Armstrong, S. (2013).

[562] Viz. Russell, S. Dewey, D. et al. (2015), p. 110; Bostrom, N. (2012); Muehlhauser, L. & Salamon, A. (2012); p. 29; Yudkowsky, E. (2008); Shulman, C. (2010).

[563] Bostrom, N. (2012), p. 5, see also Armstrong, S. (2013).

[564] According to Omohundro, S. (2012), these arguments apply equally to all types of potential architectures of intelligent systems, i.e. to neural networks, genetic algorithms, theorem provers, expert systems, Bayesian networks, fuzzy logic, evolutionary programming, etc., as long as they are sufficiently powerful, i.e. capable of far-ranging reflection and strategising.

[565] Brundage, M. (2015).

[566] See also Bostrom, N. (2014), ch. 7.

objectives.[567] Yudkowsky (2011) offers the following thought experiment to illustrate the underlying idea: “Suppose you offer Gandhi a pill that makes him want to kill people. The current version of Gandhi does not want to kill people. Thus if Gandhi predicts the effect of the pill, he will refuse to take the pill; because Gandhi knows that if he wants to kill people he is more likely to kill people, and the current Gandhi does not prefer this”.[568] Now, as argued above, the claim is not that the emergence of such drives is inevitable, the claim is that, a priori, it must be assumed that they will emerge in any sufficiently rational agent unless this is deliberately counteracted in the design of the AI’s utility function.[569]

If one combines these theses it becomes clear why existential risk researchers and an increasing number of AI researchers think of AI as harbouring potential existential threats to humanity. Taken together they imply that, by default, a sufficiently rational artificial agent would, once ‘switched on’, have reasons to resist and prevent all attempts to get it under control, to terminate it, or to retroactively alter its goal structure. Furthermore, it would seek to acquire and utilise the maximum amount of resources it can gain access to, including resources upon which humanity might depend for its survival, in order to maximise its expected utility in accordance with its final goals. If one further assumes that the agent in question is superintelligent, i.e. ‘smarter than the best human brains in practically *every* field’, it follows that such an optimisation process, once unleashed, could not be stopped by humans. Being superior to humans in all respects, it would always find ways to pursue its final goals, whatever they may be, and irrespective of what humanity might try to bring it back under control or terminate it. The implication is that, if an AGI with an overly simple final goal were to be developed, the result could easily be one in which humanity goes extinct.[570] An often invoked thought experiment in the literature is the so-called paperclip-maximiser scenario. In this hypothetical scenario an AGI is equipped with the single goal to produce as many paperclips as possible, which results in the AGI attaining superintelligence only to find ingenious ways to transform the entire biosphere, including humanity, and ultimately ever greater portions of the universe into paperclips.[571] Omohundro (2012) presents an analogous thought experiment featuring a superintelligent chess computer which finds ways to transform the entire universe into a gigantic computing machine for the sole purpose to maximise its chess-play abilities.[572] These are deliberately caricatural renditions of the problem at hand but they serve to illustrate the basic point existential risk researchers seek to make: that one needs to be careful in the specification of the utility function of intelligent machines because, as expected utility maximisers, they are designed to pursue their final

[567] Compare the above points to Omohundro, S. (2012); Bostrom, N. (2012, 2014); Muehlhauser, L. & Salamon, A. (2012); see also Soares, N. & Fallenstein, B. (2014); and Tegmark, M. (2017), specifically ch. 2.
[568] See Yudkowsky, E. (2011), p. 389.
[569] Brundage, M. (2015), Bostrom, N. (2012).
[570] Muehlhauser, L. & Salamon, A. (2012), p. 28; Bostrom, N. (2014), p. 141.
[571] See Bostrom, N. (2012).
[572] Omohundro, S. (2012).

goals in inherently expansionistic ways, irrespective of how profane, meaningless, or harmless these goals might appear.

The problem also is often illustrated with reference to tales such as that of King Midas, who, having been granted a wish by Dionysus, wished that everything he touches should turn into gold - with the result that literally *everything* he touched turned into gold, including his daughter - or the Sorcerer's apprentice, who also pays a bitter price for trying to realise a rather innocent wish by summoning forces he cannot control.[573] The morale is that if one summons an optimisation process that is more powerful than oneself one should be careful in articulating one's wishes because one might get what one asks for, not what one wishes one had asked for.[574]

The purpose of the emerging genre of AI safety research is to ensure just that. Most importantly this means finding a solution to what is generally referred to as the 'value-alignment problem', i.e. to the question how to "design methods for preventing AI systems from inadvertently acting in ways inimical to human values ".[575] This means that in order for an "autonomous system to be helpful to humans and to pose no unwarranted risks, it needs to align its values with those of the humans in its environment in such a way that its actions contribute to the maximization of value for the humans".[576] Yudkowsky calls this "utility engineering", the design of "utility functions that will give rise to consequences we desire".[577] The challenge, he suggests, is not to so much to predict what 'AI' may or may not do but to 'choose into existence' an optimisation process the good-naturedness of which can be 'legitimately asserted'.[578]

In AI safety research a variety of potential solutions to this problem are discussed. Stuart Russell suggests that it may be possible to design AI systems that are provably safe and beneficial by following three core principles:

> "1. The machine's purpose is to maximize the realization of human values. In particular, it has no purpose of its own and no innate desire to protect itself.
> 2. The machine is initially uncertain about what those human values are. […] The machine may learn more about human values as it goes along, of course, but it may never achieve complete certainty.
> 3. Machines can learn about human values by observing the choices that we humans make".[579]

The idea thus is to equip AIs with a utility functions that contain incentives to learn what and how humans value as part of its final goal structure.[580] This approach, or rather category of approaches, is

[573] See e.g. Russell, S. (2017), p. 180; FLI (2018b).
[574] Russell, S. (2017), pp. 178-179.
[575] LCFI (2018a).
[576] Hadfield-Menell, D., et al. (2016).
[577] Yudkowsky, E. as cited in Omohundro, S. (2012), p. 164.
[578] Yudkowsky, E. (2008), p. 317.
[579] Russell, S. (2017), pp. 185–186.
[580] Torres, P. (2017c).

sometimes referred to as 'indirect normativity' and is currently the most favoured approach amongst AI safety researchers.[581]

The problem of AI safety is the perhaps single most widely discussed individual technological problem in the existential risk eco-system. Not only do most of the aforementioned existential risk research institutes list AI as one of their top research priorities, over the past years a range of institutes were founded at leading universities and with the support of influential individuals from academia and the computer industry, that focus exclusively on this subject matter. These institutes are beginning to channel considerable intellectual and financial resources into the emerging field of 'AI-safety' research and into raising awareness for their concerns amongst policy makers and the wider public.[582] The most notable of these institutes are perhaps the Machine Intelligence Research Institute (MIRI), OpenAI, an independent not-for-profit research organisation based in San Francisco, which is funded by a whole range of high-profile US technologists who have pledged a total of one billion US dollars in its support, the earlier mentioned Future or Life Institute (FLI), the Leverhulme Centre for the Future of Intelligence (LCFI) at the University of Cambridge, where existential risk associated with AI is at least one of the main research areas, and the Centre for Human Compatible Artificial Intelligence (CHAI) at UC Berkeley.[583]

Many of the authors quoted in above sections are affiliated to one or more of these institutes. Eliezer Yudkowsky for instance founded and heads MIRI. Stuart Russell acts as director of CHAI and, together with Nick Bostrom, leads the 'value-alignment project' at Cambridge's LCFI. Murray Shanahan, who is professor of cognitive robotics at Imperial College London and Senior Research Scientist at Google DeepMind, is deeply involved with Cambridge's LCFI.[584] OpenAI is a special case as its work is mainly practical. That is, its focus is on actually trying to make progress towards the development of AGI, rather than on taking part in the largely still theoretical discussions portrayed above.

The institutes are united in the conviction that there is at least a non-negligible chance that AGI (and thus superintelligence), will emerge eventually, perhaps even in this century, and, given the stakes involved, they seek to contribute to the quest of making sure it will be safe, or 'human compatible'. Berkeley's CHAI for instance states that "the long-term outcome of AI research seems likely to include machines that are more capable than humans across a wide range of objectives and environments. This raises a problem of control: given that the solutions developed by such systems are intrinsically unpredictable by humans, it may occur that some such solutions result in negative and perhaps irreversible outcomes for humans. CHAI's goal is to ensure that this eventuality cannot

[581] See for instance Bostrom (2014), ch. 13; Hadfield-Menell, D. et al. (2016), Rusell, S. (2017).

[582] For an overview of different AI safety approaches in the field see Mallah, R. (2017), Baum S. (2017), or Muehlhauser, L. & Salamon, A. (2014).

[583] Please compare to the web presences of the above listed. See MIRI (2018a); OpenAI (2015); LCFI (2018b); CHAI (2018).

[584] See LCFI (2018a).

arise, by refocusing AI away from the capability to achieve arbitrary objectives and towards the ability to generate provably beneficial behavior".[585] CSER states that "superintelligence could be possible within this century" and that "AI should be developed in a safe and beneficial direction".[586] The LCFI holds that "many researchers now take seriously the possibility that intelligence equal to our own will be created in computers, perhaps within this century. Freed of biological constraints, such as limited memory and slow biochemical processing speeds, machines may eventually become more intelligent than we are – with profound implications for us all".[587] MIRI similarly states that "researchers largely agree that AI is likely to begin outperforming humans on most cognitive tasks in this century. Given how disruptive domain-general AI could be, we think it is prudent to begin a conversation about this now, and to investigate whether there are limited areas in which we can predict and shape this technology's societal impact“.[588] OpenAI does not give any indication as to whether it believes the advent of superintelligence to be imminent or not. It does, however, state that it is important to begin researching on how to make its eventual arrival safe now: "Because of AI's surprising history, it's hard to predict when human-level AI might come within reach. When it does, it'll be important to have a leading research institution which can prioritise a good outcome for all over its own self-interest. We're hoping to grow OpenAI into such an institution".[589]

In sum, the existential risk from AI is associated not with the possibility of the emergence of a malevolent or evil AI but with the inherent unpredictability and potential irreversibility of unleashing an optimisation process that is more intelligent than the humans who specified its objectives, and that furthermore must be assumed to be *inherently indifferent* to how its actions might affect humanity.[590] As Shanahan (2015) puts it: "Every action it carries out, every piece of advice it offers, will be in the ruthless pursuit of maximizing the reward function at its core. If it finds a cure for cancer, it will not be because it cares. It will be because curing cancer helps maximise its expected reward. If it causes a war, it will not be because it is greedy or hateful or malicious. It will be because a war will help maximise its expected reward".[591] Of course, this conjuncture of total moral myopia and superhuman rational intelligence may seem counterintuitive. However, Huw Price's pragmatist perspective on intelligence does undergird this perspective on intelligence. As a civilisation we are capable of creating ingenious works of engineering, art, poetry, music, etc., whilst, simultaneously, we appear to be indifferent to the misery of billions of animals in overcrowded factory farms, research laboratories, etc.

From the perspective of existential risk researchers, we are nearing a point in time after which we might find ourselves in a position equivalent to the one animals are in today - of being

[585] See CHAI (2018).
[586] See CSER (2018b).
[587] See LCFI (2018b).
[588] See MIRI (2018b).
[589] See OpenAI (2015).
[590] Cf. Russell, S. & Dafoe, A. (2017).
[591] Shanahan, M. (2015), pp. 207-208.

confronted with a superior intelligence that can do as incomprehensible, inherently unpredictable, and from our perspective pointless things to us and our environment as we can to animals. According to existential risk researchers, the main difference is that, thanks to rational choice theory, we are in a better position to predict what such superior intelligences might 'want'; that, given the nature of instrumental rationality, it must be expected to have incentives to act in ways that will conflict with our interests, unless we explicitly counteract this and ensure that the superior intelligences we create is either a 'friendly' superintelligence or firmly under humanity's control. A priori, to sum it up in Yudkowsky's words, we know only one thing: "the AI neither hates you, nor loves you, but you are made out of atoms that it can use for something else".[592]

4.2 Framing the AI-risk argument

In the above rudimentary sketch of the AI-risk argument it might already have become clear that there appears to be an argumentative leap in its discussion of AI. In fact, it would make sense to distinguish between two different argumentative levels in the argument, because it appears to rest on two very different notions of intelligence.

1) Claims pertaining to rather vague expectations regarding intelligence, surrounding notions such as 'the tipping point', 'superintelligence', AGI, or human-level intelligence. Claims of this type, on the surface, tend to leave open what intelligence actually is. However, as we have seen above, superintelligence and comparable concepts such AGI are typically defined by using 'human intelligence' as a reference point. They thereby feed off an intuitive form of understanding of the notion of 'intelligence', without, however, having to articulate in greater detail what exactly might be meant by that.

2) Claims pertaining to the potential consequences of the advent of a superintelligence. These claims are categorically different in so far as they typically rest on rather specific conceptions of intelligence, of what it is or does. Typically, as we have seen and as Russell et al. (2015) put it, "in this context, 'intelligence' is related to statistical and economic notions of rationality — colloquially, the ability to make good decisions, plans, or inferences".[593] Huw Price's pragmatic characterisation of intelligence can also be seen as an example for this type of perspective on intelligence.

In most versions of the AI-risk argument first the vague expectation that one day AI might exceed human intelligence is articulated and then, in a second step, it is explained, based on specific, economic conceptions of how intelligence typically exerts itself, why this prospect should concern us. Another way to qualify this distinction is to explain it in terms of the distinction between 'weak' and 'strong AI'. According to the textbook definition of AI the field of AI is divided into two

[592] Yudkowsky, E. (2008), p. 333.
[593] Russell, S., Dewey, D., Tegmark, M. (2015).

branches.[594] The first branch seeks to develop a better understanding of intelligence and cognition as such. It is characterised by the aim, as one author puts it, to find "the mark of the mental".[595] The second branch abstracts from such deeper theoretical problems and focuses on application, that is, on intelligent *action*. It aims, most generally put, "to engineer smart machines and applications" in order to solve specific practical problems by *emulating* intelligent behaviour. Luciano Floridi (2014) refers to these two different branches of AI research under the categories 'productive', or 'cognitive' AI' on the one hand side and 'reproductive' or 'engineering' AI' on the other hand side. Whilst cognitive AI aims to produce intelligence as such, engineering AI is not necessarily interested in what intelligence is but focuses on reproducing, i.e. emulating, its outcomes.[596] Russell and Norvig (2016) in their seminal textbook on AI similarly distinguish between AI research that focuses on 'thought processes' or 'reasoning', which they refer to as 'process-oriented AI', and AI research that focuses on behaviour, which they refer to as 'goal-oriented AI'.[597] Another common, and perhaps slightly more poignant way of phrasing this distinction is that between 'strong' and 'weak' AI. Weak AI, as Arkoudas & Bringsjord (2014) put it, "aims at building machines that act intelligently, without taking a position on whether or not the machines actually are intelligent", whereas strong AI aims at "building persons, period",[598] or, as philosopher of mind John Haugeland puts it, "the goal of this research is not merely to mimic intelligence or produce some clever fake. Not at all. AI wants only the genuine article: machines with minds, in the full and literal sense".[599] Whichever terms one decides to employ in order to label the two traditions in AI research (I will in the following use the distinction between 'strong' and 'weak AI'), the difference is clear: one seeks to study and recreate whatever intelligence *is*, whilst the other takes a behaviouristic, pragmatic stance on the problem, studying and trying to reproduce what intelligence *does*. This distinction allows us to see that the AI-risk argument oscillates between the two poles.

The 'vague expectations' of future superintelligence implicitly build on strong conceptions of AI because, by using human intelligence as a reference point, notions such as 'superintelligence' or AGI implicitly invoke and feed of an intuitive idea of what intelligence *is,* not only on what it *does.* Just consider the above quoted definitions of superintelligence: Soares and Fallenstein (2014) define superintelligence as an AI that is "smarter than the best human brains in practically *every* field",[600] Bostrom and Dafoe speak of superintelligence as a system that is "more cognitively capable than

[594] See for instance Frankish, K. & Ramsey, W. (2014), p. 1.

[595] Arkoudas, K. & Bringsjord, S. (2014), p. 34.

[596] Floridi, L. (2014), p. 140.

[597] More specifically he identifies four common categories of definitions of AI in the field that vary along these two dimensions 1) "systems that think like humans", 2) "systems that act like humans", 3) "systems that think rationally" and 4) "systems that act rationally". See Russell, S. & Norvig, P. (2016), p. 5, Figure 1.1.

[598] Arkoudas, K. & Bringsjord, S. (2014), p. 35.

[599] Haugeland, I. (1985), p. 2, p. 35, for an instructive discussion of complications in relation to such categorisations in the field of AI see Martinez-Plumed, F., et al. (2018).

[600] Soares, N. & Fallenstein, B. (2014), p. 1.

humans in all practically relevant domains",[601] and CSER defines superintelligence as being "superior to human performance in many or nearly all domains".[602] Intelligence itself, in all of these cases, is left unspecified, it is not actually explained what human intelligence consists in. However, this definition allows existential risk researchers to invoke all capacities of human intelligence, whatever their ultimate source may be, for speculations about hypothetical future AIs and to assume that they can be supercharged.

The risk itself, however, is explained based on a weak conception of AI as it treats it on purely pragmatic terms. I.e. it is argued that whether or not a superintelligence actually *is* intelligent in a substantive sense of the word is irrelevant for the argument. Intelligence is treated in line with the dictum that the "question of whether a computer can think is no more interesting than the question of whether a submarine can swim".[603] If anything, as we have seen, the reader is explicitly *discouraged* to anthropomorphise AIs,[604] that is, she is encouraged to think of superintelligence pragmatically, as a mindless and extremely powerful cross-domain optimisation process, hence along weak lines.

My reason to highlight this two-pronged argumentative strategy is not to commence a discussion of whether the AI-risk argument is consistent or not.[605] Rather, the reason is that its two dimensions and the manner in which they oscillate make the AI-risk argument profoundly interesting from a Heideggerian perspective. This is why they function as a bracket of my following discussion. Against the background of the preceding chapters, both dimensions are interesting and revealing in their own right and the argument as a whole, i.e. the oscillation between strong and weak conceptions of AI, can be seen as emblematic for the mechanisms of technological thinking as discussed by for instance Günther Anders. However, whilst the first dimension of the argument has in fact a rather long history with eschatological expectations surrounding AI dating back to the earliest days of the discipline, the second dimension, in particular its reflection in so-called 'AI safety' research, appears to be a much more recent phenomenon.

In the following sections I will proceed as follows. First, I will demonstrate that the first dimension of the argument, i.e. the vague expectations of future tipping points and discontinuities, as

[601] Bostrom, N. et al. (2016), p. 2.

[602] Cf. CSER (2018b).

[603] See Dijkstra, E. (1984), as cited in Floridi. L. (2014), p. 140.

[604] See for instance Yudkowsky, E. (2008); Bostrom, N. (2014), specifically ch. 7; Armstrong, S. (2013); Russell, S. (2017).

[605] I cannot discuss this point in greater depth in this chapter as the *validity* of the argument as such is not the subject of my discussion. My sense is, however, that this argumentative strategy is circular. Either it needs to be shown that we are getting closer to building strong AI based on weak conceptions of AI (including such things as consciousness, a sense of identity, multi-faceted goals and commitments, the ability to integrate abstract information with idiosyncratic, sensory and many more kinds of data, etc.) and thus that such a system can be built with technologies that are based on contemporary behaviouristic models of intelligence. Or one focuses on latter behaviouristic models of intelligence, in which case one cannot simply posit the capabilities of strong AI as a benchmark for thinking about the future capabilities of AIs. It will have to be demonstrated, that is, that superhuman levels of control over the environment can be achieved without understanding and reengineering strong AI first. According to several leading voices in the field, however, at the present moment in time not much points into that direction. As Floridi (2014) argues, when it comes to AGI, AI has not even joined the competition yet.

several authors have noted before, is *fideistic,* not necessarily in nature, but in structure.[606] What this tells us, however, is not that the belief cannot be true. Rather, what it reveals are a range of interesting assumptions about not only the nature of intelligence but about technology and science generally that appear to be underpinning contemporary discourses surrounding the future of AI. Against the background of my previous discussion of Heidegger, Arendt, and Anders this is interesting because it highlights that the basic assumptions they identified at the heart of the technological project are alive and well, that, in fact, they are underpinning the existential fears surrounding AI as well as wider macro-strategic discussions of existential risk.

In the second section I will then proceed to discuss the technical dimension of the AI risk argument. In the light of previous discussions, it is almost uncanny to see how closely the concerns of AI researchers echo Heidegger's, Arendt's and Anders' normative concerns about the potential effects of the technological understanding of being. We might hence understand the technical account of the existential risk from AI argument as a form of unintended concession to critical philosophers of technology.

4.3 Is the belief in the singularity fideistic?

The expectation of superintelligent machines has a rather long history. In fact, progress in what is now called AI has been accompanied by eschatological expectations ever since its official inception as an academic discipline at the 1956 Dartmouth Conference and arguably even for a longer time than that if one takes the literary genre into account.[607] The 'intelligence explosion' for instance is a concept that was famously coined in 1964 by the statistician and computer technology pioneer I.J. Good. In a paper on 'ultraintelligent machines' Good presented the reader with the following conjecture:

> "Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then

[606] See for instance Bringsjord, S., et al. (2012); Danaher, J. (2015); Lawrence, N. (2016); Floridi, L. (2015, 2016).

[607] Three pieces of science fiction conference clearly stand out as particularly prescient in the light of contemporary AI anxieties: First, Samuel Butler's 1872 novel '*Erewhon',* which portraits a society where people have become convinced by an argument that holds that machines will become autonomous and assume control unless their further development and possession is banned. Second, John Campbell's *'The Last Evolution'*, a short story of 1932, which in effect pre-empts I.J. Good's intelligence explosion. Third, the 1928 short story '*The Machine Stops'* by Edward M. Forster, which anticipates many contemporary concerns regarding potential psychological, social, cultural and political effects of the complete technification and automation of humanity's environment. 'The Machine Stops', however, should perhaps best not be understood as story about superintelligence but as a story about the reality we, according to many authors, already are beginning to live in; a world where we are embedded in a web of narrow AI's that govern each and every aspect of our lives and mediate our communication with one another, without, however, being governed by a single superior intelligence.

