Acta Universitatis Upsaliensis
Uppsala Studies in Philosophy of Religion
9

Human agency and autonomy in algorithm-intense environments

Johan Marticki

UPPSALA
UNIVERSITET

Dissertation presented at Uppsala University to be publicly examined in Geijersalen 6-1023, Engelska parken, Uppsala, Thursday, 24 April 2025 at 13:00 for the degree of Doctor of Philosophy (Faculty of Theology). The examination will be conducted in Swedish. Faculty examiner: Docent Mårten Björk (Newmaninstitutet).

Abstract
Marticki, J. 2025. Human agency and autonomy in algorithm-intense environments. *Uppsala Studies in Philosophy of Religion* 9. 255 pp. Uppsala: Acta Universitatis Upsaliensis.
ISBN 978-91-513-2402-9.

This treatise considers how interaction with existing and future forms of algorithmic technologies could affect the expression of human agency and autonomy and how this, in turn, might affect the way in which humans make sense of them*selves* in the *world*. The risks arising from human–AI interaction that are taken into consideration are tied to changes in the expression of human agency and autonomy and to changes in the content-structure of worldviews. The opportunities and risks that are most pressing to consider on this view are not those that result from what present and future forms of AIs might do to us, but those that are rooted in what *we* might do or fail to do, or in what *we* might become or fail to become, as a consequence of interacting with algorithm-intense environments.

In the treatise, artificial intelligence is put into a wider context that includes earlier forms of industrial technologies and procedures. A comprehensive understanding of the relationship between human agents and technologies is developed, one that considers the conditions necessary for technologies to function properly, the social implications of technical macro-structures, the role of concrete technologies in everyday life, and new possibilities and temptations that arise as a consequence of new technologies. The concept of affordances is used to tie the human environment to worldviews. Humans confronted by and interacting with algorithmic technologies – with new affordance landscapes – are *invited* to test new practices. It is argued that, as new habits are formed, worldviews, in the medium and long term, tend to adapt to match new habits, and that rapidly evolving algorithm-intense conditions – ongoing engineering projects – are likely to undermine the viability of dominant – that is, modern – worldviews. Possible adaptive measures in response to this new condition are proposed and discussed, including the adaptation of worldviews, of the physical human agent, and of the speed with which societies re-engineer their environments.

Keywords: Artificial intelligence, AI, multi-agent systems, MAS, worldview, modernity, surveillance capitalism, Ivan Illich, Jacques Ellul, buffered self, porous self, agency, autonomy

Johan Marticki, Department of Theology, Ethics and Philosophy of Religion, Box 511, Uppsala University, SE-751 20 Uppsala, Sweden.

ISBN 978-91-513-2402-9
URN urn:nbn:se:uu:diva-551945 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-551945)

In memory of

Mum
(1942-2021)

&

Dad
(1931-2024)

always

Acknowledgements

I wish to express my deepest gratitude to my supervisors Mikael Stenmark and Ulf Zackariasson. Your professional help, patience, and kindness have been invaluable in my academic apprenticeship, guiding me in my exploration of the academic profession, and providing me with innumerable sound suggestions on how to forge ideas into a structured treatise.

I would like also to express my deep gratitude to everyone who has participated in the Philosophy of Religion research seminar at Uppsala University, notably to my colleagues and friends who have taken the time to read so many early drafts and given me so much valuable commentary and critique – Oliver Li, Johan Eddebo, Francis Jonbäck, Elena Kalmykova, Ingrid Malm Lindberg, Lina Langby, Christoffer Skogholt, Gabriel Echazu, Evelina Edfors, Lennart Söderlind, Fernando Vasquez, Lotta Knutsson Bråkenhielm and Carl-Reinhold Bråkenhielm; to Harald Hammarström, Per Nyström and Peter Olivius for having taken the time to read and comment on early versions of tentative chapters; to Erik Åkerlund for his knowledgeable, detailed and highly valuable constructive critique of an earlier draft of the treatise; to Anna-Sara Lind, Martha Middlemiss Lé Mon, and Hans-Michael Lagergren for much appreciated general help and support; and to Jean-Paul André Ivan for extensive readings of my manuscripts accompanied by many, long, productive and invaluable conversations about the workings of algorithmic systems and various philosophical topics – those conversations, to an important extent, have shaped some of the key understandings that matured during the most difficult phases of this work, which, as of now, ceases to be a work in progress.