> unquestionably be an ‘intelligence explosion’, and the intelligence of man would be left far behind. Thus the first ultra-intelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control“.[608]

As can be told from the brief overview above, Good’s conjecture is still very much compatible with present concerns and expectations regarding AI. His definition of ‘ultraintelligence’ is essentially equivalent to today’s definitions of superintelligence and the scenario he describes foreshadows contemporary concerns about losing control and being left behind. Accordingly, Good is still regularly invoked by authors in the field.[609] Contrary to Good’s stance on the intelligence explosion, however, most contemporary accounts do not hold that ultraintelligence needs to be developed *before* an intelligence explosion can follow. Rather, it is now widely believed that a significant upsurge in general intelligence is likely to follow as soon as AI reaches levels of intelligence that are broadly comparable to human levels, i.e. once it reaches ‘human-level artificial intelligence’. In other words, the above-mentioned tipping point is now believed to be less far up the intelligence-ladder than Good might have thought.

Alan Turing, in as early as 1951, also seems to have entertained concerns comparable to today’s existential risk researchers when he argued that:

> “If a machine can think, it might think more intelligently than we do, and then where should we be? Even if we could keep the machines in a subservient position, for instance by turning off the power at strategic moments, we should, as a species, feel greatly humbled. ... [T]his new danger … is certainly something which can give us anxiety”.[610]

Generally speaking, many of the pioneers of early AI were convinced that human-level machine intelligence was imminent.[611] Herbert Simon and Alan Newell of RAND Corporation for instance stated in 1958 that “there are now in the world machines that think, that learn and that create. Moreover, their ability to do these things is going to increase rapidly until—in a visible future—the range of problems they can handle will be coextensive with the range to which human mind has been applied”.[612] Marvin Minsky, famously claimed in 1968 that “within a generation, I am convinced, few compartments of intellect will remain outside the machine’s realm – the problem of ‘artificial intelligence’ will be substantially solved […] we will have intelligent computers like HAL in the film 2001”.[613]

[608] Good, I.J. (1965), p. 32.

[609] See for instance Häggström, O. (2016), p. 102, fn. 235; Rees, M. (2017), p. 381, or Bostrom, N. (2009), p. 207, as well as Bostrom, N. (2014), p. 24.

[610] Turing, A. (1951), as cited in Russell, S. (2017), p. 179.

[611] See Armstrong, S. et al. (2014), Armstrong, S. & Sotala, K. (2015), and Dreyfus, H. (2012) for overviews of optimistic statements from the early phase of research in AI.

[612] Russell, S. (2016), p. 20.

[613] Minsky, M. (1968) as cited in Dreyfus (2012), p. 91, Minsky here refers to Stanley Kubrick’s film ‘2001: A Space Odyssey’, which features an intelligent, conscious board computer named HAL.

However, if grand expectations have accompanied AI since its very beginnings, so has criticism of such expectations. Hubert Dreyfus for instance claimed in 1968, referring directly to Minsky's above claim that

> "Minsky is typical of all workers in the area in giving two sorts of arguments in support of his view: (1) Empirical arguments based on progress achieved thus far, and (2) A priori arguments about what machines can in principle do […] I will argue that the empirical arguments gain their plausibility only on the basis of an appeal to an implicit philosophical assumption […]".[614]

The fact that we can distinguish between vague and specific conceptions of intelligence in today's AI risk argument indicates that Dreyfus's criticism is as poignant in the context of today's debates as it was at the time of his writing. The technical concerns surrounding future powerful cross-domain optimisers gain their plausibility only on the basis of an appeal to a far wider conception of superintelligence, which in turn hinges on the a priori assumption that human intelligence can be recreated by technological means. It is easy to be carried away by the technical dimension of the argument and the idea of an all-powerful optimisation process transforming the world into paperclips. But despite the fact that existential risk researchers seek to remain as technical in their analysis as possible and discourage from anthropomorphisms, the plausibility of their worst-case scenarios hinges entirely on the inherently anthropomorphic definitions of superintelligence quoted above and hence implicitly on the intelligence of actual *persons*, rather than the kind of mindless optimisation processes we find in narrow AIs. Consider Russell et al. (2015), who first relate intelligence to statistical and economic notions of rationality only to then, in order to highlight the stakes involved, claim that "everything that human civilization has to offer is a product of intelligence".[615] But is everything civilisation has to offer a product of statistical and economic notions of rationality? This is at least not entirely obvious and generations of scholars in for instance post-colonial studies have actually made the exact reverse claim, stressing the extent to which statistical and economic notions of rationality are a product of culture.[616]

We thus find the AI-risk argument meandering between two different conceptions of intelligence, a narrow and a wide one, depending on the context in question. Since the only general intelligence we know of is our own, and since we cannot yet tell with certainty what exactly it is in our brains that allows us to exert the high levels of control over the environment which is argued to be the defining characteristic of our intelligence, the AI risk argument hinges on the assumption that, whatever it may be that allows us to do this, can be artificially reproduced. Price, as we have seen, argues that modern science and technology have allowed us to attain unprecedented control over the environment. Science and technology are hence mainly understood as expressions of economic and statistical notions of rationality. However, it is somewhat of an open question what it actually is in us

[614] Dreyfus, H. (1968), p. 1.

[615] Russell, S. et al. (2015).

[616] For an interesting, critical discussion of the cultural roots of AI see Reilley, K. (2011).

that puts us into the position to engage in scientific research and how this capacity could be artificially reproduced. One only needs to think of Michael Polanyi's 'tacit knowledge',[617] or decades of research in science and technology studies in order to see how 'messy' innovation processes are, how hard it is to explicate how they unfold. Nonetheless it is a common assumption that AIs might one day be in the position to undertake research on their own and perhaps better than us. The anthropomorphic definition of superintelligence allows for this.

Empirically, however, as Floridi (2014) argues, research in strong, or 'productive' AI has so far been rather disappointing: "It does not merely underperform with respect to human intelligence; it has not joined the competition yet. Current machines have the intelligence of a toaster and we really do not have much of a clue about how to move from there".[618] As it happens, Floridi's perspective on the track-record of progress in strong AI appears to be widely shared by authors in the field.[619] The AI-risk argument then derives its plausibility mainly from its assumptions, that is, a priori arguments about what machines can in principle do, as Dreyfus puts it, rather than extrapolations from present trends.

Based on similar considerations, Bringsjord et al. (2012) argue that "the belief in the singularity is fideism". [620] The singularity is an elusive concept with a variety of different meanings within the futurist literature.[621] However, Bringsjord et al. define it narrowly as the "arrival on Earth of computing machines more intelligent, indeed vastly more intelligent, than human persons".[622] The notions 'singularity' and 'emergence of superintelligence' are thus used interchangeably. They further define 'the belief in the singularity' as the proposition that this event, i.e. the arrival of

[617] Polanyi, M. (2009).

[618] Floridi, L. (2014), p. 141.

[619] See for instance Deutsch, D. (2017); Bringsjord, S., et al. (2012); Dreyfus, H. (2012); Chalmers, D. (2010); Bostrom, N. (2014); Davis, E. (2014); Lawrence, N. (2016).

[620] Bringsjord, S. et al. (2012), p. 395.

[621] In the futurist literature it the concept of the singularity is used to describe a wide variety of different possible technological futures. The origins of the term, however, are in mathematics where it denotes a mathematical relation in which a given quantity goes to infinity in finite time [see Häggström, O. (2016), p. 109]. Because it posits an exponential growth of machine intelligence and computing speed within a short period of time, the above described intelligence explosion is often referred to under the label 'technological singularity'. A different scenario which trades under the label singularity is the transhumanist scenario, which predicts the merging of humanity and technology to the effect that all human limitations, physical and mental, will be overcome, resulting in a posthuman race and a 'biointelligence explosion' [See for instance Kurzweil, R. (2005); or Diamandis, P., et al. (2012)]. Generally speaking, accounts of technological singularities differ along several lines (regarding timescales and -lines, its nature, its causes, the transition process, whether the singularity is a state or a process, etc.) which I cannot cover here [For a more detailed discussions of differences and commonalities between different accounts of the singularity see for instance Eden, A. et al. (2012); Shanahan, M. (2015); or Callaghan, V., et al. (2017)]. At a minimum, however, accounts of the singularity appear to be united by the idea that humanity is either approaching or has already embarked onto a process of technological acceleration that will result in a discontinuity or a rupture in human history after which events will become radically unpredictable. This assessment is shared by Armstrong (2017), who claims that "a singularity in a model means that we encounter a breakdown of our ability to predict beyond that point. It need not mean that the world goes crazy, or even that the model does. But it does mean that our standard tools become inadequate for understanding and shaping what comes after. New tools are needed". We know these two characteristics already from the AI-risk argument, which, indeed, can be considered as a singularity hypothesis, based on the above characterisations. See Armstrong, S. (2017), p. 1.

[622] Bringsjord, S. et al. (2012), p. 397.

superintelligence, will occur in the not too distant future, that is, approximately within this century. As we have seen above, this proposition is widely held amongst the above listed AI-risk research institutes. We can hence understand Bringsjord's et al.'s discussion of the belief in the singularity as a discussion of the belief that the AI-risk argument should be taken seriously.

According to Bringsjord et al. the belief in the singularity rests on three propositions:

> (P1) There will be AI (created by HI), where HI refers to 'human intelligence' and AI to human-level artificial intelligence.
> (P2) If there is AI, there will be AI+ (created by AI), where AI+ stands for artificial intelligence that exceeds HI across all or most domains.
> (P3) If there is AI+, there will be AI++ (created by AI+), where AI++ refers to superintelligence, i.e. an AI that vastly exceeds HI on all levels.
> → There will be AI++, i.e. superintelligence will arrive and the singularity will occur.[623]

Bringsjord et al. then introduce a framework based on which they propose to assess whether, based on present best knowledge, the belief in the singularity qualifies as rationalist. In order for the belief to qualify as rationalist, according to the authors, its propositions must be either probable, beyond reasonable doubt, certain, or evident. Beginning with the strongest criteria, i.e. whether the propositions are either certain and beyond reasonable doubt, Bringsjord et al. arrive at the conclusion that (P1) is neither certain nor beyond reasonable doubt because the inverse proposition, i.e. that AI can *never* reach the level of HI, is neither logically inconsistent nor can it, at this point in time, be proved wrong based on empirical grounds because, according to the authors it is not beyond reasonable doubt that human minds process 'information in a manner above the Turing limit'. Turning to evidence, i.e. to the question whether the track record of machine intelligence to date suggests that we are approaching AI, the authors' judgment is even clearer, arguing that no data suggests anything pointing into that direction: "For the fact of the matter is that a sharp toddler of today makes a mockery of any computing machine with designs on natural-language communication. And even if we leave natural-language communication out of the picture and refer instead to human-level problem solving specifically in areas that would seem to be positively ideal for computing machines, we perceive not the steady advance of computing machines, but their paralysis when stacked against the capability of humans".[624] This brings us to the last criterion, the criterion for *weak* rationalism, which is the question whether (P1) - (P3) are at least more probable than not to hold.

[623] See Bringsjord, S. et al. (2012), p. 395. This schematised account of the singularity is based on Chalmers' (2010) formalisation of the argument. See Chalmers, D. (2010), p. 13.

[624] See Bringsjord, S. et al. (2012), p. 404, It is not my aim here to assess the accurateness of Bringsjord's evaluation of empirical evidence. However, it should be noted that much progress has been made since the article was published in 2012, specifically when it comes to language processing, which, from today's perspective might make Bringsjord's AI-toddler comparison appear spurious. Google for instance has recently released Google Assistant, which can autonomously execute phone calls and is capable to sustain conversations with individuals for prolonged periods of time. However, this does not necessarily take away from Bringsjords claim that evidence that AI is approaching HI is scarce, for his examples pertain to domain-specific, i.e. narrow AIs. Bringsjord's dismissal of the evidence would have been easier had he invoked the fact that barely any presently existing AI is capable of performing more than one well-specified type of task.

Bringsjord et al. list a variety of reasons based on which they seek to demonstrate that it is more probable than not that (P1) - (P3) does *not* hold, most of which are evidential and anecdotal. The most interesting one for the purposes of this chapter, however, is based on the observation that the belief in the singularity is premised on the 'concept of ever-increasing intelligence'. That is, it is based on a proposition (P4) that it is known that levels of intelligence can differ and in particular that machine intelligence can exceed human levels of intelligence.

This prompts the authors to ask the following question: "But if the proponents of the case in question know this [that there can be a qualitative difference between HI and AI++], then surely they must know what the difference in intelligence between HI and AI++ consists in. If they don't know what the difference consists in, then they aren't within their epistemic rights in asserting (P4)".[625]

Bringsjord et al. here in effect identify the same problem as the one underpinning above distinction between vague and the specific conceptions of intelligence that can be identified in the AI-risk argument. Superintelligence is defined as 'smarter than humans across all domains' without, however, specifying what that means, i.e. wherein the difference between HI and AI++ might consist. For Bringsjord et al. this inability means that those who believe that the singularity is near, or at least that it might potentially be near, are not within their epistemic rights to assert this since it means "to forge ahead and believe, in the absence of the normal prerequisites".[626] It means to make a concept without semantic content the baseline for thinking about future technological possibilities.

These considerations lead Bringsjord et al. to conclude that the belief in the singularity is fideistic. 'Fideism' is understood as the view that one ought to believe in the occurrence of an event that is pictured to be 'weighty, unseen and temporally removed' despite having little or no evidence for the proposition's correctness. 'Weighty' here is understood to indicate that the event is expected to be 'profoundly transformative', 'unseen' indicates that it is expected to involve 'beings or entities as of yet invisible', and 'temporally removed' simply means that the event is expected to occur at some unspecified point in the future so that the proposition cannot, at any given point in time, be proved wrong. The idea that we are approaching a point in time at which intelligence escapes its biological constraints clearly meets all of the authors' criteria. It is expected to be a profoundly transformative event in human history, it involves agents the precise nature of which cannot yet be specified and the point in time, at which their arrival is expected, too, is left unspecified. According to Bringsjord et al. fideism is the hallmark of religious beliefs, the belief in supernatural beings in the absence of any rational or empirical reasons that might sustain such beliefs. However, in stating that the belief in the singularity is fideistic, Bringsjord et al. do not mean to suggest that the belief in the singularity is esoteric or religious in nature, what they suggest is merely that the kind of argument

[625] Bringsjord, S. et al. (2012), p. 405.
[626] Ibid.

employed by those who think that the emergence of a superintelligence is imminent parallels fideistic types of arguments, i.e. that it is fideistic in *structure*.[627]

Bringsjord et al. are not alone in portraying contemporary expectations regarding AI along such lines. Oxford philosopher of technology Luciano Floridi speaks of those who believe that superintelligence is imminent as 'Singularitarians', claiming that they are "not unlike people wearing tin foil hats",[628] machine learning scholar Neill Lawrence calls them 'Singularians',[629] and philosopher Christian Munthe likens the logic of the AI risk argument to that of Pascal's Wager, asking why "ultimate harm advocates are not all attending mass".[630]

4.4 Fideism or ontological transparency?

However, there is a danger that in ridiculing above expectations surrounding AI and relegating them to the realms of obscurantism, the deeper significance of the underlying narrative is overlooked. What shines through in so called 'Singularitarianism' is arguably not so much a susceptibility for religious patterns of thought than what Heidegger calls 'the technological understanding of being' that finds itself confronted with the very dilemmas Heidegger, Arendt and Anders have brought out decades ago. Even if high expectations regarding the imminent arrival of AGI are currently not supported by empirical evidence, which is not the purpose of this chapter to make a judgment about, it appears worthwhile to take the AI-risk argument seriously because it gives clear sight on the mindset that, according to Heidegger, Arendt and Anders underlies scientific and technological progress in general. Once that is payed heed to, the fact that the underlying expectation can be labelled 'fideistic' becomes significant in its own right.

What sustains the AI-risk argument and the associated belief in the singularity in the first place is not so much a semi-religious sentiment than simply what Haugeland (1985) identifies as the defining assumption of research in AI *in general*. It is "the powerful suggestion that our own minds work on computational principles […] the theory that people are computers",[631] with the implication that the same scientific theory can explain processes in the brain as well as in computers.[632] In the field of AI and in the cognitive sciences this assumption is known as the computational theory of the mind.[633]

In the existential risk environment, this assumption translates almost seamlessly into the supposition that artificial intelligence performing on the level of human intelligence must in principle be possible. Huw Price calls this argument against sceptics 'the blow to the head': "the tricks are all

[627] Ibid, p. 399.
[628] Floridi, L. (2015).
[629] Lawrence, N. (2016).
[630] Munthe, C. (2015).
[631] Haugeland, J. (1985), pp. 5-6.
[632] Boden, M. (2006), p. 168.
[633] Rescorla, M. (2017).

there for our inspection: most of it is done with the glob inside our skulls. Understand that, and you understand how to do it artificially, at least in principle".[634] FHF institute adviser Olle Häggström argues along similar lines: "the blind forces of nature have succeeded in producing human-level intelligence this way, so we should be able to do it".[635] And Michael Anissimov (2012) responds to Bringsjord et al.'s article as follows:

> "Given the nearly universally accepted supposition in the cognitive sciences that intelligence is made up of a collection of mental routines that are fuzzy algorithms, plus the Church-Turing thesis, we get the conclusion that intelligence is indeed computable by standard Turing machines […] The notion that human beings are the only agents that can implement intelligence is being supplanted by the notion that intelligence is a bundle of algorithms that can be implemented by any suitable computer, whether carbon-based or silicon-based".[636]

Hence, the assumption is that, since brains and computers are considered to be the same thing, namely 'hardware', and since minds are mere 'collections of fuzzy algorithms', it should, in principle, be possible to reproduce what brains do by artificial means.

Once that assumption is made it is but a small step to propositions two and three in Bringsjord et al.'s formalised singularity argument, i.e. to the idea that AI will eventually supersede human levels of intelligence.[637] Häggström for instance proceeds to argue that "there seems to be no good reason at all to think that human-level intelligence is the maximal level attainable by a physical object in our universe: to think that no configuration of matter can, even in principle, achieve higher-than-human intelligence is just anthropo-hubristic and insane".[638] Arguments such as these typically are further backed up by an evolutionary perspectives on intelligence, i.e. the observation that intelligence levels have increased over time and vary not only between but also within species. The late Stephen Hawking for instance argued that "it's clearly possible for something to acquire higher intelligence than its ancestors: we evolved to be smarter than our ape-like ancestors, and Einstein was smarter than his parents".[639] From this perspective, the belief that AI will eventually supersede HI is not really comparable to fideism. Rather it means that basic assumptions and observations about the nature of cognition and intelligence are made explicit and the baseline for extrapolations into the future.

As Dreyfus pointed out in his response to Minsky's early hopes surrounding AI, these assumptions in turn hinge on an even deeper layer of assumptions about technology in general, on "a

[634] Price, H. (2013).

[635] Häggström, O. (2016), p. 107.

[636] Anissimov, M. (2012), p. 411, The Church-Turing thesis holds that any function that is computable at all must also be computable by a Turing machine. For more information on the Church-Turing thesis see e.g. Copeland, B.J. (2017), or Boden, M. (2006), p. 172 ff.

[637] See for instance Schneider, S. (2016, 2017) for discussions of why, based on present evidence, AI, once it has reached human levels of intelligence, is likely to rapidly exceed human intelligence levels soon after (intelligence being understood as instrumental rationality).

[638] Häggström, O. (2016), p. 101.

[639] Hawking, S. (2015).

priori arguments about what machines can in principle do".[640] The crux of the matter indeed lies in the phrase 'in principle'. Almost all of above quoted authors who argue that AGI should be possible employ that phrase, arguing that AGI should *in principle* be possible. But what does that phrase 'in principle' refer to? The principles of the universe, or nature, or intelligence? These would be very strong claims, epistemologically speaking. The most plausible answer is that it refers to the authors' own principles. What the above assertions reveal is less something about the nature of intelligence as such than about the authors' principles and assumptions regarding the relationship between science, technology and nature in general. According to these principles whatever exists in nature can be understood by the means of scientific inquiry and, in last consequence, must also be technologically reproducible. The fact that intelligence exists as a property of a physical object in the universe, by implication, means that it must also be technologically reproducible.

For O'Heigeartaigh of CSER, for instance, the fact that our brain exists is seen as a "proof of principle" that AGI is possible.[641] A priori, nature and technology are considered to be interchangeable, they are considered to be the same. How else could the fact that our brain exists count as a 'proof of principle' that *AGI* is possible? The a priori assumption about technology is that the realm of technological possibilities and the realm of natural possibilities is coextensive, governed and limited only by the laws of physics. Whatever is possible according to the laws of physics and not logically incoherent is considered technologically possible. Martin Rees's perspective on potential futures of humanity, is a case in point in that regard. In a recently published piece, Rees discusses the possibility of 'stellar-scale engineering', which might involve the technological exploitation and creation of wormholes and black holes. Rees acknowledges that, in spite of the fact that these speculative concepts are "far beyond any technological capability we can envisage" at this point, they are not in "violation of basic physical laws" and therefore fall within the space of what is deemed technologically possible. Rees takes things even one step further, wondering if the laws of physics we are presently aware of would prove to be immutable for an intelligence that is "able to draw on galactic-scale resources".[642] Hence, what the claim that AGI should '*in principle'* be possible really says, is that AGI *has* to be possible, or else the above authors' principles and assumptions regarding the very nature of reality would be false.

For, what would it mean if the production of AGI were assumed to be impossible? Theoretically speaking, there are at least two thinkable reasons for why one might do so. First, one might hold that true intelligence is too complex for us to ever fully understand how it works by scientific means. This would imply that we would also be incapable to purposefully reproduce it by technological means. Second, one might hold that, whilst intelligence may be fully explicable by scientific means, it may nonetheless be impossible to reproduce it technologically. The epistemic

[640] Dreyfus, H. (1968), p. 1.
[641] O'Heigeartaigh, S. (2017), p. 363.
[642] Cf. Rees, M. (2017), p. 393.

status of both of these two claims ultimately is an empirical matter. At this point, therefore, the question is what is reasonable to assume. If one were to assume that the first proposition holds, that would be tantamount to assuming that there are problems in the universe that are and will forever remain scientifically inexplicable. That would open the door wide for obscurantist world-views and undermine the status of science as a knowledge-seeking endeavour. If one were to assume that the second proposition holds [which is for instance Margaret Boden's position. Boden argues "that there is no obstacle in principle to human-level AI" but that most likely it is not "practically feasible" [643]], the situation would be equally dire.

As Arendt argues, it is 'the plumber', the engineer, who brings the findings of science down to earth: "the lost contact between the word of the senses and appearances and the physical world view has been re-established not by the scientist but by the 'plumber'. The technicians, who account today for the overwhelming majority of all 'researchers', have brought the results of scientists down to earth".[644] Technology provides us with empirical evidence for the fact that the results of scientific inquiry, i.e. knowledge claims about the reality "behind things as they reveal themselves naturally to our consciousness", are valid. In other words, in order to be sure that we really have understood something we need to be able to reproduce it technologically because that is our only visible evidence that we have truly identified the cause-effect relationships reigning in any given object of study. [645] Hence, if one were to purport to have understood the mechanics of human intelligence but nonetheless fail to reproduce it, that would leave us with the nagging feeling that we might actually not *really* know how it works. It is in that sense that Arendt argues that homo faber "can know only what he has made himself".[646] In that light it would not be the case, as Bringsjord et al.'s argument seems to imply, that the belief in superintelligence speaks of irrational hopes or fears, as the likening to fideism seems to suggest. Rather, the opposite appears to be the case. The belief in superintelligence speaks of modern humanity's *rational* hopes, according to which science and technology, in principle, are hoped to allow us to understand and reproduce every causal relationship that occurs in nature.

In short, in the AI-risk argument we find an unblemished articulation of what Arendt and Anders consider the base assumption of the modern belief in science and technology, namely that "everything is possible and that whatever is possible will ultimately be done".[647] Dries (2012) refers to this as the 'Pandynathos-principle' of modernity – "the 'idée fixe' of homo faber" – the diagnosis of which, according to him, was the common denominator of Arendt's and Anders' critique of modernity.[648] It is no coincidence then that the Pandynathos-principle is echoed closely in Bostrom's

[643] Boden, M. (2016), pp. 153-154.
[644] Arendt, H. (1963), p. 49.
[645] Ibid.
[646] Arendt, H. (1998), p. 295.
[647] Dries, C. (2012), p. 343.
[648] Ibid.

technological completion conjecture, which (as discussed in chapter 1, p. 40) holds that "if scientific and technological development efforts do not effectively cease, then all important basic capabilities that could be obtained through some possible technology will be obtained".[649]

However, beneath the Pandynathos-principle, of course, we find what Heidegger considers to be the essence of technology: Enframing. Everything seems possible because the technological ontological condition has us conceive of all beings as essentially the same, namely as mere material, standing reserve that can be ordered at will. The above invoked quotations could provide no clearer illustration in that regard. Intelligence is regarded as 'a fuzzy set of algorithms' and the brain as just a 'glob', "wet hardware we have inside our skulls",[650] a "thing inside our cranium",[651] a "blob of matter",[652] or a particular "configuration of matter".[653] What makes the brain and, by extension, us different from the rest of nature is merely the particular way in which matter happens to be organised in our brains.[654] Once that assumption is made, it is of course but a small step to the proposition that this configuration can be replicated and perhaps even optimised.

From a Heideggerian perspective we are encountering in AI, understood as a field of research, the final stage of the technological danger in which both, the "subject and the object are sucked up as standing reserve" because intelligence, the hallmark of the human, finds itself included in the general calculus of objectification and ordering.[655] It is therefore interesting to see that Heidegger's fears regarding the technological understanding of being are echoed in the fears of AI-safety researchers.

4.5 Technology awakes

There could arguably be no better illustration of Heidegger's fears regarding the ultimate effects of technological thinking than Bostrom's 'paperclip maximiser' - a mindless optimisation process, the embodiment of calculative thinking, that treats everything it encounters as mere standing reserve that can be mobilised for some random optimisation purpose. We find in the AI-risk argument an embryo version, a glimpse of awareness, of the type of concerns regarding technology as destiny that were entertained by Heidegger. My previous discussion of the instrumental understanding of technology, allows us to see why such a glimpse of awareness might come about in the case of AI. The reason is that, in the case of AI, 'technology awakes'. With superintelligence, the existential risk community literally envisions a technology to become a fully autonomous agent that acts freely in the world. It is

[649] Bostrom, N. (2009), p. 192.
[650] Rees, M. (2017), p. 382.
[651] Bostrom, N. (2014), p. 0.
[652] Tegmark, M. (2017), p. 70.
[653] Häggström, O. (2016), p. 101.
[654] Tegmark, M. (2017), ch. 2.
[655] Heidegger, M. (1977), p. 173.

contradictory to consider an autonomous agent a mere tool and therefore, in the case of superintelligence, the instrumental understanding of technology necessarily reaches an endpoint. In speculating about superintelligence, existential risk researchers implicitly speculate about post-tekné futures and therewith about futures which, for Heidegger, Arendt and Anders, we since long inhabit. As we have seen, in their view technology has ceased to be a form of tekné. i.e. a mere means to an end in the service of human purposes, a very long time ago.

Anders for instance argued that technology had become "the subject of history, alongside which we are merely co-historical".[656] If superintelligence were to become reality in the way envisioned by some AI researchers, this idea would obviously find its ultimate vindication – a technological artifact would become the subject of history due to its superior ability to control its environment. In the case of superintelligence, then, existential risk researchers begin to be concerned about technology along similar lines as Heidegger and Anders because they, too, begin to think about technology as force that has its own momentum, developing and reshaping the world in ways that might not only be independent of human ends but run counter to them. It is therefore interesting to see that in this specific case the normative concerns of the two camps begin to converge as well.

In the case of AI, it becomes impossible to abstract from the ontological complications which play such a pivotal role in Arendt's, Anders' and Heidegger's thinking about technology. The question AI safety researchers are asking is in effect equivalent to the one that preoccupied Heidegger, Arendt and Anders: how can we square the logic underpinning the technological mindset with the ordinary way in which humans relate to and act in the world? The reason is that AI, as a technology, is *about* ontology. It is applied ontology.[657] The very aim of the field is to create agents that perceive and process information in a given environment and are capable to act in it in order to solve specific tasks.[658] This form of interaction with their environment of course requires the AI to discriminate between aspects of reality that are deemed relevant and aspects that are deemed irrelevant in any given situation. The fears surrounding AI can be broken down to the concern that an all-powerful AI is released into the world that restructures it based on an overly reductive model of reality, regarding too many aspects of reality as irrelevant without humanity being able to intervene and correct its course of action.

This is the second reason why the chapter is called 'technology awakes'. Not only is *a* technology envisioned to come to life - it can also be argued that the *technological understanding of being* is envisioned to come to life. With some caveats, the danger which existential risk researchers have identified in AI is structurally the same as the ontologically rooted dangers that Heidegger, Arendt and Anders have associated with modern technology at large. It is the problem of

[656] Anders, G. (1992), p. 2.

[657] For a discussion of the relevance of 'applied ontology' in the context of artificial intelligence see for instance the journal 'Applied Ontology'. Here ontology is "understood as a general theory of the types of entities and relations that make up their [e.g. an AI's] respective domains of inquiry". See IOS Press (2018).

[658] Dreyfus, H. (1968); Haugeland, J. (1985).

‘metaphysical violence’, of reductionistic representations of reality, giving rise to the fear that all that is meaningful about life could be squeezed out, if not entirely destroyed.

Günther Anders argues that, if the atom bomb could speak, it would say: “It’s all the same. Whether the world exists or not - it is the same. Why should the world not just as well not exist”? [659] In other words, for Anders, the nuclear bomb embodied “pure nihilism”.[660] Anders locates the origin of nihilism in metaphysical monism, the idea that “everything is of the same kind: namely nature”,[661] and traces it back to the shock past generations felt when they were confronted with this monistic perspective on reality. Where previously the world had meaning and was thought to be governed by the laws of god, one suddenly had to come to terms with the fact that it is governed by the laws of physics, “laws without a lawmaker”,[662] for which everything is reduced to mere matter and hence essentially the same, of the same kind, irrespective of whether it is a stone, a tree, or a human being. The maxim of the nuclear bomb, Anders claims, is identical with that of monism, or nihilism – “it behaves like a nihilist in that it regards and treats everything, irrespective of whether it is a human or a machine, a loaf of bread or a book, a house or a forest, an animal or a plant, as the same, namely as nature; in this case this means: as something that yields itself to radium poisoning”.[663] Anders here closely echoes Heidegger for whom, as discussed in chapter 2, the nuclear bomb also was only the “grossest of all gross confirmations of the long since accomplished annihilation of the thing: the confirmation that the thing as a thing remains nil".[664] Hence, when Yudkowsky (2008) summarises the existential fears surrounding AI with the words “the AI neither hates you, nor loves you, but you are made out of atoms that it can use for something else”,[665] this is uncannily close to Anders’ atom bomb monism. For Yudkowsky’s AI, or Bostrom’s paperclip maximiser, just as for Anders’ bomb, everything is the same: atoms, particles of matter that can be used for some generic purpose, be it paperclip maximisation or radium poisoning and irrespective of whether they are dealing with human beings, trees, or stones.

Without being aware of it, such fears could hence serve as text-book examples for Heidegger’s, Arendt’s or Anders’ shared concerns regarding the inherently ‘dehumanising’ nature of the ontological gaze underlying modern technology and science. For, if we unpack concepts such as ‘value-alignment’ or ‘orthogonality thesis’ from a Heideggerian perspective, what do we find? What we find are the theoretical puzzles of the ‘technological understanding of being’ that Heidegger, Arendt and Anders found themselves confronted with and that were discussed under such labels as objectlessness, standing-reserve, or earth-alienation. Roden (2015) argues that “we have no *a priori* assurance that the phenomenology of a successful AGI will correspond to human

[659] Anders, G. (1958a), p. 301.
[660] Anders, G. (1962), p. 504.
[661] Ibid, p. 300.
[662] Ibid.
[663] Ibid.
[664] See Heidegger, M. (2009), p.168.
[665] Yudkowsky, E. (2008), p. 333.

phenomenology".[666] Roden here touches upon the heart of the value-alignment, or orthogonality problem, which, from a Heideggerian perspective, should in the first place not be considered as an ethical problem, as the notion of 'value-alignment' suggests, but as an ontological one.

From a Heideggerian perspective (cf. chapter 2, section 2.4) the problem is less what we consider to be valuable or not than what can serve as common reference units of value under conditions of modern technology. The inherent property of modern technology is that it dissolves everything that could serve as reference units into negotiable states of affairs, mere processes and causal relationships. Hans Jonas pointedly summarises Heidegger's characterisation of this metaphysically 'violent' gaze of the technological ontological condition in a brief piece on the analytical method, which he understood to be the basis of modern science and technology. "Analysis", Jonas argues,

> "has been the distinctive feature of physical inquiry since the seventeenth century: analysis of working nature into its simplest dynamic factors. These factors are framed in such identical quantitative terms as can be entered, combined, and transformed in equations. The analytical method thus implies a primary ontological reduction of nature, and this precedes mathematics or other symbolism in its application to nature. Once left to deal with the residual products of this reduction, or rather, with their measured values, mathematics proceeds to reconstruct from them the complexity of phenomena in a way which can lead beyond the data of the initial experience to facts unobserved, or still to come, or to be brought about. That nature lends itself to this kind of reduction was the fundamental discovery, actually the fundamental anticipation, at the outset of mechanical physics. With this reduction, "substantial forms," that is, wholeness as an autonomous cause with respect to its component parts, and therefore the ground of its own becoming, shared the fate of final causes […] the aristocracy of form is replaced by the democracy of matter".[667]

In the field of AI, the 'democracy of matter', i.e. of nature's simplest dynamic factors, is reflected in the democracy of data, of "0s" and "1s". The world of any given AI is in the first place an environment of "0s" and "1s", data-points that have no inherent meaning, relevance, or significance. The puzzle AI safety researchers find themselves confronted with is in effect the question how to reconstruct what Jonas calls an 'aristocracy of form', which comes naturally to the human being and constitutes the every-day reality we inhabit, out of the democracy of data-points, which is the base-reality of AI. The deeper issue at which the value-alignment problem and orthogonality thesis point is that, per default, the phenomenology of an AGI must be expected *not* to correspond with human phenomenology, to the effect that the reference units of its interaction with its environment must be expected to differ radically from those of humans.

It is telling in that context that, for Yudkowsky, the default reference units of his hypothetical superintelligence's interaction with physical reality are atoms, rather than for instance human beings. From a Heideggerian perspective this is no surprise at all, given that the notion of 'human being' is

[666] Roden, D. (2015), p. 171.
[667] Jonas, H. (1959), p. 140.

thoroughly imprecise and amorphous when judged by technical standards. It refers to a phenomenon, an amorphous mental construct that has immediate meaning only within the frame of reference of the intuitive, non-analytical human mode of being-in-the-world. As Braidotti (2015) argues "we assert our attachment to the species as if it were a matter of fact, a given. So much so that we construct a fundamental notion of Rights around the Human" and yet, "the concept of the human has exploded under the double pressure of contemporary scientific advances and global economic concerns".[668] Such discussions provide us with an idea of how problematic the very concept of 'human being' is once exposed to analytical scrutiny. From an AI perspective the situation is even more problematic. AI, as a scientific discipline, is left to deal with what Jonas calls 'the residual products of the analytical reduction of reality' and thus it finds itself confronted with the task to install an understanding of phenomenal reality in its products from scratch; based on the measured values of reality's simplest dynamic factors.[669] Before an AI could conceive of human beings as reference units for its interaction with reality, it would first have to know what exact configurations of the simplest dynamic factors of its models of the world qualify as 'human beings' in any given context. As Yudkowsky's characterisation of the existential risk from AI exemplifies, prima facie, for a hypothetical superintelligence all that exists is matter - patterns of atoms. So, if we want it to be 'safe', it needs to have some form of understanding of what within this otherwise meaningless ocean of data is relevant for its calculation and what is not: "The unFriendly [sic.] AI has the ability to repattern all matter in the solar system according to its optimization target. This is fatal for us if the AI does not choose specifically according to the criterion of how this transformation affects existing patterns such as biology and people".[670] Just as a self-driving car needs to be equipped with specific instructions or learning-algorithms that allow it to distinguish one sequence of pixels from another one and which tell it how this is to affect its course of action, a hypothetical superintelligence would have to learn how to distinguish one sequence of atoms from the other and how that is to affect its course of action in any given context. What this tells us is that before we can even begin to speak about 'value-alignment', we first have to speak about 'ontology-alignment'. In some AI-safety research circles this problem is beginning to be discussed under the category of 'Realistic World

[668] Braidotti, R. (2013), pp. 1-2.

[669] Depending on a given AI's task and the environment within which it is employed, the relevant factors in question may of course differ. They can for instance be pixels (as in self-driving cars), or clicks (as for instance in ad-generating algorithms), or sound signals (as for instance in speech-recognition software), etc. A hypothetical superintelligence, however, as we have seen, is defined as a superhuman *general* problem solver. That is, authors in the field envision it to have a better-than-human instrumental understanding of reality in general, not just of a digital environment, or of a board-game environment, but of reality in all its dimensions and all its intricacies. This means that the simplest dynamic factors relevant for its calculations must be assumed to be atoms, which are the simplest dynamic factors in nature we are presently aware of.

[670] Yudkowsky, E. (2008), pp. 332 – 333.

Models'.[671] Worley (2018) for instance concedes that "for an agent to be alignable it must be phenomenally conscious".[672]

This, however, amounts to no less than the task to reproduce our ordinary ways of making sense of the world based on the technological understanding of being and thus to overcome the schizophrenic condition of modern human existence. AI safety researchers, it seems, begin to become aware of exactly the kind of puzzles and perplexities Heidegger, Anders and Arendt had begun to raise attention for decades ago, asking in effect the same question: How can we align the different kinds of knowledge, the different relations to reality, underlying modern human existence, if we are to avert catastrophe?

4.6. Existential fears and existential hopes

But it is not only the potential for catastrophe that existential risk researchers see in AI. Rather, the typical position in the existential risk eco-system is that the emergence of superintelligence can either result in an existential catastrophe, or in what Cotton-Barratt and Ord (2015) call an eucatastrophe: "an event which causes there to be much more expected value after the event than before".[673] That is, the expectations surrounding AI are extremely binary; they are saturated both with existential fears and existential hopes. The late Stephen Hawking for instance (who acted as an advisor to CSER as well as the FLI) claimed that superintelligence likely "will be either the best, or the worst thing, ever to happen to humanity".[674] Olle Häggström of the FHF similarly states that "it may well be that we are standing at or very near a decisive turning point that can lead either to our prompt extinction or to a future where we flourish beyond our wildest dreams, perhaps on cosmic scales. Let us not sit idly by as the future unfolds".[675] And Max Tegmark, speaking on behalf of the participants of an AI-safety conference that brought together the most important names of the scene,[676] states that "we might create societies that flourish like never before, on Earth and perhaps beyond, or a Kafkaesque global surveillance state so powerful that it could never be toppled".[677] The one thing that appears to be generally agreed upon then is that, if superintelligence materialises, it will be what Bringsjord et

[671] For recent research on 'Realistic World Models' see for instance Soares, N. (2015b); Legg, S. & Hutter, M. (2007b); Orseau, L. & Ring, M. (2012); Bensinger, R. (2013).

[672] See Worley III, G. (2018).

[673] Cotton-Barratt, O. & Ord, T. (2015), p. 3.

[674] Hawking, S. (2015).

[675] Häggström, O. (2016), p. 149.

[676] In 2015 the FLI organised a conference on "The Future of AI: Opportunities and Challenges", which by many is regarded as the birth-date of the AI-safety research community. The conference brought together some of the most important names of the existential risk community, such as Nick Bostrom, Huw Price, Seán Ó hÉigeartaigh, Anders Sandberg, Eliezer Yudkowsky, Stuart Armstrong, or Luke Muehlhauser, as well as some of the most prominent AI scholars and tech-entrepreneurs, such as Margaret Boden, Demis Hassabis of Google DeepMind, Francesca Rossi of IBM, Elon Musk, and even Google-founder Larry Page. See Tegmark, M. (2017), p. 404, fig. 9.1.

[677] Tegmark, M. (2017), p. 52.

al. call a ‘weighty’ event, a cataclysmic event as a result of which humanity will find itself in a categorically different state of existence.

These expectations can best be understood by linking them back to Bostrom’s technological completion conjecture. As discussed in section 4.4, for many authors in the field, given their assumptions about what technology can in principle do, epitomised perhaps most pointedly in Bostrom’s technological completion conjecture, the arrival of superintelligence is assumed to be in principle inevitable (under the condition that we are spared an existential catastrophe). In line with the Pandynathos principle, the belief is that “everything is possible and that whatever is possible will ultimately be done”.[678] From this perspective, the question is not *if* we are going to develop superintelligence but only *when* and, above all, *how*. In other words, given that the arrival of superintelligence is expected to transform the human condition categorically, the entire future of humanity is collapsed into a single coding problem: either we get the utility function of the first AGI (the so-called ‘seed AI’) right and “we flourish beyond our wildest dreams”, or we get it wrong, in which case we may have to face “our prompt extinction”.[679]

As touched upon in previous chapters, in the macro-strategy of existential risk research, superintelligence assumes a central role. Given that it is understood as a superhuman problem solver, it is also expected to be better at dealing with existential risk (under the condition that it is a benign superintelligence). Bostrom (2014) for instance claims that superintelligence could:

> “reduce many other existential risks. Risks from nature—such as asteroid impacts, supervolcanoes, and natural pandemics—would be virtually eliminated, since superintelligence could deploy countermeasures against most such hazards, or at least demote them to the non-existential category (for instance, via space colonization) […] But superintelligence would also eliminate or reduce many anthropogenic risks. In particular, it would reduce risks of accidental destruction, including risk of accidents related to new technologies. Being generally more capable than humans, a superintelligence would be less likely to make mistakes, and more likely to recognise when precautions are needed, and to implement precautions competently“.[680]

Yudkowsky (2008) goes even further than that, arguing that:

> “To survive any appreciable time, we need to drive down each risk to nearly zero. ' Fairly good' is not good enough to last another million years […] Such competence is not historically typical of human institutions […] If we postulate that future minds exhibit the same mixture of foolishness and wisdom […] as the minds we read about in history books - then the game of existential risk is already over; it was lost from the beginning. We might survive for another decade, even another century, but not another million years. But the human mind is not the limit of the possible […] With luck, future historians will look back and describe the present world as an awkward in-between stage of adolescence, when humankind was smart enough to create tremendous problems for itself, but not quite smart enough to solve them. Yet before we can pass out of that stage of adolescence, we must, as adolescents, confront an adult problem: the

[678] Dries, C. (2012), p. 343.
[679] Häggström, O. (2016), p. 149.
[680] Bostrom, N. (2014), p. 265-266.