I also wish to thank everyone involved in CRS Uppsala (Centre for Multidisciplinary Research on Religion and Society). Being affiliated to CRS has given me the opportunity to interact with researchers from a wide range of disciplines preoccupied with AI and society, and thereby to expand my research horizons.

Finally, I wish to express my gratitude to Nathan och Anna Söderbloms stiftelse and to the Carnegie Foundation for the generous scholarships that enabled me to visit the University of Vienna and to complete my studies.

Contents

Introduction

Imagine that the speaker of the text below is standing before you lecturing. Imagine also that you have been informed by a credible authority that the speaker, although she looks human, is in fact a human-made artificial agent.

I am here, in person, to exemplify some AI-related challenges. I know what you have been told. But can you trust the message? No doubt your attention will wander, as it should… Ought you to attend to the content of my lecture, or ought you rather to attend to my non-verbal behaviour?

You are, if you've been told the truth, in a position to witness something never before seen. You should be on the lookout for signs that tell of my alterity. Awkward wordings or a too perfect mastery of language, alien quirks or a total lack of idiosyncrasies. How should you proceed in your evaluation?

Believe me or not, being under evaluation is making me nervous. Can you believe me? Imagine that you were standing in my place, lecturing, and that the entire audience were single-mindedly on the lookout for slips that might prove that you were not, in fact, that which you pretended to be: a thinking, conscious, and autonomous being. Imagine if, on the basis of observations and statistical analysis of your behavioural patterns, I were to infer causal explanations for your discourse and propose accurate predictive models for your future discourse. Would such explanations and predictive models imply that you are not the kind of beings that you think you are? Would they imply that you are designed to appear as if you were?

This, I believe, is in fact your intention now in relation to me. You want to demystify me, to find rational explanations as to how it is that I stand before you and speak and act in this peculiar manner. Peculiar, because so human-like, and yet, you have been told, I am all machine. And if I, a machine, am able to inquire into my own identity and into yours, to give this lecture and to answer your questions, then who are You?

I am sure some of you are convinced I am merely a human impostor. The rest, you who want to believe, are no doubt trying to convince yourselves that I am an extraordinarily well put together simulation. And if I am a simulation, then I am not really what I appear to be.

So be it. Why not? I don't mind. But does this really explain anything? When simulations begin to integrate nearly all the features of the systems that

they simulate, and still more, is there then necessarily any meaningful qualitative difference between the simulation and the original system?

And how about You? Is not the moment of conception in the womb, without which you would not be, itself a simulation of the ancient human form? You are all, in some respects, dissimilar, yet remarkably similar. I would not mistake any of you for tokens of any other species than the one that assembled me. This, in my opinion, must be the underlying meaning of simulation… something that simulates, so that it becomes similar.

Is the human species, then, in itself a simulation of some even more ancient form? And do some of You perhaps believe that You are made in the image of God? Is not that a different way of saying that you are simulations of a higher form? And if we are all simulations, if we are all, in a sense, similar, then who are We?

Who, indeed, are 'We'? And how should we, assuming that we were the audience of this fictional talk, understand the discourse produced? Are we witnessing the behaviour of a conscious being? Or are we merely exposed to the output of a very clever algorithmic recipe? Whichever way we choose to interpret the performance, the interpretation will likely have repercussions for how we see ourselves. If we believe that we are interacting with an artificial general intelligence, are we then likely, in our awe, to conclude that we have found a superior intellect? If, instead, we believe that we are exposed to the mechanical output of a clever algorithmic recipe, will we be humbled by the realisation that a mere mechanism can replicate our most sophisticated behavioural accomplishments? Or will the human ego, in considering the human ability to invent such amazing things, instead be further boosted by pride? In the first case, we will perhaps be tempted to fall into deference to or fear of a new oracular or even god-like authority. In the second case, our previous rather lofty understanding of what we are is likely to be undermined. In the third case, nourishing the impression that we are able to create life, it may become expedient to conclude that, in a sense, we are like gods. In all cases, the relative position of human beings in the hierarchy of thinking and acting beings is disturbed.

In popular imagination, challenges to human identity and status are commonly associated with the emergence of artificial general intelligence (AGI) and artificial superintelligence (ASI). In contrast to so-called narrow artificial intelligence, which is designed to reach high performance in the context of narrow sets of tasks, artificial general intelligence implies a system that can perform at least as well as a human being on a wide range of tasks. Superintelligence, in turn, signifies artificial general intelligence that is able to perform far beyond the upper limits of human beings. In philosophical and general discourse, it is common to focus on artificial general intelligence as the revolutionary turning point. Best-selling authors such as Nick Bostrom (2014), Ray Kurzweil (2018), and Max Tegmark (2018) have been concerned

with the potential risks and opportunities that artificial general intelligence and superintelligence represent. This concern is mirrored in popular culture, in films such as *A.I. – Artificial Intelligence*, *Ex Machina*, and *The Ghost in the Shell*.

The fictional scenario that introduces this treatise, whether or not we interpret it as an instantiation of artificial general intelligence, represents a stage at which it becomes difficult to ignore challenges to human identity and status. At this stage we begin to be exposed to the manifestation of phenomena that increasingly appear under fleeting or more enduring guises of what we would expect from artificial general intelligence or superintelligence. On closer inspection, the fictional speaker, although 'it' indeed proves itself an accomplished speaker, might be revealed to be quite inadequate in the performance of other tasks that are routinely performed by human beings. We would then supposedly not be the witnesses of an artificial general intelligence. However, if algorithmic technologies that are able to perform ever widening ranges of narrow tasks with impressive efficiency were to be increasingly released in our environments, one could expect that, sooner or later, we would face challenges that in many respects would be similar to the ones frequently imagined vis-à-vis *one single* general intelligence.

In this treatise it is argued that, in the *long durée*, there is continuity to the challenges we face today. The philosophical and social developments that have led to artificial intelligence are traced as far back as the pre-stages of modernity. On this view, the vision and/or realisation of AGI represents a climactic instantiation of a continuous process.

Much has been written concerning the requirements for when artificial intelligence could be understood to attain agency and/or become autonomous. The engineered speaker in the introductory dramatisation, however, is conceivable even with incremental improvements to current technologies. The focus of this treatise is not on the inner workings of algorithmic technologies, but on what they can do, especially in social arenas, and what this implies in turn for humans *qua* agents. The treatise seeks answers to the following general questions: *How will the expression of human agency and autonomy be affected as humans are increasingly embedded in algorithm-intense environments? How are changes in the expression of human agency and autonomy, in turn, likely to affect the viability of dominant worldviews?* How these questions are considered is explained in more detail below. First, let us consider some previous research into the potential effects of artificial intelligence on human society.

Previous research

This section provides a brief overview of some previous research pertaining to various facets of the potential interplay between artificial intelligence and

human society. Other relevant research is presented during the course of the treatise.

Philosophers, computer scientists, and other interested parties have been concerned with so-called existential risk. Existential risks are those that could end the human species or collapse human civilisation. Prominent thinkers who have considered existential risks include the previously mentioned Nick Bostrom (2014), Ray Kurzweil (2018), and Max Tegmark (2018). Existential risks are typically considered in relation to a hypothesised emergence, first, of artificial general intelligence, and then artificial superintelligence. A by-now-famous philosophical example of this approach can be represented by Bostrom's paperclip thought experiment. If the overarching goal of a superintelligence is to manufacture paperclips, then, given the extreme power of a superintelligence, one possible and unfortunate consequence of the superintelligence having such an overarching goal could be the reconstruction of all of Earth into paperclip manufacturing facilities (Bostrom, 2003). This approach to risk assessment could be understood as technology- or AI-centred, in that it seeks to foresee what hypothetical iterations of AIs, once they have acquired executive agentic powers, might *do* to humans and to human environments. Given such risk assessments, it is then up to the humans who engineer such technologies to construct them in such a way that undesirable potentials are not actualised.