> challenge of smarter-than-human intelligence. This is the way out of the high-mortality phase of the life cycle, the way to close the window of vulnerability […].[681]

In brief, artificial intelligence is hoped to provide us with a path out of Anders' 'Age of Respite'. For Anders the 'Age of Respite' has to be considered as humanity's last age because, no matter for how long it would last, its "differencia specifica, the possibility of self-extinction can never end but by the end itself". [682] In 1980, in *Die Antiquiertheit des Menschen II*, Anders reinforced this point, arguing that his "portrait of the contemporary human […] depicts not only the human of today but also the human of tomorrow and the human of the day after tomorrow". He thus claims that his philosophical anthropology of technocracy is "a final and definitive portrait" of the human precisely because the defining feature of modern humanity's mode of existence – that we have to think of ourselves as living in the end-time – cannot end but by the end itself and therefore "will forever remain a final time".[683]

I have argued in chapter 3 that existential risk theory might initially be conceived of as an embodiment of this new temporality and that existential risk research's mission could be summarised as 'trying to make the time of the end endless'. However, macro-strategic conjectures about AI, such as Bostrom's or Yudkowsky's, demonstrate that they have in fact not fully accepted the temporality of the 'Age of Respite'. Of course, existential risk researchers do not articulate the hope that humanity may one day be able to regain the sense of eternity that characterised the temporality of past generations according to Anders and Arendt, in which the future could simply be taken for granted and 'came' by itself. But visions of 'technological maturity' and of a benevolent superintelligence do speak of the hope that the phase of imminent danger which characterises humanity's situation since the beginning of the atomic age may one day be left behind and that the 'Age of Respite' will have been but a 'high-mortality phase' and a 'window of vulnerability'. What is hoped for is, if not collective immortality, at least that humanity can find a modus-vivendi that allows for a degree of permanence and stability which would allow 'the human of tomorrow and the human of the day after tomorrow' to feel as part of a world more permanent than herself again.

To put it in Anders' words, it speaks of "a nostalgia for finitude, the good old finitude of the past" in which mortality was reserved for the human individual.[684] This nostalgic hope underscores Anders' analysis that we are in fact unable to accept the nihilism which atomic bombs and existential risk embody. It underscores his analysis that modern man "could appropriately be described as the titan who strives desperately to recover his humanity".[685] Anders does not indicate if he thought of any specific titan in this allegory, but the titan who intuitively comes to mind is Atlas – the titan whom Zeus condemned to carry the skies without salvation in sight. As Atlas, under Anders'

[681] Yudkowsky, E. (2008), p. 341-342.
[682] See Anders, G. (1962), p. 494.
[683] Anders, G. (1992), p. 10.
[684] Anders, G. (1956b), p. 147.
[685] Ibid.

conception, humanity faces an open-ended end-time, where an end to its total responsibility would be tantamount to the end of the world. With AI, Yudkowsky and Bostrom appear to hope, we might be able to invent an artificial Atlas who would take our place, liberate us from our infinite responsibility, and allow us to be human again.

This throws light on the deep irony at the heart of macro-strategic existential risk research, because what would it mean to be human in the age of superintelligence? Would humans melt with the machines or perhaps even become machines themselves? Would humans live alongside machines more intelligent than themselves? What kind of existence would that that be? The scope of the chapter does not allow for a discussion of positive visions of post-intelligence-explosion or singularity futures. But one thing appears to be clear, namely that they necessarily will be post-human futures too.[686] From an existential risk perspective, then, humanity's best and perhaps only hope to rescue itself is to end itself.

[686] For instructive scholarly discussions of the concept of the singularity see for instance Callaghan, V. et al. (2017) and Eden, A. et al. (2012).

Conclusion

Throughout the past chapters, I hope to have demonstrated that the emerging genre of existential risk research presents us with a rich, new background for political-theoretical reflection and that it can be meaningfully related to long-standing debates in political theory and philosophy. Specifically, I hope to have shown that it resonates with and presents us with a new starting point for reflecting about old puzzles regarding the role of technology in our lives and that, indeed, existential risk research highlights the lasting relevance and insightfulness of the works of authors such as Martin Heidegger, Hannah Arendt and Günther Anders. In conclusion one might argue that what existential risk research allows us to see is that, whilst the exact risks it discusses, and the integrative approach of studying them, i.e. its macro-strategy, might be new, the underlying problem it confronts us with, the problem of 'technology as destiny', is not qualitatively different today than it has been in the past. If anything, existential risk research highlights that the central puzzles regarding the role of technology in human affairs which Heidegger, Arendt and Anders uncovered and the categories which they introduced to discuss them, categories such as 'objectlessness', 'Promethean shame' or 'technology as action', are almost more applicable now than then.

One of the first analytical philosophers of technology, Henryk Skolimowski (1966), held that technology is categorically different from science in so far as science concerns itself with "what is" whereas technology concerns itself with "what is to be".[687] Skolimowski's characterisation of technology highlights that technology is an inescapably ontological as well as normative enterprise. It requires an (oft unarticulated) positioning towards the 'what', i.e. towards what we consider as real and relevant in any particular moment in time, as well as towards the 'is to be', i.e. towards what we want reality to be like. Heidegger, Arendt, and Anders most certainly would not have disagreed with Skolimowski's assertion that technology is concerned with the question 'what is to be?' However, they would have stressed that it is concerned with that question in a highly problematic fashion, because, from their perspective, the *primary* concern of technology is with 'what is possible'. Modern technology mainly concerns itself with providing us with more options, i.e. with more efficient and faster access to material goods, information, etc. Since, from the perspective of the modern technologist the space of technological possibilities is limited only by the laws of logic and physics,[688] the base assumption regarding 'what is to be' is, prima facie, that "everything is possible".[689] "If there is anything that modern man regards as infinite", Günther Anders argues, it is

[687] Skolimowsky, H. (1966), p. 375, For a discussion of Skolimowski's important role in the history of philosophy of technology see Franssen, M., Lokhorst, G., et al. (2018).

[688] See for instance Bostrom, N. (2013), Rees, M. (2017), or Häggström, O. (2016).

[689] As discussed at a previous point Martin Rees (2017) for instance discusses the possibility of 'stellar-scale engineering', involving for instance the deliberate creation of wormholes and black holes. In spite of the fact that these speculative concepts are "far beyond any technological capability we can envisage", Rees claims,

no longer God; nor is it nature, let alone morality or culture; it is his own power".[690] As a result of this ontological shift in perspectives, all givens are conceived of as an in principle negotiable states of affairs.

From the perspective of the authors I have covered here, the central puzzle of modern technology thus resides in the fact that technology's concern with 'what is possible' has inherently undermining implications for the concern with 'what is to be'. Put simply: if everything is possible, why should anything be what it is and not some other way? Beneath this problem we find a more basic ontological problem, namely, as Hannah Arendt puts it, the shift in attention "from the search after the 'What' to the investigation of 'How'" and thus "from interest in things to interest in processes, of which things were soon to become almost accidental by-products".[691] It is only with this ontological shift in interest from the 'what' to the 'how', from things to an awareness of their process-character that the idea of 'everything is possible' became possible. It led to the realisation that, once causalities are fully understood, in principle everything in nature can be reproduced, altered and manipulated, to the effect that no state of affairs is set in stone. Hence, the idea of prima facie omnipotence goes hand in hand with an ontological transformation whereby things and objects recede and are replaced by processes and causal-functional relationships, to the effect that the reference units of the investigative pronoun 'what' recede and dissolve into a collection of causal processes. As a result, the question 'what is to be' has, by definition, become anachronistic under conditions of modern technology. Modern technology, being concerned with processes, cannot make sense of the very idea of 'whatness', or 'wholeness', echoing Heidegger's claim that the technological understanding of being undermines the very conditions from which any *ordo*, any rank and recognition, any normative orientation in the world, could arise.

In that light it is both consistent and ironic that, once infused with historical and political-theoretical context, we immediately find existential risk research confronting us with this very puzzle. It is ironic because 'what is to be?' without doubt can be considered as existential risk research's central concern: What is to be? Is there to be life on Earth or nothingness? Existential risk research is born out of a concern for *something,* for humanity and the preservation of value and of sources of meaning in the universe. Yet, as a result of its ambition to be as scientifically rigorous in its methods as possible,[692] it becomes difficult to determine what notions such as 'humanity', 'value', or 'meaning', refer to and, in consequence, to determine how the very concept of existential risk should be defined.

From a Heideggerian perspective the source of this irony is obvious. In its aspiration to become a new scientific discipline, existential risk research approaches the question of the future of

they are not in violation of "basic physical laws" and that therefore these visions do move within the space of technologically possible futures. Cf. Rees, M. (2017), p. 393.

[690] Anders, G. (1956b), p. 146.

[691] Arendt, H. (1958), p. 585.

[692] Torres, P. (2017b).

humanity from what Arendt calls 'the point of the universe'. As such, it is based on the process-oriented ontology of modern science and technology and uses the abstract space of technological possibilities and its defining principles "everything is negotiable" as a benchmark for thinking about the future. From this perspective, humanity is transcended and turned into but another negotiable process, assumed to change over time due to biological as well as technological evolution, and thus reduced to but one of many thinkable forms of meaningful conscious experience that could theoretically be technologically produced.

In other words, by choosing technology as its benchmark for thinking about the future, existential risk research draws on a logic which makes it impossible for it to identify a specific reference unit of its concerns, an Archimedean point of value, apart from technological development itself. Lacking a reference unit for visions of desirable futures, for 'what is to be', technology and its prima facie limitless possibilities turn into a conceptual placeholder for the very idea of future value. In that light it is no coincidence that the concepts of 'existential catastrophe' and 'end of technological development' are synonyms in existential risk theory (any existential catastrophe would involve an end of technological development, otherwise it would not qualify as an existential catastrophe, and an end of technological progress would deprive humanity of its chance to evade its otherwise naturally preordained doom).[693] The central irony, or paradox that we find at the heart of existential risk research is, then, as Dupuy (2009) puts it, that the "overweening ambition and pride of a certain scientific humanism leads directly to the obsolescence of humankind".[694] Heidegger, Arendt and Anders would have gone even further than that . For them it is not due to its ambitions and its pride that scientific humanism has that effect; it follows immediately from its ontological nature. Anders calls this the 'telescopical gradient' – the disjunction between the magnitude of what we can produce and the significance we attribute to us and our existence in the universe. Nobody, he argues, when gazing through the telescope into the universe suddenly feels larger than before. On the contrary, confronted with the infinite expanses of the universe, it is as if the universe through the telescope stares back at humankind, shrinking it by the same measure by which it was expanded in our telescopically enhanced vision.[695] From that perspective, the label 'scientific humanism' would have to be considered an oxymoron to begin with.

[693] Cf. Chapter 1.

[694] Dupuy, J.-P. (2009a), p. xiv.

[695] Anders, G. (1969), p. 822, translated by the author. Cf.: "Was ich meine, ist die Differenz zwischen der Größe dessen, was wir herstellen können, und der Relevanz, die wir uns selbst und unserer eigenen Existenz im Weltganzen einräumen. Es kann nämlich keine Rede davon sein, daß wir, je mehr wir leisten, in unseren Augen um so wichtiger werden. Die Regel, die hier gilt, besagt sogar umgekehrt: Je höher unsere naturwissenschaftlichen und technischen Leistungen steigen, je enormer diese werden, um so kleiner ist die Funktion, die wir uns selbst als Mitspielern im Universum zugestehen. Niemand, der durch ein Teleskop blickt, fühlt sich angesichts des plötzlich maßlos erweiterten maßlos erweiterten Himmelsraums und der vielfach vergrößerten Himmelskörper größer als vorher. Die Wirkung ist umgekehrt so, als wenn der Himmelsraum durch das Rohr auf uns zurück blickte und uns um so viel kleiner machte, als er durch unseren teleskopischen Blick auf ihn größer geworden war. Aus diesem Grunde dürfen wir von einem »teleskopischen Gefälle« sprechen.

Existential risk researchers could justifiably object that one needs to distinguish more carefully between their long-term concerns and their short- to medium-term concerns. Whilst above characterisation of its implications might apply to their speculations about the long-term consequences and potentialities of technological development, in particular regarding AI, in the short-term, their agenda and their reference unit is abundantly clear, common-sensical and straight-forward: Their reference unit of concern is humanity and an existential risk is one that threatens humanity's continued existence. However, from the perspective of Arendt, Anders and Heidegger, this kind of argument would have to be considered a red herring.

In a way it suggests that the deep, ontologically rooted, puzzles and tensions science and technology confront us with should become a matter of interest only once the human being itself becomes an object of actual technological decision making, i.e. once we are practically confronted with the choice whether humanity should wilfully adopt some form of post-human existence, whether we should create new artificial people or not and, if so, in 'whose image'. From a Heideggerian perspective, this is a dangerous misunderstanding because it suggests that until we have reached this unspecifiable moment in time, modern technology could be dealt with under abstraction of the ontological puzzles it confronts us with and as if it were a mere means to an end at the free disposal of humanity. What Anders, Arendt and Heidegger demonstrate is that this approach not only cannot work but is itself expressive of a mistaken perspective on modern technology. It conceives of technology as being concerned with 'what is to be', not noticing that the mindset underlying modern technology undermines the conditions under which we can make sense of that very question.

Existential risk research so far has not systematically addressed this puzzle at its core. However, as part of this thesis' aim to bring out what facets of this emerging genre of research might be new or distinctive, I hope to have demonstrated that it indeed presents this puzzle in a new light and therewith revivifies central aspects of Heidegger's, Arendt's and Anders' thinking about technology.

What existential risk research shows is that the idea of human value and technology necessarily throw each other into question - that one cannot simultaneously take the first as a given and posit the neutrality of the second. If humanity has a value, then technology is not a neutral instrument and if technology is a neutral instrument it throws human value into a question. Thus, either, one takes some notion of human value as a given and thus makes it one's benchmark for thinking about present and future, in which case the all-transcending logic of modern technology cannot be regarded as something neutral, or we take technology and the unlimited space of possibilities it represents as a benchmark, in which case one cannot bracket the problem of human value. On the contrary, the very idea of human value is thrown into question because making the space of technological possibilities our benchmark for thinking about the future renders it impossible for us to ascertain what our reference units of value should be. In other words, by attempting to do both, taking the idea of human value as a given, whilst simultaneously thinking of technology as a neutral collection of instruments, existential risk research ends up in a paradoxical situation where it attempts to secure something, humanity, by drawing on a logic which cannot make sense of this term to begin with, nor, in fact, of any *thing* at all.

Bibliography

1. Primary literature on Anders, Arendt and Heidegger

Günther Anders:

Anders, G. (1948). On the Pseudo-Concreteness of Heidegger's Philosophy. *Philosophy and Phenomenological Research 8*(3): 337–371.

Anders, G. (1956a). *Die Antiquiertheit des Menschen*. München: C.H. Beck.

Anders, G. (1956b). Reflections on the H-Bomb. *Dissent 3*(2): 146–155.

Anders, G. (1962). Theses for the Atomic Age. *The Massachusetts Review 3*(3): 493–505.

Anders, G. (1965). Being without time. In Esslin, M. (ed.), *Samuel Beckett. A collection of critical essays.* Englewood Cliffs: Prentice-Hall.

Anders, G. (1969). Der Blick vom Mond. *Merkur 23*(9): 817–835.

Anders, G. (1972). *Endzeit und Zeitenende: Gedanken über die atomare Situation*. München: C.H. Beck.

Anders, G. (1982a). *Hiroshima ist überall*. München: C.H. Beck.

Anders, G. (1982b). *Ketzereien*. München: C.H. Beck.

Anders, G. (1992). *Die Antiquiertheit des Menschen 2 - Über die Zerstörung des Lebens im Zeitalter der dritten industriellen Revolution* (4th edition). München: C.H. Beck.

Anders, G. (2001). *Über Heidegger*. (G. Oberschlick & W. Reimann, eds.). München: C.H. Beck.

Anders, G. (2006). *Tagesnotizen: Aufzeichnungen 1941 - 1979*. Frankfurt am Main: Suhrkamp.

Anders, G. (2016). Promethean Shame. In Müller, C.J. (ed., trans.) *Prometheanism*. London, New York: Rowmann & Littlefield.

Anders, G. (2018). *Die Weltfremdheit des Menschen: Schriften zur philosophischen Anthropologie*. (C. Dries & H. Gätjens, eds.) (1. Auflage.). München: C.H. Beck.

Anders, G. & Arendt, H. (2016). *Schreib doch mal 'hard facts' über Dich: Briefe 1939 bis 1975, Texte und Dokumente*. (K. Putz, ed.). München: C.H. Beck.

Anders, G. & Eatherly, C. (1961). *Burning Conscience*. London: Weidenfeld and Nicholson.

Hannah Arendt:

Arendt, H. (1958). The Modern Concept of History. *The Review of Politics 20*(4): 570–590.

Arendt, H. (1961). *Between Past and Future*. New York: Viking Press.

Arendt, H. (1963). Man's Conquest of Space. *The American Scholar 32*(4): 527–540.

Arendt, H. (1964). "Cybernetics". Lecture. *Hannah Arendt Papers*. Manuscript Division. Library of Congress, Washington, D.C.

Arendt, H. (1970). *On Violence*. New York: Harcourt, Brace & World.

Arendt, H. (1973). *The Origins of Totalitarianism* (New ed.). New York: Harcourt Brace Jovanovich.

Arendt, H. (1990). *On Revolution* (Reprinted.). London: Penguin Books.

Arendt, H. (1994). *Essays in Understanding, 1930-1954: Formation, Exile, and Totalitarianism*. (J. Kohn, ed.). New York: Schocken Books.

Arendt, H. (1998). *The Human Condition* (2nd ed.). Chicago: University of Chicago Press.

Arendt, H. (2000). *The Portable Hannah Arendt*. (P.R. Baehr, ed.). New York: Penguin Books.

Arendt, H. (2005). *The Promise of Politics*. (J. Kohn, ed.) (1st ed.). New York: Schocken Books.

Arendt, H. (2007). The Conquest of Space and the Stature of Man. *The New Atlantis* (18): 43–55.

Arendt, H. & Jaspers, K. (1987). *Briefwechsel 1926 - 1969* (2. Aufl.). München: Piper.

Martin Heidegger:

Heidegger, M. (1933). Die Selbstbehauptung der deutschen Universität. In: Heidegger, M. (2000a). *Martin Heidegger Gesamtausgabe (HGA)* Vol. 16. (Hermann Heidegger, ed.). Frankfurt am Main, Germany: Vittorio Klostermann.

Heidegger, M. (1949). *Über den Humanismus*. Frankfurt am Main, Germany: Vittorio Klostermann.

Heidegger, M. (1955). *Der Satz vom Grund*. (P. Jaeger, ed.). Frankfurt am Main, Germany: Vittorio Klostermann.

Heidegger, M. (1966). *Discourse on Thinking*. (Anderson, J. & Freund, H., trans.). New York, USA: Harper & Row.

Heidegger, M. (1977). *The Question Concerning Technology and other Essays*. (W. Lovitt, trans.). New York, USA: Harper and Row.

Heidegger, M. (1981). 'Only a God Can Save Us': The Spiegel Interview (1966). (W. Richardson, trans.), In Sheehan*,* T. (ed.). (1981). *Heidegger: The Man and the Thinker*: 45-67.

Heidegger, M. (1996). *Being and time: a translation of Sein und Zeit.* (J. Stambaugh, trans.). Albany, USA: State University of New York Press.

Heidegger, M. (2000a). *Martin Heidegger Gesamtausgabe.* (Hermann Heidegger, ed.). Frankfurt am Main, Germany: Vittorio Klostermann.

Heidegger, M. (2000b). *Introduction to Metaphysics*. (Fried, G. and Polt, R. (trans.)). New Haven, USA & London, UK: Yale Nota Bene, Yale University Press.

Heidegger, M. (2002). *Off the Beaten Track*. J. Young & K. Haynes (eds.). Cambridge, UK; New York, USA: Cambridge University Press.

Heidegger, M. (2009). *Poetry, language, thought* (20. print.). New York, USA: Perennical Classics.

Heidegger, M. (2012). *Contributions to philosophy (of the event)*. (R. Rojcewicz & D. Vallega-Neu, eds.). Bloomington, USA: Indiana University Press.

Heidegger, M. (1993). *Martin Heidegger: Basic Writings: from Being and Time (1927) to the Task of Thinking (1964)*. (D.F. Krell, ed.). San Francisco, USA: Harper

2. Primary literature on existential risk

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. & Mané, D. (2016). Concrete Problems in AI Safety. *arXiv:1606.06565 [cs]*. Retrieved from http://arxiv.org/abs/1606.06565

Annual Report of the Government Chief Scientific Adviser (2014). *Innovation: Managing Risk, Not Avoiding It. Evidence and Case Studies.* Government Office for Science. Retrieved from https://www.gov.uk/government/publications/innovation-managing-risk-not-avoiding-it

Armstrong, S. (2013). General purpose intelligence: Arguing the orthogonality thesis. *Analysis and Metaphysics* 12: 68–84.

Armstrong, S., & Sandberg, A. (2013). Eternity in six hours: Intergalactic spreading of intelligent life and sharpening the Fermi paradox. *Acta Astronautica* 89: 1-13.

Armstrong, S. & Sotala, K. (2015). How We're Predicting AI – or Failing to. In Romportl, J., Zackova, E. & Kelemen, J. (eds.) (2015), *Beyond Artificial Intelligence: The Disappearing Human-Machine Divide*. New York, USA: Springer International: 11-29

Armstrong, S. (2017). Introduction to the Technological Singularity. In Callaghan, V. et al. (eds): *The Technological Singularity: Managing the Journey*. New York, USA: Springer International: 1-10.

Armstrong, S., Sotala, K. & Ó hÉigeartaigh, S. (2014). The errors, insights and lessons of famous AI predictions–and what they mean for the future. *Journal of Experimental & Theoretical Artificial Intelligence 26*(3): 317–342.

Avin, S., Wintle, B.C., Weitzdörfer, J., Ó hÉigeartaigh, S.S., Sutherland, W.J. & Rees, M.J. (2018). Classifying global catastrophic risks. *Futures 102*: 20-26.

Baum, S. (2017). A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy. *Global Catastrophic Risk Institute Working Paper 17-1.* Global Catastrophic Risk Institute. Available at https://papers.ssrn.com/abstract=3070741

Baum, S. & Barrett, A. (2017). Global Catastrophes: The Most Extreme Risks. In Bier, V. (ed.) (2017). *Risk in Extreme Environments: Preparing, Avoiding, Mitigating, and Managing.* New York, USA: Routledge: 174-184.