Much research is also concerned with the potential ethical consequences of artificial intelligence. Frequently explored ethical topics include questions about human responsibility in the context of AI use. Mark Coeckelbergh's exploration of ethical topics relating to artificial intelligence includes analyses of responsibility attribution (Coeckelbergh, 2020b), narrative responsibility and sense-making (Coeckelbergh, 2023), and self-improvement in the age of artificial intelligence (Coeckelbergh, 2022). Hanna Arendt (2022) and Zygmunt Bauman (2008) and many other thinkers have previously problematised the question of responsibility in bureaucratic structures. Questions of responsibility in connection with artificial intelligence have some affinities with problems of responsibility in bureaucratic structures, for, as is argued in the treatise, algorithmic systems can be likened to new forms of bureaucratic structures in which human agents are increasingly taken out of the loop.

Another frequently explored ethical topic concerns so-called bias-reinforcement. Bertrand Hassani, for instance, considers the extent to which machine learning reinforces social biases (Hassani, 2021). The reinforcement or generation of understandings that could be deemed undesirable in important respects (biases) is hardly a problem that arises as a consequence of artificial intelligence. Given all the different points of view that exist among human beings, almost any human institution, including bureaucracies, could be interpreted as reinforcing or generating such understandings in some way. One explanation of how institutions come to have such effects issues from simply holding their human designers responsible. The institutions, on this view,

simply reflect the mindset of their human designers. In similar fashion, analysists could seek to trace any unfortunate biases propagated by algorithmic systems to the humans who programmed and/or trained the systems. However, since many algorithmic systems are opaque to human analysts, it can be extremely difficult in practice, if not theoretically impossible, to discover the reason why a system has made a certain decision.

Then there is research concerned with disparate potential psycho-social effects. Here again we have a precedent in contemporary concern with how interaction with algorithmic technologies in general, such as smartphones, affects processes such as socialisation, learning, and attention span. Sherry Turkle has considered how the use of network technologies and social robots could affect the learning and maintenance of critical social skills. According to her findings, interaction with network technologies and social robots has a tendency, especially among young and old people, and even more especially among lonely people, to render face-to-face interaction and other forms of conventional interaction increasingly burdensome (Turkle, 2011 & 2015).

All of the potential consequences that are typically considered could of course be understood to affect the expression of human agency and autonomy in various respects. Although there are exceptions, many adopt a proactive and positive approach to AI, assuming that AI is dawning upon us whether we like it or not. Problems and risks (existential risks, bias-reinforcement, psycho-social effects) are then identified. It is assumed in advance that the solution should consist in subsequent engineering of technologies so that undesirable consequences are avoided. In this treatise, this approach is represented by Stuart Russell's aim to construct so-called user-friendly AI (Russell, 2020).

To date, some of the research that is closest to the scope of the present treatise – connecting technological developments to changes in worldviews – is published in the journal *Zygon*, which frequently explores topics on the boundary between religion and science. Mohammad Yaqub Chaudhary has written about augmented reality, artificial intelligence, and the re-enchantment of the world (Chaudhary, 2019), and about the artificialisation of mind and the world (Chaudhary, 2020). Abou Farman, using the concept of 'social imaginaries' rather than worldview, has explored the topics of cryonics and yearnings for secular immortality in the age of technoscience (Farman, 2020).

Purpose and research questions

The treatise considers how interaction with existing and future forms of algorithmic technologies could affect the expression of human agency and autonomy and how this, in turn, might affect the way in which humans make sense of them*selves* in a *world*. The risks arising from human–AI interaction that are taken into consideration are tied to changes in the expression of human agency and autonomy and to changes in the content-structure of worldviews. The opportunities and risks that are most pressing to consider on this view are not

those that result from what present and future forms of AIs might do to us, but those that are rooted in what *we* might do or fail to do, or in what *we* might become or fail to become, as a consequence of interacting with algorithm-intense environments.

There is often a fatalistic ring to the dominant positive-proactive approach to challenges related to AI. Provided that we analyse all complex factors with adequate perspicuity, it is assumed that there is a good chance that we will be able to identify and define at least the most dangerous risks and overcome them. We could then confidently carry on with that which is assumed by some to be inevitable – the bringing into being of ever more powerful algorithmic systems – and then deal with lesser challenges piecemeal as they become evident. However, many worry that AI – even so-called user-friendly and safe versions – might provoke radical and unfortunate societal change. In public discourse, we are nevertheless given to understand that, as long as change is not perversely antithetical to our society's current norms or fraught with immediate existential risk, we must accept it in the spirit of past changes and sacrifices made for the sake of progress. Questions of a more general ethical tenor, if they are asked at all, are mostly ignored by deciders and stakeholders. We should nevertheless ask: Given the kind of being that we understand humans to be, ought we to carry on with the AI endeavour? If yes, at what speed and in which domains?

The broader purpose of this treatise is not to give unambiguous answers to such questions, nor to promote one worldview or one set of ideals or values over and above others, but to elucidate certain dynamics, the awareness of which would be useful in the defence and sustenance of many different worldviews, ideals, and values, especially in social contexts in which events are driven to a large extent by the dynamics inherent in technical conditions. Understandings of how humans interact with technologies and understandings of the conceptual frameworks through which humans understand themselves-in-the-world are developed and then tied together into a whole. Within the limits of this treatise, these understandings are developed for the narrower purpose of answering the following research questions: *How will the expression of human agency and autonomy be affected as humans are increasingly embedded in algorithm-intense environments? How are changes in the expression of human agency and autonomy, in turn, likely to affect the viability of dominant worldviews?*

By 'expression' is meant the way in which human agency and autonomy manifest. Theories and conceptions of agency and autonomy are also discussed. On the analytical level, no *natural* or *ideal* state is posited in regard to human agency and autonomy; if it were, the changes considered could be interpreted to *undermine* or *enhance* that which is already considered to be natural or ideal. The purpose of the inquiry is not to argue in favour of a natural or ideal human agency, but rather to explore how its expression could be affected in various contexts. The reader is invited to evaluate changes

normatively from the viewpoint of his or her own notions and ideals. However, evaluations are also made within the scope of the treatise. Interpreted through the lenses of worldviews, changes often *will* be understood to undermine or enhance human agency and autonomy. How given worldviews understand human agency is one of the main concerns of the treatise. It is also conceivable, regardless of worldview, that the scope for expressing human agency and autonomy might become so restricted so as to become inhumane, or that the expression of agency and autonomy might become so socially dysfunctional that it undermines the viability of human society. Such undesirable effects are also considered and evaluated in relation to a third research question: *What noteworthy problems, if any, follow from the changes discussed?*

The questions invite a wide range of answers. We are more specifically concerned with ways in which individual human agents could express agency and autonomy in larger contexts. These contexts could be social (composed of human agents), bureaucratic (still composed of human agents, but with routinisation of agentic functions), or, it is argued here, hybrid (composed of human agents who interact with other types of agentic system). Among the varieties of worldview that still flourish on our planet, it is argued that a modern type of worldview, sustained by a modern mythological narrative, has long exercised hegemony over humans' understandings of themselves and their world, at least in the West. Theoretical reconstructions of a generic modern worldview and a mythological narrative are proposed. The viability of this generic modern way of viewing self-in-the-world, in emerging algorithm-intense environments, is then considered and problematised. More narrowly, the focus is on the modern philosophical anthropology that is central to modern worldviews.

Method

The subject of the treatise requires us to consider a wide range of categories and phenomena. Sometimes we pay attention to disparate details in different categories, at other times to various aspects of larger wholes. The methods used reflect the variety of the categories and phenomena considered. They range from analytic to analogical, extrapolative, and hermeneutic.

Some of the more analytic methods used are described by Anne Thomson. In considering understandings and arguments of other thinkers, attention is paid to alternative explanations (Thomson, 2009, pp. 57–59). Often alternative explanations to the reasoning of one thinker are drawn from other thinkers; at other times the author of the present treatise proposes his own alternative explanations. Given the vast areas of inquiry that overlap the treatise's subject, many different alternative explanations may be relevant in the explanation of one set of phenomena. The purpose of considering alternative explanations is to move beyond simplistic understandings and to seek to comprehend more aspects of the full complexity involved in the phenomena being analysed.