Baum, S., Farquhar, S., Rhodes, C., Reghezza, M., Van Danzig, A., Mehra, M., et al. (2016). 'Resetting the frame'. *Global Challenges Quarterly Risk Report August 2016*. Global Challenges Foundation. Retrieved from https://www.cser.ac.uk/media/uploads/files/Global-Challenges-Quarterly-Risk-Report-August-2016.pdf

Baum, S. (2010). Is Humanity Doomed? Insights from Astrobiology. *Sustainability 2*(2): 591–603.

Baum, S. (2015). The far future argument for confronting catastrophic threats to humanity: Practical significance and alternatives. *Futures* 72: 86–96.

Beckstead, N. (2015). Differential technological development: some early thinking. *The GiveWell Blog*. Retrieved from http://blog.givewell.org/2015/09/30/differential-technological-development-some-early-thinking/

Beckstead, N., Bostrom, N., Cotton-Barratt, O., Bowerman, N., MacAskill, W., Ó hÉigeartaigh, S. & Ord, T. (2014). Unprecedented Technological Risks. *Joint report by the Future of Humanity Institute, CSER, and the Global Priorities Project*. Retrieved from https://www.fhi.ox.ac.uk/publications/beckstead-n-bostrom-n-bowerman-n-cotton-barratt-o-macaskill-w-o-heigeartaigh-s-ord-t-2014-unprecedented-technological-risks-future-of-humanity-institute/

Beckstead, N. & Ord, T. (2014). Managing Existential Risk from Emerging Technologies. In Annual Report of the Government Chief Scientific Adviser (2014). *Innovation: Managing Risk, Not Avoiding It. Evidence and Case Studies.* Government Office for Science*:* 115-120.

Bensinger, R. (2013). Building Phenomenological Bridges. *Less Wrong*. Retrieved 17 September 2018, from http://lesswrong.com/lw/jd9/building_phenomenological_bridges/.

Bensinger, R. (2015). Davis on AI capability and motivation. *Machine Intelligence Research Institute*. Retrieved 8 August 2018, from https://intelligence.org/2015/02/06/davis-ai-capability-motivation/

BERI (2018). Home. *Berkeley Existential Risk Initiative.* Retrieved 6 November 2018, from https://existence.org

Bostrom, N. (2002). Existential risks and related hazards. *Journal of Evolution and Technology* 9(1): 1–31.

Bostrom, N. (2003). Are We Living in a Computer Simulation? *The Philosophical Quarterly (1950-) 53*(211): 243–255.

Bostrom, N. (2004). The Future of Human Evolution. In Tandy, C. (ed.) (2004). *Death and Anti-Death: Two Hundred Years After Kant, Fifty Years After Turing*. Palo Alto, USA.: Ria University Press: 339-371.

Bostrom, N. (2006a). Dinosaurs, Dodos, Humans? *Global Agenda*. World Economic Forum, January 2006: 230-231. Retrieved from https://nickbostrom.com/papers/globalagenda.pdf

Bostrom, N. (2006b). What is a Singleton? *Linguistic and Philosophical Investigations 5*(2): 48–54.

Bostrom, N. (2008, April 22). Where Are They? *MIT Technology Review*. Retrieved 26 July 2018, from https://www.technologyreview.com/s/409936/where-are-they/

Bostrom, N. (2009). The Future of Humanity. In Olsen, J., Selinger, E. & Riis, S. (eds.) (2009), *New Waves in Philosophy of Technology*: 186-215.

Bostrom, N. (2012). The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds and Machines 22*(2): 71–85.

Bostrom, N. (2013). Existential Risk Prevention as Global Priority. *Global Policy 4*(1): 15–31.

Bostrom, N. (2014). *Superintelligence: paths, dangers, strategies*. Oxford, UK: Oxford University Press.

Bostrom, N. & Circovik, M. (eds.) (2008). *Global atastrophic risks*. Oxford, UK: Oxford University Press.

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In Frankish, K. & Ramsey, Y. (eds.) (2014). *The Cambridge Handbook of Artificial Intelligence* Cambridge, UK: Cambridge University Press: 316-334.

Bostrom, N., Dafoe, A. & Carrick, F. (2016). Policy Desiderata in the Development of Machine Superintelligence. Working Paper. Future of Humanity Institute. Retrieved from https://www.fhi.ox.ac.uk/wp-content/uploads/Policy-Desiderata-in-the-Development-of-Machine-Superintelligence.pdf

Brundage, M. (2015). Taking superintelligence seriously: Superintelligence: Paths, Dangers, Strategies by Nick Bostrom. *Futures 72*: 32–35.

CHAI. (2018). About. *Center for Human-Compatible AI*. Retrieved 15 November 2018, from https://humancompatible.ai/about

Chalmers, D. (2010). The singularity: A philosophical analysis. *Journal of Consciousness Studies* 17(9–10): 7–65.

Cotton-Barratt, O. & Ord, T. (2015). Existential Risk and Existential Hope: Definitions. Technical report No. #2015-1. Future of Humanity Institute, University of Oxford. Retrieved from https://www.fhi.ox.ac.uk/Existential-risk-and-existential-hope.pdf

Cotton-Barratt, O., Farquhar, S., Halstead, J. & Schubert, S. (2016). *GLOBAL CATASTROPHIC RISKS 2016*. A joint report by the Global Challenges Foundation and Global Priorities Institute. Retrieved from globalprioritiesproject.org/2016/04/global-catastrophic-risks-2016/

CSER (2016). Cambridge Conference on Catastrophic Risk 2016. Retrieved 19 October 2018, from https://www.cser.ac.uk/events/CCCR-2016/

CSER (2018a). About. *Centre for the Study of Existential Risk.* Retrieved 20 October 2016, from https://www.cser.ac.uk/about-us/

CSER (2018b). Risks from Artificial Intelligence. *Centre for the Study of Existential Risk.* Retrieved 14 November 2018, from https://www.cser.ac.uk/research/risks-from-artificial-intelligence/

CSER (2018c). Team. *Centre for the Study of Existential Risk.* Retrieved 14 November 2018, from https://www.cser.ac.uk/team/

Currie, A. (2018a). Existential Risk, Creativity & Well-Adapted Science. *Studies in the History & Philosophy of Science* (forthcoming). Retrieved from http://philsci-archive.pitt.edu/14800/

Currie, A. (2018b). Geoengineering tensions. *Futures 102*: 78–88.

Dafoe, A. & Russell, S. (2016, October 19). Analysis of Müller and Bostrom 2016. *Harvard Dataverse*. Retrieved from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/JHR1GX

Dasgupta, P., Ramanathan, V., Raven, P., Rees, M., et al. (2015). *Climate Change and The Common Good: A Statement of The Problem and The Demand For Transformative Solutions*. The

Pontifical Academy of Sciences and the Pontifical Academy of Social Sciences. Retrieved from https://www.cser.ac.uk/media/uploads/files/climate-change-and-the-common-good.pdf

Dasgupta, P. (2017). Birth and Death. *Apollo - University of Cambridge Repository.* Retrieved from https://www.hbs.edu/faculty/conferences/2016-newe/Documents/Dasgupta_POPULATION-October%202016.pdf

De Blanc, P. (2011). Ontological Crises in Artificial Agents' Value Systems. The Singularity Institute, San Francisco, CA, May 19. http://arxiv.org/abs/1105.3821.

DeGrey, A. et al. (2017). *The Next Step: Exponential Life.* Madrid: Turner.

Denkenberger, D. and Pearce, J. (2015). *Feeding Everyone No Matter What*. London, UK: Academic Press.

Dewey, D. (2011). Learning What to Value. In Schmidhuber, J. Thórisson, K. & Looks, M. (eds.) (2011), *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3-6, 2011, Proceedings*. Berlin, Germany: Springer: 309-314.

Dewey, D. (2016). Long-term strategies for ending existential risk from fast takeoff. In Müller, V. (ed.) (2016). *Risks of Artificial Intelligence.* New York, USA: Taylor & Francis: ch.11.

Diamandis, P.H. & Kotler, S. (2012). *Abundance: the future is better than you think* (1st Free Press hardcover ed.). New York, USA: Free Press.

Drexler, E. (1986). *Engines of Creation*. Garden City, USA: Anchor Press/Doubleday.

Drexler, E. (2006). *Engines of Creation 2.0: The Coming Era of Nanotechnology*. Los Angeles: WOWIO.

Drexler, K.E. (2013). *Radical abundance: how a revolution in nanotechnology will change civilization*. New York, USA: BBS PublicAffairs.

Duettmann, A., Peterson, C. & Miller, M.S. (2017). Cyber, Nano, and AGI Risks: Decentralized Approaches to Reducing Risks. In Garrick, B.J. (ed.) (2017). *Proceedings of The First Colloquium On Catastrophic And Existential Risk.* B. John Garrick Institute or the Risk Sciences. University of California Los Angeles: 144-183.

Farquhar, S., Haltstead, J., Cotten-Barratt, O., Schubert, S., Belfield, H. & Snyder-Beattie, A. (2017). *Existential Risk: Diplomacy and Governance*. Global Priorities Project 2017, Global

Priorities Institute, Ministry of Foreign Affairs Finland. Retrieved from http://globalprioritiesproject.org/2017/02/existential-risk-diplomacy-and-governance/

FHI (2018a). Future of Humanity Institute. *Oxford Martin School.* Retrieved 12 October 2017 from https://www.oxfordmartin.ox.ac.uk/research/programmes/future-humanity

FHI (2018b). Team. *Future of Humanity Institute*. Retrieved 12 October 2017 from https://www.fhi.ox.ac.uk/the-team/

FHI (2018c). Research Areas. *Future of Humanity Institute*. Retrieved 12 October 2017 from https://www.fhi.ox.ac.uk/research/research-areas/

FLI (2018a). AI Open Letter. *Future of Life Institute*. Retrieved 3 August 2018 from https://futureoflife.org/ai-open-letter/

FLI (2018b). AI Safety Myths. *Future of Life Institute*. Retrieved 14 November 2018 from https://futureoflife.org/background/aimyths/

FLI (2018c). AI Principles. *Future of Life Institute*. Retrieved 6 August 2018 from https://futureoflife.org/ai-principles/

FLI (2018d). Team. *Future of Life Institute*. Retrieved 6 August 2018 from https://futureoflife.org/team/

FRI (2018). Home. *Foundational Research Institute*. Retrieved 15 November 2018 from https://foundational-research.org/

GCRI (2018). About. *Global Catastrophic Risk Institute*. Retrieved 15 October 2017 from http://gcrinstitute.org/about/

Goulding, I. (2014). Future Global Trends in Innovation. In Annual Report of the Government Chief Scientific Adviser (2014). *Innovation: Managing Risk, Not Avoiding It. Evidence and Case Studies.* Government Office for Science: 25-34.

GPI (2018). About us. *Global Priorities Institute*. Retrieved 20 November 2018, from https://globalprioritiesinstitute.org/about-us/

GPP (2018). Home. *Global Priorities Project*. Retrieved 20 November 2018, from http://globalprioritiesproject.org

Grace, K., Salvatier, J., Dafoe, A., Zhang, B. & Evans, O. (2018). When Will AI Exceed Human Performance? Evidence from AI Experts. *arXiv:1705.08807 [cs]*. Version 3. Retrieved from http://arxiv.org/abs/1705.08807

Green, B.P. (2014). Are science, technology, and engineering now the most important subjects for ethics? Our need to respond. Presented at the 2014 IEEE International Symposium on Ethics in Science, Technology and Engineering, ETHICS 2014.

Green, B.P. (2016). "Emerging technologies, catastrophic risks, and ethics: three strategies for reducing risk". IEEE International Symposium on Ethics in Engineering, Science and Technology. ETHICS 2016 Symposium Record. IEEE Xplore, 13-14 May 2016, Vancouver, British Columbia, Canada.

Hadfield-Menell, D., Russell, S.J., Abbeel, P. & Dragan, A. (2016). Cooperative Inverse Reinforcement Learning. In Lee, D. et al. (eds.) (2016). *Advances in Neural Information Processing Systems 29*: 30th Annual Conference on Neural Information Processing Systems 2016 (1). Red Hook, USA: Curran Associates: 3916-3925. Retrieved from https://arxiv.org/abs/1606.03137

Häggström, O. (2016). *Here be dragons: science, technology and the future of humanity* (First edition.). Oxford, UK; New York, USA: Oxford University Press.

Haqq-Misra, J. (2016). Here be dragons: science, technology and the future of humanity. Book Review. *Law, Innovation and Technology 8*(2): 268–270.

Hassabis, D. (2016). AlphaGo: using machine learning to master the ancient game of Go. *Official Google Blog*. Retrieved 21 September 2018 from https://googleblog.blogspot.com/2016/01/alphago-machine-learning-game-go.html

Hassabis, D. (2017, April 27). Artificial Intelligence: Chess match of the century. *Nature* (544): 413-414.

Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron 95*(2): 245–258.

Hawking, S. (2015). Hawking Reddit AMA on AI. *Future of Life Institute*. Retrieved 12 October 2018 from https://futureoflife.org/2015/10/11/hawking-reddit-ama-on-ai/

Hawking, S. (2016, December 1). This is the most dangerous time for our planet. *The Guardian*. Retrieved 12 October 2018 from https://www.theguardian.com/commentisfree/2016/dec/01/stephen-hawking-dangerous-time-planet-inequality

Hawking, S., Tegmark, M., Russell, S. & Wilczek, F. (2014, May 1). Stephen Hawking: 'Transcendence looks at the mplications of artificial intelligence – but are we taking AI seriously enough?'. *The Independent*. Retrieved 12 October 2018 from http://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-9313474.html

Hotz, J. (2018, January 29). To Save Generations Of Tomorrow, We Need To Change Political Decision-making Today. *HuffPost UK*. Retrieved 31 January 2018, from http://www.huffingtonpost.co.uk/entry/to-save-generations-of-tomorrow-we-need-to-change_uk_5a6f4347e4b0290826014b1c

Humanity+ (2018). About. *Humanity+*. Retrieved 11 March 2018, from https://humanityplus.org

IEET (2018). Institute for Ethics and Emerging Technologies. Retrieved 15 November 2018, from https://www.ieet.org/

Jebari, K. (2014). Of Malthus and Methuselah: does longevity treatment aggravate global catastrophic risks? *Physica Scripta 89*(12): 128005.

Jebari, K. (2015). Existential Risks: Exploring a Robust Risk Reduction Strategy. *Science and Engineering Ethics 21*(3): 541–554.

Kareiva, P. & Carranza, V. (2018). Existential risk due to ecosystem collapse: Nature strikes back. *Futures 102*: 39–50.

Karnofsky, H. (2016). Potential Risks from Advanced Artificial Intelligence: The Philanthropic Opportunity. *Open Philanthropy Project*. Retrieved from https://www.openphilanthropy.org/blog/potential-risks-advanced-artificial-intelligence-philanthropic-opportunity

Khan, R. (2018, January 10). 5 reasons why we need to start talking about existential risks. *World Economic Forum*. Retrieved from https://www.weforum.org/agenda/2018/01/5-reasons-start-talking-existential-risks-extinction-moriori/

Kurzweil, R. (2005). *The singularity is near: when humans transcend biology*. New York, USA: Viking.

Kurzweil, R. (2013). Progress and Relinquishment. In More, M. & Vita-More, N. (eds.) (2013). *The Transhumanist Reader: Classical and Contemporary Essays on the Science, Technology, and Philosophy of the Human Future*. New York, USA: Wiley-Blackwell: 451-453.

LCFI. (2018a). The Value Alignment Problem. *Leverhulme Centre for the Future of Intelligence*. Retrieved 6 November 2018, from http://lcfi.ac.uk/projects/ai-futures-and-responsibility/value-alignment-problem/

LCFI. (2018b). Preparing for the age of intelligent machines. *Leverhulme Centre for the Future of Intelligence*. Retrieved 15 November 2018, from http://lcfi.ac.uk/about/

LCFI (2018c). Murray Shanahan. *Leverhulme Centre for the Future of Intelligence*. Retrieved 15 November 2018, from http://lcfi.ac.uk/team/murray-shanahan/

Leslie, J. (1996). *The End of the World - The Science and Ethics of Human Extinction*. London, UK: Routledge.

Lin, F. (2017). The First Colloquium on Catastrophic and Existential Risk. *The B. John Garrick Institute for the Risk Sciences*. Retrieved 20 October 2018, from https://www.risksciences.ucla.edu/news-events/2017/1/31/the-first-colloquium-on-catastrophic-and-existential-threats

Mallah, R. (2017). The Landscape of AI Safety and Beneficence Research: Input for Brainstorming at Beneficial AI 2017. *Future of Life Institute.* Retrieved 17 September 2018, from https://futureoflife.org/landscape/ResearchLandscapeExtended.pdf

Martinez-Plumed, F., Bao Sheng, L., Ó hÉigeartaigh, S., Vold, K. & Orallo, H. (2018). The Facets of Artificial Intelligence: A Framework to Track the Evolution of AI. In Lang, J. (ed.). *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. International Joint Conference on Artificial Intelligence: 5180-5187.

Matheny, J.G. (2007). Reducing the Risk of Human Extinction. *Risk Analysis 27*(5): 1335–1344.

McEuen, P. & Dekker, C. (2008). Synthesizing the Future. *ACS Chemical Biology 3*(1): 10–12.

Mhamdi, E.M.E., Guerraoui, R., Hendrikx, H. & Maurer, A. (2017). Dynamic Safe Interruptibility for Decentralized Multi-Agent Reinforcement Learning. *arXiv:1704.02882 [cs, stat]*. Retrieved from http://arxiv.org/abs/1704.02882

MIRI (2018a). About the Machine Intelligence Research Institute. *Machine Intelligence Research Institute*. Retrieved from https://intelligence.org/about/

MIRI (2018b). Mission. *Machine Intelligence Research Institute*. Retrieved from http://www.rationality.org/about/mission/

Muehlhauser, L. & Salamon, A. (2012). Intelligence Explosion: Evidence and Import. In Eden, A., Moor, J., Søraker, J. & Steinhart, E. (eds.) (2012). *Singularity Hypotheses*. Berlin, Germany: Springer: 15-42.

Muehlhaeuser, L. (2014). Two mistakes about the threat from artificial intelligence. *World Economic Forum*. Retrieved from https://www.weforum.org/agenda/2014/12/two-mistakes-about-the-threat-from-artificial-intelligence/

Müller, V.C. (2014). Risks of general artificial intelligence. *Journal of Experimental & Theoretical Artificial Intelligence 26*(3): 297–301.

Ó hÉigeartaigh, S. (2017). Technological Wild Cards: Existential Risk and a Changing Humanity. In DeGrey, A. et al. (2017). *The Next Step: Exponential Life*. Madrid: Turner. 344-371.

Ó hÉigeartaigh, S., et al. (2018, February 2). Preparing for the future: artificial intelligence and us. *Research*. University of Cambridge blog. Retrieved 25 July 2018, from https://www.cam.ac.uk/research/discussion/preparing-for-the-future-artificial-intelligence-and-us

Omohundro, S. (2012). Rational Artificial Intelligence for the Greater Good. In Eden, A., Moor, J., Søraker, J. & Steinhart, E. (eds.) (2012), *Singularity Hypotheses*. Berlin, Germany: Springer: 161-180.

OpenAI. (2015). Introducing OpenAI. *OpenAI Blog*. Retrieved 25 August 2018, from https://blog.openai.com/introducing-openai/

OpenAI. (2017). Learning from Human Preferences. *OpenAI Blog*. Retrieved 23 August 2018, from https://blog.openai.com/deep-reinforcement-learning-from-human-preferences/

OpenAI. (2018a). Learning Dexterity. *OpenAI Blog*. Retrieved 20 August 2018, from https://blog.openai.com/learning-dexterity/

OpenAI. (2018b). Learning Dexterous In-Hand Manipulation. *arXiv:1808.00177 [cs, stat]*. Retrieved from http://arxiv.org/abs/1808.00177

OpenAI. (2018c). OpenAI Charter. *OpenAI Blog*. Retrieved 20 August 2018, from https://blog.openai.com/openai-charter/

OpenAI. (2018d). Preparing for Malicious Uses of AI. *OpenAI Blog*. Retrieved 24 August 2018, from https://blog.openai.com/preparing-for-malicious-uses-of-ai/

Ord, T., Hillerbrand, R. & Sandberg, A. (2010). Probing the improbable: methodological challenges for risks with low probabilities and high stakes. *Journal of Risk Research* 13(2): 191-205.

Orseau, L. & Ring, M. (2012). Space-Time Embedded Intelligence. In Bach, J. Goertzel, B. & Ikle, M. (eds.) (2012). *Artificial General Intelligence: 5th International Conference, AGI 2012, Oxford, UK, December 2012: Conference Proceedings*. New York, USA: Springer: 209-218.

Pamlin, D. & Armstrong, S. (2015). *Global Challenges: 12 Risks that Challenge Human Civilization: The Case for a new Category of Risk*. Global Challenges Foundation. Retrieved from https://api.globalchallenges.org/static/wp-content/uploads/12-Risks-with-infinite-impact.pdf

Persson, I. & Savulescu, J. (2008). The Perils of Cognitive Enhancement and the Urgent Imperative to Enhance the Moral Character of Humanity. *Journal of Applied Philosophy 25*(3): 162–177.

Phoenix, C. & Drexler, E. (2004). Safe exponential manufacturing. *Nanotechnology 15*(8): 869.

Price, H. (2013, January 27). Cambridge, Cabs and Copenhagen: My Route to Existential Risk. *The New York Times*. Retrieved from https://opinionator.blogs.nytimes.com/2013/01/27/cambridge-cabs-and-copenhagen-my-route-to-existential-risk/

Price, H. & Ó hÉigeartaigh, S. (2014). CASE STUDY: POLICY, DECISION-MAKING AND EXISTENTIAL RISK. In *Annual Report of the Government Chief Scientific Adviser 2014. Innovation: Managing Risk, Not Avoiding It. Evidence and Case Studies*: 117

Price, H. & Tallinn, J. (2012). Artificial intelligence – can we keep it in the box? *The Conversation*. Retrieved 25 August 2018, from http://theconversation.com/artificial-intelligence-can-we-keep-it-in-the-box-8541

Price, H. & Vold, K. (2018). Living With AI. *Research Horizons* (35), February 2018 Issue.