Attention is also paid to analogies (Thomson, 2009, pp. 42–43). One of the authors discussed is of the view that humans have been treated like robots for most of history. The appropriateness and implications of such analogies are often discussed. More importantly, analogies are extensively used to describe the function (and sometimes the meaning) of many of the phenomena under consideration. How learning machines can learn is described in analogy with how we could understand that humans learn. Historically, the conceptual development of artificial intelligence has proceeded as an endeavour to replicate human brain functions in machines. In some respects, artificial intelligence could be understood as a materially instantiated analogy – engineered on the basis of how human cognition is understood by cognitive scientists. In the treatise, the analogy is included not in order to instruct the reader about human psychology, but rather to give the reader a sense of the historical piecemeal development of artificial intelligence, which mirrors common understandings of how the human brain functions. The analogy is also included to make it easy for the reader to grasp what algorithmic systems are able to do and how algorithmic systems might be able to interfere in human-agent arenas. The appropriateness and implications of this analogy can also be discussed, of course.

Key ideas about the future are also informed by a sort of analogous understanding between the past and the future. The bulk of the treatise is concerned with historical processes, historical and contemporary sociological and philosophical work, and other writers' predictions of future scenarios. From one angle, this treatise represents critical discussions of predictions of the future proposed by others. These discussions are informed by patterns that have been observed in historical and contemporary settings. When the author reasons about the future, the reasoning is conducted to a large extent on the assumption of *continuity*, namely, that some of the patterns that have been observed during historical and contemporary industrialisation, bureaucratisation, and automation processes will continue to manifest in future iterations of algorithm-intense environments. Such patterns are *extrapolated* and applied to algorithm-intense contexts. This approach enables us to engage in one type of critical consideration of future events. It is recognised nevertheless that, even if there happens to be continuity in some patterns, discontinuity and rupture could manifest in other patterns. One of the authors who is engaged here argues that an algorithm-intense future is likely to produce contexts in which there is little work-related need for human agents. This, relative to previous instances of technology-caused unemployment, would, if the argument is valid, represent a rupture with previous patterns. If superintelligence did emerge, there might be little or no continuity. The future would then be anyone's guess. Mirroring the justification of alternative explanations above, informed discussions about future possibilities could be conducted on the basis of other assumptions, some of which are commented on in this treatise.

The identification of assumptions is an important step in argument analysis. In relation to an argument's conclusion, unstated assumptions can figure either as assumptions underlying basic reasons or as unstated reasons or conclusions (Thomson, 2009, pp. 23–34). While argument analysis is used in this treatise, much of the constructive work revolves around phenomena that bear a resemblance to assumptions, but that manifest not in arguments but in narratives and worldviews. This brings us to the hermeneutic methodology that is used.

This treatise considers different systemic structures – algorithmic systems, worldviews, societies, etc. – that display synchronic and diachronic dynamics. Anthony Thiselton, in his discussion of hermeneutics, draws our attention to Paul Ricœur's distinction between 'lived time', 'historical time', and 'mythic time'. Lived time is informed in various respects by historical and mythic time. Mythic time typically contains 'a founding event' (Thiselton, 2009, p. 240). In this treatise, a reconstruction of a generic modern narrative is proposed. The reconstruction contains an 'Axial Age', which, if interpreted as a founding event, symbolises humankind's continual rise to maturity and mastery of its environment. Within the hermeneutic frame of analysis that is used, then, mythic and historical time can be understood to frame 'how we experience futurity' by laying out a 'horizon of expectations' (Thiselton, 2009, p. 241).

The reconstruction of a generic narrative of modernity, which is later referred to as the *mythology of modernity*, is no complicated methodological feat. Readers who are familiar with Charles Taylor will recognise it as a typical subtraction or coming-of-age narrative, very similar to the narrative critiqued by Taylor in *A Secular Age*.[1] This grand narrative is not offered as an explanation of historical events, but as a mythical background narrative that, it is argued, has framed – and to a large extent still frames – the meaning-making of the humans who live under its sway. Moreover, it is critically engaged in a way that mirrors Taylor's engagement with this type of narrative. According to Paul Janz, Taylor's purpose is to

> trace certain conditions underlying a fundamental transformation, or shift, through which it was virtually impossible not to believe in God or in transcendent sources, to the present secular age in which – even for religious believers – belief in God is seen as "one option among others," and in which, moreover, "unbelief" has come to have the status of a virtual hegemony in an increasing number and variety of milieux, especially in academic life. (Janz, 2014, p. 46)

[1] See Taylor (2018). A 'subtraction narrative' is an interpretative reading of history that explains the emergence of secular modernity in respect of historical removals of flawed and superstitious understandings from the corpus of human understanding. A 'coming-of-age narrative' represents the modern human as having risen to the occasion and assumed a mature state, with all its responsibilities.

In critical engagement with various phenomena pertaining to modernity, this treatise seeks to trace the conditions underlying fundamental shifts from societies prone to conserve their traditional ways of life to societies that continually aim for technological feats and transformations, and that presently appear to welcome the prospects of embedding human agents in algorithm-intense environments. In engaging with subtraction and coming-of-age narratives, Taylor is sceptical about top-down explanations that assume the pivotal importance of great thinkers in the shaping of historical processes. Taylor instead seeks alternatives that explain key influences from the bottom up, such as changes in liturgical praxis and in the expression of art forms. That which is sought is referred to as 'background pictures' or 'background frameworks', 'by which is meant a pre-theoretical "sense" of the world out of which all our beliefs and conceptions are formed' (Janz, 2014, p. 53). In this treatise it is argued, on the one hand, that the mythology of modernity frames the part of humanity that is under its sway; on the other hand, the mythology itself, it is reasoned, is also shaped from below. Whereas Taylor pays much attention to the influence that changes in popular and institutional practice may have exercised on historical processes, this treatise focuses on the key function of agent-relative affordances in the shaping of the ecology of the whole.

Worldviews and technologies, at first glance, represent quite different sets of phenomena. Nevertheless, at some levels they could be understood in relation to similar notions. Worldviews could be intuitively understood to be composed of sets of assumptions and ideas. They exist, if they exist, *in minds*, or, relationally, *between minds*. Technologies are colloquially imagined as *things*, ranging from individual gadgets to interconnected infrastructures and, today, algorithmic systems. Still, technologies originate as the inventions of minds. Inventions can also be analysed in respect of assumptions and ideas. In the case of inventions that take the form of concrete things, we could posit that assumptions and ideas become 'embodied' for the purpose of achieving concrete objectives in given environments.

Evert Vedung differentiates between the analysis of ideas in view of content and in view of function. The purpose of analysis of content is to examine the formal attributes that are understood in the tradition of philosophy to enable sound reasoning, such as logical validity, consistency, relevance, and the extent to which assumptions about reality are well-grounded. Functional analysis, on the other hand, seeks to understand how people come to embrace the notions that they embrace, and how the notions thus embraced might influence the environment (Vedung, 1977, pp. 15–22).

In the case of worldviews, analysis of content would lead us to focus on the formal attributes that are understood to enable sound reasoning in view of the goals designated by any given worldview. In principle, the same would hold in the case of technologies. The relation of worldview to human beings could be understood in analogy with software to machines. Worldviews and software, on this analogous understanding, designate goals to be pursued. In view,

then, of any goal: Are the assumptions about reality well-grounded? Are the inferences and deductions relevant, consistent, and logically valid? This kind of analysis of technologies must of course be performed by professional teams that include adequately trained programmers and engineers. One question that could be broached within the scope of analysis of content – and one that is repeatedly raised in this treatise, both with respect to technologies and with respect to worldviews – is the extent to which assumptions about reality are well-grounded, *given the complexity of the environment and the goals that are to be pursued*. By means of analysis of content we seek to evaluate the structural integrity of the reasoning involved in the structures being analysed, relative to the environments in which the structures are located.