Randle, M. & Eckersley, R. (2015). Public perceptions of future threats to humanity and different societal responses: A cross-national study. *Futures 72*: 4–16.

Rees, M. (2004). *Our final hour: a scientist's warning: how terror, error, and environmental disaster threaten humankind's future in this century on earth and beyond*. New York, USA: Basic Books.

Rees, M. (2008). Foreword. In Bostrom, N. & Circovik, M. (eds.) (2008). *Global catastrophic risks*. Oxford, UK: Oxford University Press: vii-xi.

Rees, M. (2013). Denial of Catastrophic Risks. *Science 339*(6124): 1123–1123.

Rees, M. (2014, November 26). Martin Rees: The world in 2050 and beyond. *New Statesman*. Retrieved from https://www.newstatesman.com/sci-tech/2014/11/martin-rees-world-2050-and-beyond

Rees, M. (2017a). Foreword. In Torres, P. (2017b). *Morality, Foresight, and Human Flourishing: An Introduction to Existential Risks*. Durham, USA: Pitchstone Publishing.

Rees, M. (2017b). Interstellar Travel and Post-Humans. In DeGrey, A. et al. (2017). *The Next Step: Exponential Life.* Madrid: Turner: 372-496.

Rees, M. (2018, October 3). Martin Rees brings 'On the Future: Prospects for Humanity' to Harvard. *Harvard Gazette*. Retrieved from https://news.harvard.edu/gazette/story/2018/10/martin-rees-brings-on-the-future-prospects-for-humanity-to-harvard/

Russell, S. (2017). Provably Beneficial Artificial Intelligence. In DeGrey, A. et al. (2017). *The Next Step: Exponential Life.* Madrid: Turner: 175-192.

Russell, S. (2018). Q&A: The future of artificial intelligence. Retrieved 14 November 2018, from https://people.eecs.berkeley.edu/~russell/research/future/q-and-a.html

Russell, S. & Dafoe, A. (2017). Yes, the experts are worried about the existential risk of artificial intelligence. *MIT Technology Review*. Retrieved 13 August 2018, from https://www.technologyreview.com/s/602776/yes-we-are-worried-about-the-existential-risk-of-artificial-intelligence/

Russell, S., Dewey, D. & Tegmark, M., et al. (2015). AI Principles. *AI Magazine 4*(36).

Russell, S. & Norvig, P. (2016). *Artificial intelligence: a modern approach* (Third edition). Boston, USA: Pearson.

Sandberg, A. & Bostrom, N. (2008): Global Catastrophic Risks Survey, *Technical Report* #2008-1, Future of Humanity Institute, Oxford University: 1-5.

Sandberg, A. (2012, November 2). Personalised weapons of mass destruction: governments and strategic emerging technologies. *Practical Ethics blog.* University of Oxford. Retrieved 15 May 2018, from http://blog.practicalethics.ox.ac.uk/2012/11/personalised-weapons-of-mass-destruction-governments-and-strategic-emerging-technologies/

Sandberg, A., Matheny, J. & Ćirković, M. (2008, September 9). How can we reduce the risk of human extinction? *Bulletin of the Atomic Scientists*. Retrieved 15 May 2018, from https://thebulletin.org/how-can-we-reduce-risk-human-extinction

Shanahan, M. (2015). *The technological singularity*. Cambridge, USA: The MIT Press.

Shulman, C. (2010). Omohundro's "Basic AI Drives" and Catastrophic Risks. San Francisco, CA, USA: The Singularity Institute. Retrieved from https://intelligence.org/files/BasicAIDrives.pdf

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A. Hassabis, D., et al. (2017). Mastering the game of Go without human knowledge. *Nature 550*(7676): 354–359.

Singer, P., Beckstead, N. & Wage, M. (2013). Preventing human extinction. *Effective Altruism*. Retrieved from http://effective-altruism.com/ea/50/preventing_human_extinction/.

Singer, P. (2015). *The most good you can do: how effective altruism is changing ideas about living ethically*. New Haven, USA: Yale University Press.

Singer, P. (2016). *Ethics in the real world: 86 brief essays on things that matter*. Princeton, USA: Princeton University Press.

Singer, P. (2016). Can artificial intelligence be ethical? *World Economic Forum*. Retrieved 9 May 2018, from https://www.weforum.org/agenda/2016/04/can-artificial-intelligence-be-ethical/.

Smil, V. (2008). *Global Catastrophes and Trends: The next 50 Years*. Cambridge, USA.: MIT Press.

Snyder-Beattie, A. (2015, May 3). Small groups, dangerous technology: Can they be controlled? *Bulletin of the Atomic Scientists*. Retrieved 13 April 2018, from https://thebulletin.org/small-groups-dangerous-technology-can-they-be-controlled8270.

Soares, N. (2015a). Research Guide. *Machine Intelligence Research Institute*. Retrieved from https://intelligence.org/research-guide/

Soares, N. (2015b). Formalizing Two Problems of Realistic World-Models. *Machine Intelligence Research Institute.* Technical report 2015 (3). Berkeley, CA, USA. Retrieved from https://intelligence.org/files/RealisticWorldModels.pdf.

Soares, N. & Fallenstein, B. (2014). Aligning Superintelligence with Human Interests: A Technical Research Agenda. *Machine Intelligence Research Institute.* Technical report 2014(8). Berkeley, CA, USA: Machine Intelligence Research Institute. https://intelligence.org/files/TechnicalAgenda.pdf.

Soares, N. & Fallenstein, B. (2017). Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda. In Callaghan, V. et al. (eds.) (2017): *The Technological Singularity: Managing the Journey*. New York, USA: Springer International: 103-126.

Sotala, K. & Yampolinsky, R. (2017). Risks of the Journey to the Singularity. In Callaghan, V. et al. (eds.) (2017): *The Technological Singularity: Managing the Journey.* New York, USA: Springer International: 11-24.

Sotala, K. & Yampolskiy, R.V. (2015). Responses to Catastrophic AGI risk: A Survey. *Physica Scripta 90*(1).

Tegmark, M. (2015, October 12). Elon Musk donates $10M to keep AI beneficial. *Future of Life Institute*. Retrieved 6 August 2018, from https://futureoflife.org/2015/10/12/elon-musk-donates-10m-to-keep-ai-beneficial/

Tegmark, M. (2017). *Life 3.0: being human in the age of artificial intelligence* (First edition.). New York: Alfred A. Knopf.

Todd, B. (2017). Why despite global progress, humanity is probably facing its most dangerous time ever. *80,000 Hours*. Retrieved 12 January 2018 from https://80000hours.org/articles/extinction-risk/

Torres, P. (2016, June 29). Existential Risks Are More Likely to Kill You Than Terrorism. *Future of Life Institute*. Retrieved 17 May 2018, from https://futureoflife.org/2016/06/29/existential-risks-likely-kill-terrorism/

Torres, P. (2017a). Agential risks and information hazards: An unavoidable but dangerous topic? *Futures 95*: 86–97.

Torres, P. (2017b). *Morality, Foresight, and Human Flourishing: An Introduction to Existential Risks*. Durham, NC, USA: Pitchstone Publishing.

Torres, P. (2017c). Who Would Destroy the World? Omnicidal Agents and Related Phenomena. *Aggression and Violent Behavior* (Forthcoming).

Torres, P. (2017d, October 24). Why superintelligence is a threat that should be taken seriously. *Bulletin of the Atomic Scientists*. Retrieved from https://thebulletin.org/2017/10/why-superintelligence-is-a-threat-that-should-be-taken-seriously/

Turchin, A. & Denkenberger, D. (2018). Classification of global catastrophic risks connected with artificial intelligence. *AI and Society*, 2018: 1-17. Retrieved from https://doi.org/10.1007/s00146-018-0845-5

Worley III, G. (2018, February 19). Formally Stating the AI Alignment Problem. *Map and Territory*. Retrieved on November 12, 2018, from https://mapandterritory.org/formally-stating-the-ai-alignment-problem-fe7a6e3e5991

Yudkowsky, E. (2008). Artificial Intelligence as Positive and Negative Factor in Global Risk. In Bostrom, N. & Circovik, M. (eds.) (2008). *Global Catastrophic Risks*. Oxford, UK; New York, USA: Oxford University Press: 308-346.

Yudkowsky, E. (2011). Complex Value Systems in Friendly AI. In Schmidhuber, J. Thórisson, K. & Looks, M. (eds.) (2011), *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3-6, 2011, Proceedings*. Berlin, Germany: Springer: 388-393.

3. Secondary literature

Achterhuis, H. (ed.). (2001). *American Philosophy of Technology: The Empirical Turn*. Bloomington, USA: Indiana University Press.

Adam, B., Beck, U., Loos, J.v. (eds.) (2000). *The Risk Society and Beyond: Critical Issues for Social Theory*. London, UK: Sage.

Agar, J. (2003). *The Government Machine: A Revolutionary History of the Computer*. Cambridge, USA: MIT Press.

Albertson, B. & Gadarian, S. (2015). *Anxious Politics: Democratic citizenship in a threatening world.* Cambridge, UK: Cambridge University Press

Aleksievich, S. (2016). *Chernobyl prayer: a chronicle of the future.*: London, UK: Penguin.

Alfonseca, M., Cebrian, M., Anta, A.F., Coviello, L., Abeliuk, A. & Rahwan, I. (2016). Superintelligence cannot be contained: Lessons from Computability Theory. *arXiv:1607.00913 [cs]*. Retrieved from http://arxiv.org/abs/1607.00913

Amadae, S.M. (2016). *Prisoners of Reason: Game Theory and Neoliberal Political Economy*. New York, USA: Cambridge University Press.

Amadae, S.M. (2018). Computable Rationality, NUTS, and the Nuclear Leviathan. In D. Bessner & N. Guilhot (eds.), *The Decisionist Imagination: Sovereignty, Social Science, and Democracy in the 20th Century*. New York, USA: Berghahn Books.

Andersen, R. (2012, March 6). We're Underestimating the Risk of Human Extinction. *The Atlantic*. Retrieved from https://www.theatlantic.com/technology/archive/2012/03/were-underestimating-the-risk-of-human-extinction/253821/

Andersen, R. (2013, February 25). Will humans be around in a billion years? Or a trillion? *Aeon*. Retrieved from https://aeon.co/essays/will-humans-be-around-in-a-billion-years-or-a-trillion

Andersen, R. (2014, September 30). Exodus. *Aeon*. Retrieved from https://aeon.co/essays/elon-musk-puts-his-case-for-a-multi-planet-civilisation

Andersson, J. (2012). The Great Future Debate and the Struggle for the World. *The American Historical Review 117*(5): 1411–1430.

Andersson, J. (2018): *The Future of the World: Futurology, Futurists, and the Struggle for the Post Cold War Imagination.* Oxford, UK; New York, USA: Oxford University Press

Angell, I.O. (1993). Intelligence: logical or biological. *Communications of the ACM 36*(7): 15 ff.

Aradau, C. & Münster, R. v. (2011). *Politics of Catastrophe: Genealogies of the Unknown.* London, UK: Routledge.

Arkoudas, K. & Bringsjord, S. (2014). *Philosophical Foundations*. In Frankish, K. & Ramsey, W. (eds.) (2014), *The Cambridge Handbook of Artificial Intelligence.* Cambridge, UK: Cambridge University Press: 34-63.

Atkinson, R.D. (2015). The 2015 ITIF Luddite Award Nominees: The Worst of the Year's Worst Innovation Killers. *Information Technology and Innovation Foundation*. Retrieved from https://itif.org/publications/2015/12/21/2015-itif-luddite-award-nominees-worst-year%E2%80%99s-worst-innovation-killers

Audi, R. (ed.). (2009). *The Cambridge dictionary of philosophy* (2. ed., 11. printing.). Cambridge, UK: Cambridge Univ. Press.

Babich, B. (2013a). Angels, the Space of Time, and Apocalyptic Blindness: On Günther Anders' Endzeit–Endtime. *Etica & Politica / Ethics & Politics XV*(2): 144–174.

Babich, B. (2013b). O, Superman! Or being Towards Transhumanism: Martin Heidegger, Günther Anders, and Media Aesthetics. *Divinatio* (36): 41–99.

Bambach, C. (2003). Heidegger, Technology, and the Homeland. *The Germanic Review: Literature, Culture, Theory* 78(4): 267-282.

Barbrook, R. & Cameron, A. (1996). The Californian ideology. *Science as Culture* 6(1): 44–72.

Barnodsky, A. et al. (2011). Has the Earth's sixth mass extinction already arrived? *Nature* 471(7336): 51-57.

BBC (2018, May 8). Google AI to make phone calls for you. Retrieved on October 14, 2018 from https://www.bbc.com/news/technology-44045424.

Beck, U. (1992). *Risk Society: Towards a New Modernity.* London, UK: Sage.

Beck, U. (2007). *World at Risk*. Cambridge, UK; Malden, USA: Polity.

Benedict, K. (2018). FAQ. *Bulletin of the Atomic Scientists*. Retrieved from https://thebulletin.org/doomsday-clock/faq/

Benhabib, S. (1990). Critical theory and postmodernism: on the interplay of ethics, aesthetics, and utopia in critical theory. *Cardozo Law Review 11*: 1435 ff.

Benhabib, S. (2003). *The Reluctant Modernism of Hannah Arendt* (New ed.). Lanham, USA: Rowman & Littlefield.

Benhabib, S. (ed.). (2010). *Politics in dark times: encounters with Hannah Arendt.* Cambridge, UK; New York, USA: Cambridge University Press.

Benson, R. (2017, February 12). Meet Earth's Guardians, the real-world X-men and women saving us from existential threats. *Wired UK*. Retrieved from https://www.wired.co.uk/article/earth-guardians-existential-risk

Bernasconi, R. (2002). Hannah Arendt, Phenomenology and Political Theory. In Tymieniecka, A.-T. (ed.) (2002), *Phenomenology World-Wide*. Dordrecht, Netherlands: Springer Netherlands: 645-647.

Bernstein, R. (2006). Arendt on thinking. In Villa, D.R. (ed.). (2006). *The Cambridge Companion to Hannah Arendt*: 277-292.

Bernstein, R.J. (2013). *Hannah Arendt and the Jewish Question*. Hoboken, USA: Wiley.

Bertman, S. (2007). Chariots in the Sky. *The New Atlantis* (18): 71–75.

Binns, C. (2010). *Introduction to Nanoscience and Nanotechnology*. Hoboken, USA: Wiley.

Blitz, M. (2014). Understanding Heidegger on Technology. *The New Atlantis* (41): 63–80.

Boden, M. (1984). Impacts of artificial intelligence. *Futures 16*(1): 60–70.

Boden, M.A. (ed.). (1996). *Artificial Intelligence*. San Diego, USA: Academic Press.

Boden, M. (2006). *Mind as Machine: A History of Cognitive Science*. Oxford, UK: New York, USA: Clarendon Press: Oxford University Press.

Boden, M. (2016). *AI: Its Nature and Future*. Oxford, UK; New York, USA: Oxford University Press.

Bohannon, J. (2015). Fears of an AI pioneer. *Science 349*(6245): 252–252.

Böhme, G. (2012). *Invasive technification: critical essays in the philosophy of technology*. London, UK: Bloomsbury.

Bordoni, C. (2017). *State of fear in a liquid world*. London, UK ; New York, USA: Routledge.

Borgman, A. (2005). Technology. In: Dreyfus, H. & Wrathall, M. (eds.) (2005). *A Companion to Heidegger*: 420-432.

Borrie, J., Caughley, T. & Wan, W. (eds.) (2017). Understanding Nuclear Weapons Risk. *UNIDIR Research*, 2017. United Nations Institute for Disarmament Research. Retrieved from http://www.unidir.org/files/publications/pdfs/understanding-nuclear-weapon-risks-en-676.pdf

Botsman, R. (2017, October 21). Big data meets Big Brother as China moves to rate its citizens. *WIRED UK*. Retrieved 13 November 2017 from http://www.wired.co.uk/article/chinese-government-social-credit-score-privacy-invasion

Braidotti, R. (2013). *The Posthuman*. Cambridge, UK; Malden, USA: Polity.

Bringsjord, S. (2015). A Refutation of Searle on Bostrom (re: Malicious Machines) and Floridi (re: Information). *APA Newsletters* 15(1).

Bringsjord, S., Bringsjord, A. & Bello, P. (2012). Belief in The Singularity is Fideistic. In Eden, A., Moor, J., Søraker, J. & Steinhart, E. (eds.), *Singularity Hypotheses*: 395-412.

Broers, A. (2005). *The triumph of technology*. Cambridge, UK: Cambridge University Press.

Bronson, R. (2018). 2018 Doomsday Clock Statement. *Bulletin of the Atomic Scientists.* Retrieved from https://thebulletin.org/2018-doomsday-clock-statement/

Brown, A. (2015). The Silicon State. *Centre for Public Impact*. Retrieved 9 February 2016, from http://www.centreforpublicimpact.org/person/brownadrian/

Bryson, J.J. (2018). Patiency is not a virtue: the design of intelligent systems and systems of ethics. *Ethics and Information Technology 20*(1): 15–26.

Buchanan, B.G. (2005). A (Very) Brief History of Artificial Intelligence. *AI Magazine 26*(4): 53–60.

Callaghan, V., Miller, J., Armstrong, S. & Yampolinsky, R. (eds.). (2017). *The Technological Singularity: Managing the Journey*. New York, USA: Springer International.

Canovan, M. (1977). *The Political Thought of Hannah Arendt*. London: Methuen.

Canovan, M. (1995). *Hannah Arendt: A Reinterpretation of her Political Thought* (Reprinted.). Cambridge, UK; New York, USA: Cambridge University Press.

Canovan, M. (2006). Arendt's theory of totalitarianism: a reassessment. In Villa, D.R. (ed.). (2006). *The Cambridge Companion to Hannah Arendt*: 25-43.

Carson, C. (2010). Science as instrumental reason: Heidegger, Habermas, Heisenberg. *Continental Philosophy Review 42*(4): 483–509.

Catherine, C., Wachter, S., Mittelstadt, B., Taddeo, M. & Floridi, L. (2018). Artificial Intelligence and the 'Good Society': the US, EU, and UK Approach. *Science and Engineering Ethics 24*(2): 505–528.

Ceballos, G., Ehrlich, P., Barnosky, A., García, A., Pringle, R. & Palmer, T. (2015). Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Science Advances 1*(5): e1400253.

Cholbi, M. (2015). Time, Value, and Collective Immortality. *The Journal of Ethics 19*(2): 197–211.

Christman, J. (2015). Autonomy in Moral and Political Philosophy. . In Zalta, E.N. (ed.) *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Retrieved from http://plato.stanford.edu/archives/spr2015/entries/autonomy-moral/

Cimbala, S.J. (2014). Revisiting the Nuclear 'War Scare' of 1983: Lessons Retro- and Prospectively. *The Journal of Slavic Military Studies 27*(2): 234–253.

Cohen, A. & Lee, S. (eds.). (1986). *Nuclear weapons and the future of humanity: the fundamental questions*. Totowa, USA: Rowman & Allanheld.

Collins, H. & Pinch, T. (2010). *The Golem at large: what you should know about technology* (6. print.). Cambridge, UK: Cambridge University Press.

Cookson, C. (2016, January 27). Google computer triumphs in complex board game battle. *Financial Times*. London. Retrieved from http://www.ft.com/cms/s/0/b8e38a28-c4fa-11e5-b3b1-7b2481276e45.html#axzz3zVfnYLiN

Cooper, A., Brown, T., Price, S., Ford, J. & Waters, C. (2018). Humans are the most significant global geomorphological driving force of the 21st Century. *Anthropocene Review* 5(3): 222-229.

Cooper, D.E. (1997). Wittgenstein, Heidegger and Humility. *Philosophy 72*(279): 105–123.

Copeland, B.J. (2017). The Church-Turing Thesis. In Zalta, E.N. (ed.) *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/church-turing/

Coughlan, S. (2013, April 24). How are humans going to become extinct? *BBC News*. Retrieved from https://www.bbc.co.uk/news/business-22002530.

Coyne, J.A. (2010, November 5). Better all the time. Book Review - What Technology Wants - By Kevin Kelly. *The New York Times*. Retrieved 12 January 2018 from http://www.nytimes.com/2010/11/07/books/review/Coyne-t.html

Crawford, K. (2016). Can an Algorithm be Agonistic? Ten Scenes from Life in Calculated Publics. *Science, Technology, & Human Values 41*(1): 77–92.

Creighton, J. (2018, March 15). OpenAI wants to make safe AI, but that may be an impossible task. *Futurism*. Retrieved 25 April 2018 from https://futurism.com/openai-safe-ai-michael-page/

Crutzen, P.J. (2002). Geology of mankind. *Nature 415*(6867): 23–23.

Danaher, J. (2015). Why AI Doomsayers Are Like Sceptical Theists and Why It Matters. *Minds and Machines 25*(3): 231–246.

Dakers, J. (ed.) (2006). *Defining Technological Literacy*: *Towards an Epistemological Framework*. New York, USA: Palgrave Macmillan.

Davis, E. (2015). Ethical Guidelines for A Superintelligence. *Artificial Intelligence 220*(C): 121-124.

Dawsey, J. (2017). Ontology and Ideology: Günther Anders's Philosophical and Political Confrontation with Heidegger. *Critical Historical Studies 4*(1): 1–37.

Decartes, R. (1911). Discourse on the Method of Rightly Conducting the Reason. In *The Philosophical Works of Descartes*. Vol. I. Haldane, E. S. & Ross, G. R. T. (trans.). Cambridge, UK: Cambridge University Press: 79-130.

Derrida, J. (1984a). Of an Apocalyptic Tone Recently Adopted in Philosophy. *Oxford Literary Review* 6(2): 3-37.