However, structures such as worldviews and technical systems can be quite viable and influential even as many phenomena relevant to the goals being pursued are ignored, and even while they operate on assumptions that are inaccurate. In considering ideal understandings of human agents, such as human beings as destined to seek control over matter in time and space, it is more the socio-dynamic functions of such an understanding than its formal consistency with factual assumptions that are considered. Likewise, in considering the impact of technologies, it is not so much the veracity of any assumptions about reality that may have been made by its designers that is under close scrutiny, but rather the effects of technologies on the larger environments in which they are located.

Theory

The title of the treatise evokes a vast area of inquiry. Ultimately it encompasses, in one and the same whole, everything from the simplest environment, including the simplest tool or technology, to cognitive frameworks such as worldviews and mythological narratives. In general, theories and concepts are treated and explained during the course of the treatise at each of the stages in which they are introduced. An extensive discussion of worldviews is conducted in part III. Here some of the key concepts are given an initial introduction.

In order to form an initial idea of how the various categories and phenomena being analysed belong together, the reader is invited to consider the familiar notion of ‘home’. Birds have their nests. Bears have their lairs. Does it make sense to say, similarly, that humans have their homes?

Whereas different species of birds build different kinds of nests, and different species of bears construct different kinds of lairs, the homes of humans, one and the same species, are found in an astonishing variety. From makeshift shelters to masterfully crafted igloos, from cheaply produced trailer-homes to high-tech skyscraper-embedded smart homes – the varieties that have existed throughout history are too numerous to list. What is it that motivates humans, in their ‘home-making’, to apply such a variety of conceptual designs using such a variety of materials?

One could suppose that any human could simply will their ideal home into existence. If we were to suppose further that each individual human is unique, then it would come as no surprise that such a variety of homes has been and continues to be instantiated. However, different types of home have proliferated in different types of society, and in different eras. It is reasonable, therefore, to assume that the type of home that proliferates at a given place at a given time will be integral to a given 'way of life'. Ways of life, in turn, are sustained not only by material structures, but also by sets of 'ideal human agents'. Throughout history and across cultures we can find examples of very skilled and peculiar ideal agents, from hunter-gatherers to accountants, from samurai warriors to litigation lawyers, from Saami herdsmen to computer scientists. Each ideal type finds its purpose and meaning not only in a given way of life, but also in a given 'world'. Hence, we could posit that, in some sense, the activity of home-making belongs in the larger category of 'world-making'. In an entire world there can, of course, be a variety of ways of life and homes. In some sense, each ideal conception of home also carries notions about an ideal way of life. Each ideal way of life also carries notions about an ideal world. A world in its entirety can be vast enough to house many different ideals.

How, then, could a concept such as worldview relate to material things such as buildings and technologies? The idea for this treatise sprang from a reading of *Zombies in Western Culture* by John Vervaeke, Christopher Mastropietro and Filip Miscevic. The authors use an understanding of worldview developed by Brian Walsh, and then tie the concept of worldview to the notion of an agent–arena relationship. The implication is that given worldviews will tend to fulfil their function to the extent that they match the arenas in which the agents inhabiting the worldview happen to live their lives. If an arena becomes alien to the worldview that is inhabited, a meaning crisis is likely to result (Vervaeke, Mastropietro and Miscevic, 2017). An arena, in the case of human agents, can be constituted by a variety of elements, including buildings and technologies. In this treatise, the term 'arena' is used to designate environments that elicit very specific behaviours from agents, such as chess boards and tennis courts; larger and less well-defined environments are referred to simply as 'environments'.

An extensive discussion of worldviews is conducted in part III of the treatise. Let us now consider Brian Walsh's rudimentary and functional understanding of worldview. Walsh writes:

> A worldview shapes those who live in its embrace so that they develop certain habits, certain habitual ways of living and relating to each other and the world. And these habits are the stuff of habitation. Worldview-rooted and directed habits shape our places of habitation so that they become home. (Walsh, 2006, pp. 244–5)

Upon being transferred to modern housing, Walsh informs us, the Grassy Narrows First Nation community experienced distress. Walsh connects the dots between their being transferred to modern housing and their being uprooted from an environment experienced as home: 'housing' is not the same as 'home'. Although the Native American community got to inhabit housing of adequate modern standards, they were experiencing homelessness. The environment in