Derrida, J. (1984b). No Apocalypse, Not Now (Full Speed Ahead, Seven Missiles, Seven Missives). *Diacritics 14*(2): 20–31.

Deutsch, D. (2011). *The Beginning of Infinity: Explanations that Transform the World*. London, UK: Allen Lane.

Deutsch, D. (2012, October 3). How close are we to creating artificial intelligence? *Aeon*. Retrieved from https://aeon.co/essays/how-close-are-we-to-creating-artificial-intelligence

Dijk, P. v. (2000). *Anthropology in the age of technology: the philosophical contribution of Günther Anders*. Amsterdam, Netherlands: Rodopi.

Dreyfus, H. (1965). Alchemy and Artificial Intelligence. Santa Monica, CA, USA. RAND Corporation, 1965. Retrieved from https://www.rand.org/pubs/papers/P3244.html

Dreyfus, H. (1968). Cybernetics as the Last Stage of Metaphysics. *Akten des XIV. Internationalen Kongresses für Philosophie*. Retrieved 22 August 2018, from https://www.pdcnet.org/pdc/bvdb.nsf/purchase?openform&fp=wcp14&id=wcp14_1968_0002_0000_0493_0499

Dreyfus, H. (1991). *Being-in-the-world: a commentary on Heidegger's Being and time, division I.* Cambridge, USA: MIT Press.

Dreyfus, H. (1993). Heidegger on the connection between nihilism, art, technology, and politics. In C. Guignon (ed.) (1993), *The Cambridge Companion to Heidegger*: 298-316.

Dreyfus, H. (2007). Why Heideggerian AI Failed and How Fixing it Would Require Making it More Heideggerian. *Philosophical Psychology 20*(2): 247–268.

Dreyfus, H. (2009). Heidegger on Gaining a Free Relation to Technology. In: Kaplan, D.M. (ed.). (2009). *Readings in the Philosophy of Technology*: 25-33.

Dreyfus, H. (2012). A History of First Step Fallacies. *Minds and Machines 22*(2): 87–99.

Dreyfus, H. & Spinosa, C. (1997). Highway bridges and feasts: Heidegger and Borgmann on how to affirm technology. *Man and World 30*(2): 159–178.

Dreyfus, H. & Spinosa, C. (2003). Further Reflections on Heidegger, Technology, and the Everyday. *Bulletin of Science, Technology & Society 23*(5): 339–349.

Dreyfus, H. & Wrathall, M. (eds.) (2005). *A Companion to Heidegger.* New York, USA: Blackwell Publishing.

Dries, C. (2009). *Günther Anders*. Paderborn, Germany: Fink.

Dries, C. (2012). *Die Welt als Vernichtungslager: eine kritische Theorie der Moderne im Anschluss an Günther Anders, Hannah Arendt und Hans Jonas*. Bielefeld, Germany: Transcript Verlag.

Dupré, J. (2001). *Human Nature and the Limits of Science*. Oxford, UK; New York, USA: Clarendon Press; Oxford University Press.

Dupuy, J.-P. (2009a). *On the Origins of Cognitive Science: The Mechanization of the Mind*. Cambridge, USA: MIT Press.

Dupuy, J.-P. (2009b). Technology and Metaphysics. In Friis, J., Pedersen, S. & Hendricks, V. (eds.). (2009). *A companion to the philosophy of technology*: 214-218.

Dupuy, J.-P. (2012). Enlightened Doomsaying and the Concern for the Future. In *Ritsumeikan studies in language and culture*. Presented at the The 8th International Conference of the Graduate School of Core Ethics and Frontier Sciences, Catastrophe and Justice, Ritsumeikan University, Japan. Retrieved from https://ci.nii.ac.jp/naid/110009659938/en

Dupuy, J.-P. (2013). *The Mark of the Sacred*. Stanford, USA: Stanford University Press.

Dupuy, J.-P. (2015). *A Short Treatise on the Metaphysics of Tsunamis*. East Lansing, USA: Michigan State University Press.

Eagleton, T. (2018). *Radical sacrifice*. London, UK; New Haven, USA: Yale University Press.

Eaves, E. (2017, February 28). IARPA Director Jason Matheny advances tech tools for US espionage. *Bulletin of the Atomic Scientists*. Retrieved 15 May 2018, from https://thebulletin.org/2017/march/iarpa-director-jason-matheny-advances-tech-tools-us-espionage10556

Eden, A., Moor, J., Søraker, J. & Steinhart, E. (eds.). (2012). *Singularity Hypotheses: A Scientific and Philosophical Assessment*. Berlin, Germany: Springer.

Edwards, P.N. (2010). *A vast machine: computer models, climate data, and the politics of global warming*. Cambridge, Mass: MIT Press.

Ellis, E., Maslin, M., Boivin, N. & Bauer, A. (2016). Involve social scientists in defining the Anthropocene. *Nature News 540*(7632): 192.

Ellul, J. (1964). *The Technological Society*. New York, USA: Vintage Books.

Esslin, M. (ed.). (1965). *Samuel Beckett. A collection of critical essays.* Englewood Cliffs, USA: Prentice-Hall.

Etzioni, O. (2016, September 20). No, the Experts Don’t Think Superintelligent AI is a Threat to Humanity. *MIT Technology Review*. Retrieved 14 August 2018 from https://www.technologyreview.com/s/602410/no-the-experts-dont-think-superintelligent-ai-is-a-threat-to-humanity/

Farías, V. (1989). *Heidegger and Nazism*. Philadelphia, USA: Temple University Press.

Farmer, J.D. & Lafond, F. (2016). How predictable is technological progress? *Research Policy 45*(3): 647–665.

Feenberg, A. (2002). *Transforming Technology: A Critical Theory Revisited*. Oxford, UK: Oxford University Press.

Feenberg, A. (2005). *Heidegger and Marcuse: the catastrophe and redemption of history*. New York, USA: Routledge.

Feenberg, A. (2006). What Is Philosophy of Technology? In Dakers, J. (ed) (2006). *Defining Technological Literacy*: *Towards an Epistemological Framework*: 5-16.

Felt, U., Fouché, R. & Miller, C. (eds.). (2017). *The handbook of science and technology studies* (Fourth edition.). Cambridge, USA: MIT Press.

Ferré, F. (1988). *Philosophy of Technology*. London: Prentice Hall.

Featherstone, J. (1978). Rousseau and Modernity. *Daedalus 107*(3): 167-192.

Field, A. (2014). Schelling, von Neumann, and the Event that Didn't Occur. *Games 5*(1): 53–89.

IPCC (2014). *Climate Change 2014: Impacts, Adaptation and Vulnerability.* Working Group II Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Field, C.B., V.R. Barros, D.J. Dokken, K.J. Mach, M.D. Mastrandrea, T.E. Bilir, M. Chatterjee, K.L. Ebi, Y.O. Estrada, R.C. Genova, B. Girma, E.S. Kissel, A.N. Levy, S. MacCracken, P.R. Mastrandrea, and L.L.White (eds.). Cambridge, UK; New York, USA: Cambridge University Press.

Floridi, L. (2014). *The 4th revolution: how the infosphere is reshaping human reality* (First edition.). New York, USA; Oxford, UK: Oxford University Press.

Floridi, L. (2015). Singularitarians, AItheists, and Why the Problem with Artificial Intelligence is H.A.L. (Humanity at Large) and not HAL. *APA Newsletter on Philosophy and Computers 14*(2): 8-11.

Floridi, L. (2016, May 9). True AI is both logically possible and utterly implausible. *Aeon*. Retrieved 23 May 2018 from https://aeon.co/essays/true-ai-is-both-logically-possible-and-utterly-implausible

Foot, P. (2002). *Virtues and Vices*. Oxford, UK; New York, USA: Oxford University Press.

Forster, E.M. (2011). *The Machine Stops*. London: Penguin.

Frankish, K. & Ramsey, W. (eds.) (2014), *The Cambridge Handbook of Artificial Intelligence.* Cambridge, UK: Cambridge University Press.

Franssen, M., Lockhorst, G.-J. & Van de Poel, I. (2018). Philosophy of Technology. In Zalta, E.N. (ed). *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/entries/technology/

Freitas, R.A. (1985). There is no Fermi Paradox. *Icarus 62*(3): 518–520.

Frey, C. & Osborne, M. (2017). The future of employment: how susceptible are jobs to computerisation. *Technological Forecasting and Social Change 114*: 254-280.

Frey, C., Osborne, M. & Holmes, C. (eds.). (2016). Technology at work v2.0: The Future Is Not What It Used to Be. *Citi GPS: Global Perspectives & Solutions*. Citigroup and Oxford Martin School. Retrieved from https://www.oxfordmartin.ox.ac.uk/downloads/reports/Citi_GPS_Technology_Work_2.pdf

Friis, J., Pedersen, S. & Hendricks, V. (eds.). (2009). *A companion to the philosophy of technology*. Chichester, UK; Malden, USA: Wiley-Blackwell.

Fukuyama, F. (2002). *Our posthuman future: consequences of the biotechnology revolution* (1st ed.). New York: Farrar, Straus and Giroux.

Garber, D. (2001). *Descartes embodied: reading Cartesian philosophy through Cartesian science*. Cambridge, UK; New York, USA: Cambridge University Press.

Gaus, G., Courtland, S. & Schmidtz, D. (2018). Liberalism. In Zalta, E.N. (ed). *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=liberalism

Geman, S., Bienenstock, E. & Doursat, R. (1992). Neural Networks and the Bias/Variance Dilemma. *Neural Computation 4*(1): 1–58.

Gershenfeld, N. (2012). How to Make Almost Anything: The Digital Fabrication Revolution. *Foreign Affairs 91*(6): 43–57.

Gershgorn, D. (2016, December 14). Microsoft's new plan is to flood your entire life with artificial intelligence. *Quartz*. Retrieved 12 August 2018, from https://qz.com/863058/microsofts-new-plan-is-to-flood-your-entire-life-with-artificial-intelligence/

Glendinning, C. (1990). Notes toward a Neo-Luddite Manifesto. *The Anarchist Library*. Retrieved 22 September 2018, from https://theanarchistlibrary.org/library/chellis-glendinning-notes-toward-a-neo-luddite-manifesto

Gobbo, F. (2016). The Unavoidable Charm of the Superintelligence and Its Risk. *APA Newsletter on Philosophy and Computers 15*(2): 11-12.

Goldstein, J.S. (2011). *Winning the war on war: the decline of armed conflict worldwide*. New York, USA: Dutton.

Good, I.J. (1965). Speculations Concerning the First Ultraintelligent Machine. *Advances in computers 6*: 31–88.

Gray, J. (2007). *Black Mass: Apocalyptic Religion and the End of Utopia.* New York, USA: Farrar, Straus and Giroux.

Gray, J. (2010). *Politik der Apokalypse: wie Religion die Welt in die Krise stürzt.* (C. Trunk, trans.) (3. Aufl.). Stuttgart, Germany: Klett-Cotta.

Gray, J. (2012). *The Immortalization Commission: The Strange Quest to Cheat Death*. London, UK: Penguin.

Gray, J. (2015). *Soul of the Marionette: A Short Enquiry into Human Freedom*. London, UK: Allen Lane.

Gray, J.D. (2017). Scheffler's "Afterlife Conjecture" is Not That Compelling: How His "Doomsday" and "Infertility" Scenarios Might Robustly Preserve Value and Meaning. *Philosophia 45*(2): 637–646.

Gray, K. (2017, November 1). Inside Silicon Valley's new non-religion: consciousness hacking. *WIRED UK*. Retrieved 13 January 2018 from http://www.wired.co.uk/article/consciousness-hacking-silicon-valley-enlightenment-brain

Greenberg, M. (1999, July 1). Apocalypse Not Just Now. *London Review of Books 21*(13): 19–22.

Groĭs, B. (2012). *Introduction to Antiphilosophy*. London, UK; New York, USA: Verso Books.

Groĭs, B. (ed.). (2018). *Russian Cosmism*. Cambridge, USA: MIT Press.

Guez, A., Weber, T., Antonoglou, I., Simonyan, K., Vinyals, O., Wierstra, D., … Silver, D. (2018). Learning to Search with MCTSnets. *arXiv:1802.04697 [cs, stat]*. Retrieved from http://arxiv.org/abs/1802.04697

Guignon, C. (ed.). (1993). *The Cambridge Companion to Heidegger*. Cambridge, UK: Cambridge University Press.

Haas, L. (2012, February 23). A More Crocodile Crocodile. *London Review of Books 34*(4): 28–31.

Habermas, J. (1989). Work and Weltanschauung: The Heidegger Controversy from a German Perspective. *Critical Inquiry 15*(2): 431–456.

Habermas, J. (2003). *The Future of Human Nature*. Cambridge, UK: Polity.

Habermas, J. (2007). *The Postnational Constellation: Political Essays* (Reprinted.). Cambridge, UK: Polity.

Hackett, E.J. & Society for Social Studies of Science (eds.). (2008). *The handbook of science and technology studies* (3rd ed.). Cambridge, USA: MIT Press

Hall, H. (1993). Intentionality and World. In Guignon, C. (ed.), *The Cambridge Companion to Heidegger*: 122-141.

Hansson, S.O. (2014). Risk. In Zalta, E.N. (ed.) *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/spr2014/entries/risk/

Harrari, Y. (2017). *Homo Deus: a brief history of tomorrow* (Revised edition.). London, UK: Vintage.

Harrington, C. (2016). The Ends of the World: International Relations and the Anthropocene. *Millennium 44*(3): 478–498.

Hart, H. (1951). Some Cultural-Lag Problems Which Social Science Has Solved. *American Sociological Review 16*(2): 223–227.

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. Cambridge, USA: MIT Press.

Headrick, D.R. (2010). *Power over peoples: technology, environments, and Western imperialism, 1400 to the present*. Princeton, USA: Princeton University Press.

Heer, J. (2015, November 10). The New Utopians. *The New Republic*. Retrieved from https://newrepublic.com/article/123217/new-utopians

Himmelfarb, M. (2010). *The Apocalypse: A Brief History*. Malden, USA: Wiley-Blackwell.

Hinchman, L.P. & Hinchman, S.K. (1984). In Heidegger's Shadow: Hannah Arendt's Phenomenological Humanism. *The Review of Politics 46*(2): 183–211.

Honig B. (2011). *Emergency Politics: Paradox, Law, Democracy.* Princeton, USA: Princeton University Press

Honneth, A. (2014). Foreword. In: Jaegi, R. (2014). *Alienation*.

Horkheimer, M. & Adorno, T.W. (2002). *Dialectic of Enlightenment: Philosophical Fragments*. Stanford, USA: Stanford University Press.

Hörl, E. (2015). The technological condition. *Parrhesia* (22): 1–15.

Howe, L. & Wain, A. (eds.). (1993). *Predicting the future*. Cambridge, UK; New York, USA: Cambridge University Press.

Hume, D. (2007). *An enquiry concerning human understanding*. (Millican, P.F. ed.). Oxford, UK; New York, USA: Oxford University Press.

Ihde, D. (1990). *Technology and the Lifeworld: From Garden to Earth*. Bloomington, USA: Indiana University Press.

Ihde, D. (2010). *Heidegger's technologies: Postphenomenological Perspectives*. New York, USA: Fordham University Press.

IOS Press (2018). *Applied Ontology: An Interdisciplinary Journal of Ontological Analysis and Conceptual Modeling*. Retrieved 13 October 2018, from https://www.iospress.nl/journal/applied-ontology/

Irpan, A. (2016). AlphaGo vs Lee Sedol: Post Match Commentaries. *alexirpan.com*. Retrieved 18 September 2018, from http://www.alexirpan.com/2016/03/17/alphago-lsd.html

Irving, G., Christiano, P. & Amodei, D. (2018). AI safety via debate. *arXiv:1805.00899 [cs, stat]*. Retrieved from http://arxiv.org/abs/1805.00899

Jaeggi, R. (2014). *Alienation*. New York, USA: Columbia University Press.

Jaspers, K. (1960). *Die Atombombe und die Zukunft des Menschen*. München, Germany: Piper & Co.

Jonas, H. (1953). A Critique of Cybernetics. *Social Research 20*(2): 172–192.

Jonas, H. (1959). The Practical Uses of Theory. *Social Research 51*(1/2): 65–90.

Jonas, H. (1964). Heidegger and Theology. *The Review of Metaphysics 18*(2): 207–233.

Jonas, H. (1973). Technology and Responsibility: Reflections on the New Tasks of Ethics. *Social Research 40*(1): 31–54.

Jonas, H. (1979). Toward a Philosophy of Technology. *The Hastings Center Report 9*(1): 34–43.

Jonas, H. (1981). Reflections on Technology, Progress, and Utopia. *Social Research 48*(3): 411–455.

Jonas, H. (1984). *The Imperative of Responsibility: In Search of an Ethics for the Technological Age*. Chicago, USA: University of Chicago Press.

Jonas, H., Fox, B. & Wolin, R. (2006). Hannah Arendt: An Intimate Portrait. *New England Review* 27(2): 133–142.

Joy, B. (2000). Why the Future Doesn't Need Us. *Wired*. Retrieved from https://www.wired.com/2000/04/joy-2/

Kagan, S. (2012). *Death*. Yale, USA: Yale University Press.

Kalyvas, A. (2008). *Democracy and the Politics of the Extraordinary: Max Weber, Carl Schmitt, and Hannah Arendt.* Cambridge, UK; New York, USA: Cambridge University Press.

Kaplan, D.M. (ed.). (2009). *Readings in the philosophy of technology* (2nd ed.). Lanham, USA: Rowman & Littlefield Publishers.

Kateb, G. (1984). *Hannah Arendt: Politics, Conscience, Evil*. Oxford: Robertson.

Kateb, G. (1997). Technology and Philosophy. *Social Research 64*(3): 1225–1246.

Kateb, G. (2006). *Patriotism and Other Mistakes*. New Haven: Yale University Press.

Kelly, K. (2010). *What technology wants*. New York: Viking.

Knight, W. (2017, Febrauary 22). Paying With Your Face: 10 Breakthrough Technologies 2017. *MIT Technology Review*. Retrieved from https://www.technologyreview.com/s/603494/10-breakthrough-technologies-2017-paying-with-your-face/

Koganzan, R. (2007). Science and Totalitarianism. *The New Atlantis* (18). Retrieved from https://www.thenewatlantis.com/publications/science-and-totalitarianism

Kolbert, E. (2014). *The Sixth Extinction: An Unnatural History.* New York: Henry Holt and Company, LLC

Kolodny, N. & Brunero, J. (2016). Instrumental Rationality. In Zalta, E.N. (ed.). *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/win2016/entries/rationality-instrumental/

Koselleck, R. (2004). *Futures Past: On the Semantics of Historical Time.* New York: Columbia University Press.

Krieger, Z. & Roth, A.I. (2007). Nuclear Weapons in Neo-Realist Theory. *International Studies Review 9*(3): 369–384.

Kroker, A. (2002). Hyper-Heidegger. *CTheory*. Retrieved from http://www.ctheory.net/articles.aspx?id=348

Kroker, A. (2004). *The will to technology and the culture of nihilism: Heidegger, Nietzsche and Marx*. Toronto: University of Toronto Press.

Lambert, F. (2016, July 1). Understanding the fatal Tesla accident on Autopilot and the NHTSA probe. *Electrek*. Retrieved from https://electrek.co/2016/07/01/understanding-fatal-tesla-accident-autopilot-nhtsa-probe/

Lanchester, J. (2015, March 5). The Robots Are Coming. *London Review of Books* 37(5*)*: 3–8.

Lanchester, J. (2017, August 17). You Are the Product. *London Review of Books* 39(16): 3–10.

Latour, B. & Weibel, P. eds. (2005) *Making Things Public: Atmospheres of Democracy*. Cambridge, USA [Karlsruhe, Germany]: MIT Press [ZKM/Center for Art and Media in Karlsruhe]

Law, J. (2017). STS as Method. In Felt, U., Fouché, R. & Miller, C. (eds.). (2017). *The handbook of science and technology studies*: 31-58.

Lawler, A. (n.d.). Our Proud Human Future. *The New Atlantis*. Retrieved 2 February 2018, from https://www.thenewatlantis.com/publications/our-proud-human-future

Lawrence, N. (2016). Future of AI 6. Discussion of ‘Superintelligence: Paths, Dangers, Strategies’. *http://inverseprobability.com*. Retrieved on 12 September 2018, from http://inverseprobability.com/2016/05/09/machine-learning-futures-6

Lawson, C. (2017). *Technology and isolation*. Cambridge, UK; New York, USA: Cambridge University Press.

Lebedev, A. (2011, July 21). Stanislav Petrov. Retrieved 2 August 2018, from https://web.archive.org/web/20110721000030/http://www.worldcitizens.org/petrov2.html

Lee, C.-S., Wang, M.-H., Yen, S.-J., Wei, T.-H., Wu, I.-C., Chou, P.-C., … Yang, T.-H. (2016). Human vs. Computer Go: Review and Prospect. *arXiv:1606.02032 [cs]*. Retrieved from http://arxiv.org/abs/1606.02032

Legg, S. & Hutter, M. (2007a). A Collection of Definitions of Intelligence. *arXiv:0706.3639 [cs]*. Retrieved from http://arxiv.org/abs/0706.3639

Legg, S., & Hutter, M. (2007b). Universal Intelligence: A Definition of Machine Intelligence. *Minds and Machines* 17(4): 391-444.

Lewis, C.S. (1943/2013). *The Abolition of Man*. England: Important Books.

Lewis, P. & Pelopidas, B. (2014). *Too Close for Comfort: Cases of Near Nuclear Use and Options for Policy*. London: Chatham House - The Royal Institute for International Affairs. Retrieved from https://www.chathamhouse.org//node/13981

Liessmann, K.P. (2002). *Günther Anders: philosophieren im Zeitalter der technologischen Revolutionen*. München: C.H. Beck.

Liessmann, K. P. (2011). Thought after Auschwitz and Hiroshima: Günther Anders and Hannah Arendt. *Enranohar 43*: 123–135.

Loscerbo, J. (1981). *Being and technology: a study in the philosophy of Martin Heidegger*. The Hague ; Boston : Hingham, Mass: Nijhoff ; Distributed by Kluwer Boston.

Machery, E. (2008). A Plea for Human Nature. *Philosophical Psychology 21*(3): 321–329.

Mackay, R. & Avanessian, A. (eds.). (2014). *Accelerate: the accelerationist reader*. Falmouth, UK: Urbanomic Media Ltd.

Martin, S. (2017, April 1). Humanity will go EXTINCT in 100 years, as prominent scientist warns 'IT'S TOO LATE' | Science | News | Express.co.uk. Retrieved from https://www.express.co.uk/news/science/785903/Humanity-EXTINCT-TOO-LATE-frank-fenner.

McCorduck, P. (2004). *Machines who think: a personal inquiry into the history and prospects of artificial intelligence* (25th anniversary update.). Natick, Mass: A.K. Peters.

McCormick, J.P. (1994). Fear, Technology, and the State: Carl Schmitt, Leo Strauss, and the Revival of Hobbes in Weimar and National Socialist Germany. *Political Theory 22*(4): 619–652.

McMahan, J. (1986). Nuclear Deterrence and Future Generations. In Cohen. A. and Lee, S. (eds) (1986). *Nuclear Weapons and the Future of Humanity: The Fundamental Questions*: 319-339.

Meyer, D. (n.d.). Vladimir Putin Says Whoever Leads in Artificial Intelligence Will Rule the World. *Fortune*. Retrieved 13 July 2018 from http://fortune.com/2017/09/04/ai-artificial-intelligence-putin-rule-world/

Misa, T.J., Brey, P. & Feenberg, A. (eds.). (2003). *Modernity and technology*. Cambridge, MA: MIT Press.

Mitcham, C. (1987). Responsibility and Technology: The Expanding Relationship. In P. T. Durbin (ed.), *Technology and Responsibility*. Dordrecht: Springer Netherlands.

Mitcham, C. (1994). *Thinking through technology: the path between engineering and philosophy*. Chicago: University of Chicago Press.

Mitchell, A. (2017). Is IR going extinct? *European Journal of International Relations 23*(1): 3–25.

Moore, G.E. (1965). Cramming more components onto integrated circuits. *Electronics Magazine* 38(8): 114 ff.

Morozov, E. (2012, April 5). In Your Face. *London Review of Books* 34(7), pp. 25–27.

Müller, C. (2015). Desert Ethics: Technology and the Question of Evil in Günther Anders and Jacques Derrida. *Parallax 21*(1): 42–57.

Müller, C.J. (2016). *Prometheanism: technology, digital culture, and human obsolescence*. London New York: Rowman & Littlefield International.

Müller, V.C. (2014). Risks of general artificial intelligence. *Journal of Experimental & Theoretical Artificial Intelligence 26*(3): 297–301.

Munster, R. van & Sylvest, C. (2016a). *Nuclear realism: global political thought during the thermonuclear revolution*. London, New York: Routledge.

Munster, R. van & Sylvest, C. (2016b). *The politics of globality since 1945: assembling the planet*. London, New York: Routledge.

Munthe, C. (2015). Philosophical Comment: Why Aren't Existential Risk / Ultimate Harm Argument Advocates All Attending Mass? *Philosophical Comment*. Retrieved from http://philosophicalcomment.blogspot.com/2015/02/why-arent-existential-risk-ultimate.html

Munthe, C. (2017). The Black Hole Challenge: Precaution, Existential Risks and the Problem of Knowledge Gaps. *Ethics, Policy & Environment.* forthcoming.

Nagel, T. (1974). What Is It Like to Be a Bat? *The Philosophical Review 83*(4): 435–450.

Nancy, J.-L. (2000). *Being singular plural*. Stanford, Calif: Stanford University Press.

Nancy, J.-L. (2015). *After Fukushima: the equivalence of catastrophes*. New York: Fordham University Press.

Newell, A., Shaw, J.C. & Simon, H.A. (1958). Chess-playing Programs and the Problem of Complexity. *IBM J. Res. Dev. 2*(4): 320–335.

Nisbet, R.A. (1994). *History of the idea of progress* (4. printing.). New Brunswick, NJ [u.a.]: Transaction Publ.

Nye, D.E. (2006). *Technology matters: questions to live with*. Cambridge, USA: MIT Press.

Olsen, J., Selinger, E. & Riis, S. (eds.). (2009). *New Waves in Philosophy of Technology*. New York, USA: Palgrave Macmillan.

OpenAI (2018). OpenAI Five. *OpenAI Blog*. Retrieved 20 August 2018, from https://blog.openai.com/openai-five/

Oremus, W. (2016, January 3). Who Controls Your Facebook Feed. *Slate*. Retrieved from http://www.slate.com/articles/technology/cover_story/2016/01/how_facebook_s_news_feed_algorithm_works.single.html

Parfit, D. (1984). *Reasons and Persons.* Oxford, UK: Clarendon Press.

Pellissier, H. (2015, December 23). Should Politicians be Replaced by Artificial Intelligence? Interview with Mark Waser. *Ethical technology*. Institute for Ethics and Emerging Technologies. Retrieved 22 February 2018 from https://ieet.org/index.php/IEET2/more/pellissier20151223

Pelopidas, B. (2013). Why nuclear realism is unrealistic. *Bulletin of the Atomic Scientists*. Retrieved 1 January 2018 from https://thebulletin.org/why-nuclear-realism-unrealistic

Petrina, S. (2017). 'Critique of Technology'. In: Williams, P. & Stables, K. (eds.) (2017). *Critique in Design and Technology Education*: 31 - 49

Pinker, S. (2018). *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress*. New York, USA: Viking, Penguin Random House.

Pöggeler, O. (1994). *Der Denkweg Martin Heideggers* (4th ed.). Stuttgart, Germany: Neske.

Polanyi, M. (2009). *The tacit dimension*. Chicago, USA; London, UK: University of Chicago Press.

Popper, K.R. (2013). *The Open Society and Its Enemies*. Princeton, USA: Princeton University Press.

Posner, R. (2006). Efficient Responses to Catastrophic Risk. 6 *Chicago Journal of International Law* (511): 1-17.

GiveWell (2015, July 2). Potential global catastrophic risk focus areas. *The GiveWell blog*. Retrieved from http://blog.givewell.org/2014/06/26/potential-global-catastrophic-risk-focus-areas/

Raffoul, F. (ed.). (2013). *The Bloomsbury companion to Heidegger*. London, New York: Bloomsbury.

Reilley, K. (2011). *Automata and Mimesis on the Stage of Theatre History*. New York: Palgrave Macmillan.

Rescorla, M. (2017). The Computational Theory of Mind. In Zalta, E.N. (ed.). *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/spr2016/entries/frame-problem/

Retrieved from https://plato.stanford.edu/entries/computational-mind/

Reville, W. (2016, December 8). How long will the human species survive on Earth? *The Irish Times*. Retrieved on October 10th 2017, from https://www.irishtimes.com/news/science/how-long-will-the-human-species-survive-on-earth-1.2885564

Roden, D. (2015). *Posthuman Life: Philosophy at the Edge of the Human*. USA: Routledge

Rorty, R. (2005). 'Heidegger and the Atom Bomb'. In: *Making Things Public. Atmospheres of Democracy*, Latour, B. & Weibel, P (eds.): 274–75.

Rorty, R. (2007). *Philosophy as cultural politics*. Cambridge, UK ; New York: Cambridge University Press.

Rubin, C. (2007). Thumos in Space. *The New Atlantis* (18): 66–71.

Runciman, D. (2012). What Is Realistic Political Philosophy? *Metaphilosophy 43*(1–2): 58–70.

Runciman, D. (2015) Digital Politics: Why Progressives Need to Shape the Digital Economy. *Juncture 22*(1): 11-16.

Russell, A. & Vinsel, L. (2018). Is a mission to Mars morally defensible given today's real needs? – Andrew Russell & Lee Vinsel | Aeon Essays. *Aeon*. Retrieved 14 August 2018, from https://aeon.co/essays/is-a-mission-to-mars-morally-defensible-given-todays-real-needs

Rusell, B. (1924). *Icarus: Or the Future of Science*. London: K. Paul, Trench, Trubner & Co., ltd.

Russell, B. & Einstein, A. (1955). The Russell-Einstein Manifesto. Retrieved from https://pugwash.org/1955/07/09/statement-manifesto/

Sagan, C. (1983). Nuclear War and Climatic Catastrophe: Some Policy Implications. *Foreign Affairs*. Winter 1983/1984 Issue. Retrieved on 13th May 2016, from https://www.foreignaffairs.com/articles/1983-12-01/nuclear-war-and-climatic-catastrophe-some-policy-implications

Sample, I. (2008). CERN throws switch on largest machine ever built. *The Guardian*. Retrieved on 21st January 2018, from https://www.theguardian.com/science/blog/2008/sep/10/cern.large.hadron.collider

Sandel, M.J. (2004). The Case Against Perfection. *The Atlantic*. April 2014 Issue. Retrieved from https://www.theatlantic.com/magazine/archive/2004/04/the-case-against-perfection/302927/

Sauer, E. (1968). *Deutsche Philosophen - Von Eckhart bis Heidegger*. Göttingen: Musterschmidt.

Scharf, C. (2016, March 22). Where do minds belong? *Aeon*. Retrieved 11 March 2018 from https://aeon.co/essays/intelligent-machines-might-want-to-become-biological-again

Scharff, R.C. & Dusek, V. (eds.). (2014). *Philosophy of technology: the technological condition: an anthology*. Malden, MA: Wiley Blackwell.

Scheffler, S. (2016). *Death and the afterlife*. New York, NY: Oxford University Press.

Schell, J. (1982). *The Fate of the Earth.* New York, NY: Alfred Knopf.

Schell, J. (2010). In Search of a Miracle: Hannah Arendt and the Atomic Bomb. In S. Benhabib (ed.), *Politics in Dark Times*. Cambridge: Cambridge University Press.

Schiff, J. (2013). The varieties of thoughtlessness and the limits of thinking. *European Journal of Political Theory 12*(2): 99–115.

Schilpp, P. (1983). Heidegger: 'Nur noch ein Gott kann uns retten': In *Der 16. Weltkongress für Philosophie*. Peter Lang. Retrieved from http://www.pdcnet.org/oom/service?url_ver=Z39.88-2004&rft_val_fmt=&rft.imuse_id=wcp16_1983_0002_1242_1248&svc_id=info:www.pdcnet.org/collection

Schmidt, E. & Cohen, J. (2014). *The new digital age: reshaping the future of people, nations and business*. London: Murray.

Schmitt, C. (1991). *Glossarium: Aufzeichnungen der Jahre 1947-1951*. Berlin: Duncker & Humblot.

Schmitt, C. (1996). *The Concept of the Political*. Chicago: University of Chicago Press.

Schneider, J. (1945). Cultural Lag: What Is It? *American Sociological Review 10*(6): 786–791.

Schneider, S. (2017). Superintelligent AI and the Postbiological Cosmos Approach. In Losch, A. (ed.) (2017). *What is Life? On Earth and Beyond.* Cambridge, UK: Cambridge University Press: 178-198.

Schneider, S. & Gee, M. (2016). Extraterrestrials May Be Robots Without Consciousness. *Cosmos on Nautilus*. Retrieved 23 February 2018 from http://cosmos.nautil.us/feature/72/it-may-not-feel-like-anything-to-be-an-alien

Schneier, B. (2013). Our Security Models Will Never Work—No Matter What We Do. *Schneier on Security*. Retrieved from https://www.schneier.com/essays/archives/2013/03/our_security_models.html

Schraube, E. (2005). ‘Torturing things until they confess’: Günther Anders’ critique of technology. *Science as Culture 14*(1): 77–85.

Schürmann, R. (1978). Political Thinking in Heidegger. *Social Research 45*(1): 191–221.

Schwab, K. (2015, December 16). The Fourth Industrial Revolution. *Foreign Affairs*. Retrieved 19 January 2016, from https://www.foreignaffairs.com/articles/2015-12-12/fourth-industrial-revolution

Seefried, E. (2017). Globalized Science. The 1970s Futures Field: Globalized science. *Centaurus 59*(1–2): 40–57.

Shanahan, M. (2016). The Frame Problem. In (E.N. Zalta, ed.)*The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/spr2016/entries/frame-problem/

Sheehan, T. (ed.). (1981). *Heidegger: The Man and the Thinker*. New York, USA: Routledge.

Sheehan, T. (2012). The turn. In B. W. Davis (ed.), *Martin Heidegger: key concepts*. Durham, UK: Acumen Publishing Limited.

Sheehan, T. (2013). The Turn: All Three of Them. In Francois Raffoul & E. S. Nelson (eds.), *The Bloomsbury Companion to Heidegger*. London, UK: Bloomsbury Academic.

Shuckburgh, E. (ed.). (2008). *Survival: the survival of the human race*. Cambridge, UK; New York, USA: Cambridge University Press.

Simbirski, B. (2016). Cybernetic Muse: Hannah Arendt on Automation, 1951–1958. *Journal of the History of Ideas 77*(4): 589–613.

Simon, H.A. (1996). *The sciences of the artificial* (3. ed.). Cambridge, MA.: MIT Press.

Simonite, T. (2016, March 31). How Google Plans to Solve Artificial Intelligence. *MIT Technology Review*. Retrieved 25 May 2018, from https://www.technologyreview.com/s/601139/how-google-plans-to-solve-artificial-intelligence/

Skolimowski, H. (1966). The Social Character of Technological Problems: Comments on Skolimowski’s Paper. *Technology and Culture 7*(3): 384–390.

Soares, N. (2014). Research Guide. *Machine Intelligence Research Institute*. Retrieved 7 July 2018, from https://intelligence.org/research-guide/

Sodikoff, G.M. (ed.). (2012). *The anthropology of extinction: essays on culture and species death*. Bloomington: Indiana University Press.

Srinivasan, A. (2015). Stop the Robot Apocalypse. *London Review of Books* 37(18): 3–6.

Srinivasan, A. (2017, January 6). Remembering Derek Parfit. *LRB blog*. Retrieved from https://www.lrb.co.uk/blog/2017/01/06/amia-srinivasan/remembering-derek-parfit/

Steffen, W., Crutzen, P., Grinevald, J. & McNeill, J. (2011). The Anthropocene: conceptual and historical perspectives. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 369*(1938): 842–867.

Stern, N. (2016a). Economics: Current climate models are grossly misleading. *Nature Comment 530*(7591): 407.

Stern, N. (2016b). *Why are we waiting? the logic, urgency, and promise of tackling climate change.* Cambridge, MA: MIT Press.

Strong, T. (n.d.). Foreword: Dimensions of the new debate around Carl Schmitt. In Schmitt, C. (1996). *The Concept of the Political*. Chicago: University of Chicago Press.

Strong, T.B. (2012). *Politics without vision: thinking without a banister in the twentieth century*. Chicago: University of Chicago Press.

Sunstein, C. (2005). *Laws of Fear: Beyond the Precautionary Principle.* New York: Cambridge University Press.

Sunstein, C. (2009). *Worst-Case Scenarios*. Cambridge, MA: Harvard University Press.

Talbott, S. (2007). Ghosts in the Evolutionary Machinery. *The New Atlantis* (18): 26–40.

Taleb, N.N. (2007). *The Black Swan: The Impact of the Highly Improbable* (1st ed.). New York: Random House.

Tandy, C. (2004). *Death and Anti-Death: Two Hundred Years After Kant, Fifty Years After Turing.* Palo Alto, USA.: Ria University Press.

Thiele, L.P. (2016). Twilight of Modernity: Nietzsche, Heidegger, and Politics. *Political Theory* 2(3): 468-490.

Thomson, I. (2009). Understanding Technology Ontotheologically, or: The Danger and the Promise of Heidegger, an American Perspective. In Olsen, J., Selinger, E. & Riis, S. (eds.) (2009), *New Waves in Philosophy of Technology.* New York: Palgrave Macmillan.

Thomson, I. (2015). Heidegger's Aesthetics. In (E.N. Zalta, ed.). *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/fall2015/entries/heidegger-aesthetics/

Todd, B. (2017). Why despite global progress, humanity is probably facing its most dangerous time ever. *80,000 Hours*. Retrieved from https://80000hours.org/articles/extinction-risk/

Toscano, A. (2016). The Promethean Gap: Modernism, Machines, and the Obsolescence of Man. *Modernism/modernity 23*(3): 593–609.

Unal, B. & Lewis, P. (2018). *Cybersecurity of Nuclear Weapons Systems* (Research Paper). London: Chatham House - The Royal Institute for International Affairs. Retrieved from https://reader.chathamhouse.org/cybersecurity-nuclear-weapons-systems-threats-vulnerabilities-and-consequences

Unal, B., Lewis, P. & Royal Institute of International Affairs. (2017). *Cybersecurity of nuclear weapons systems: threats, vulnerabilities and consequences*. Retrieved from https://www.chathamhouse.org/sites/files/chathamhouse/publications/research/2018-01-11-cybersecurity-nuclear-weapons-unal-lewis-final.pdf

Yudkowsky, E. (2009). Value is Fragile. *The less wrong blog.* Retrieved 30 August 2018, from https://www.lesswrong.com/posts/GNnHHmm8EzePmKzPk/value-is-fragile

Villa, D.R. (1996). *Arendt and Heidegger: the fate of the political*. Princeton, USA: Princeton University Press.

Villa, D.R. (1999). *Politics, philosophy, terror: essays on the thought of Hannah Arendt*. Princeton, N.J. Chichester: Princeton University Press.

Villa, D.R. (ed.). (2006). *The Cambridge Companion to Hannah Arendt*. Cambridge, UK: Cambridge University Press.

Vogel, L. (1995). Hans Jonas's diagnosis of nihilism: The case of Heidegger. *International Journal of Philosophical Studies 3*(1): 55–72.

Vosicky, L.M. (2005). Anders' Heidegger – Heidegger anders. *Phenomenology 4*(2): 895-935.

Vuori, J.A. (2010). A Timely Prophet? The Doomsday Clock as a Visualization of Securitization Moves with a Global Referent Object. *Security Dialogue 41*(3): 255–277.

Wagner, T. (2018, February 7). Den Menschen überwinden und die Welt retten | NZZ. *Neue Zürcher Zeitung*. Retrieved from https://www.nzz.ch/feuilleton/den-menschen-ueberwinden-und-die-welt-retten-ld.1353304

Wake, D., Vredenburg, V. (2008). Are we in the midst of the sixth mass extinction? A view from the world of amphibians. *Proceedings of the National Academy of Sciences of the United States of America (Supplement 1)*: 11466-11473

Wallas, G. (1908). *Human Nature in Politics*. London, UK: Constable & Company LTD.

Walsh, T. (2017). *Android Dreams: The Past, Present and Future of Artificial Intelligence*. London, UK: C Hurst & Co Publishers Ltd.

Weitzman, M. (2011). Fat-Tailed Uncertainty in the Economics of Climate Change. *Review of Environmental Economics and Policy* 5(2): 275–292.

Wenar, L. (2016). Is Humanity Getting Better? *Opinionator*. Retrieved 17 February 2016, from http://opinionator.blogs.nytimes.com/2016/02/15/is-humanity-getting-better/

Wheeler, M. (2011). Martin Heidegger. In Zalta, E.N. (ed.), *The Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/entries/heidegger/

White, S. (2000). *Sustaining Affirmation: The Strengths of Weak Ontology in Political Theory.* Princeton, Oxford: Princeton University Press.

Wiener, J. (2016). The Tragedy of the Uncommons: On the Politics of Apocalypse. *Global policy* 7(1): 67-80.

Wiese, C. (2007). *The life and thought of Hans Jonas: Jewish dimensions*. Waltham, MA: Brandeis University Press.

Williams, P. and Stables, K. (eds.) (2017). *Critique in Design and Technology Education*, Singapor: Springer Singapore.

Winner, L. (1979). The political philosophy of alternative technology: Historical roots and present prospects. *Technology in Society* 1(1): 75–86.

Winner, L. (1980). Do Artifacts Have Politics? *Daedalus 109*(1): 121–136.

Winner, L. (2001). *Autonomous technology: technics-out-of-control as a theme in political thought* (9. printing.). Cambridge, Mass.: MIT Press.

Winner, L. (2014). Technologies as Forms of Life. In R. L. Sandler (ed.), *Ethics and Emerging Technologies*. London: Palgrave Macmillan UK.

Wittes, B. & Blum, G. (2015). *The future of violence: robots and germs, hackers and drones: confronting a new age of threat*. New York: Basic Books.

Wittgenstein, L. (1965). Lecture on Ethics. *The Philosophical Review 74*(1): 3–12.

Wolin, R. (1990). *The Politics of Being: The Political Thought of Martin Heidegger*. New York: Columbia University Press.

Wolin, R. (ed) (1998). *The Heidegger Controversy: A Critical Reader*. (3rd MIT Press edition). Cambridge, MA: MIT Press.

Wolin, R. (2001). *Heidegger's Children: Hannah Arendt, Karl Lowith, Hans Jonas, and Herbert Marcuse*. Princeton, NJ: Princeton University Press.

Working Group on the 'Anthropocene' (2018). Subcommission on Quaternary Stratigraphy. Retrieved 3 November 2018, from http://quaternary.stratigraphy.org/working-groups/anthropocene/

Wyatt, S. (2008). Technological determinism is dead; long live technological determinism. In Hackett, E.J. et al. & Society for Social Studies of Science (eds.). (2008). *The handbook of science and technology studies* (3rd ed.). Cambridge, MA: MIT Press: 165–180.

Yaqoob, W. (2014). The Archimedean point: Science and technology in the thought of Hannah Arendt, 1951–1963. *Journal of European Studies 44*(3): 199–224.

Zalasiewicz, J., Waters, C. & Head, M.J. (2017). Anthropocene: its stratigraphic basis. *Nature 541*: 289.

Zalta, E.N. (ed.). (n.d.). *Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu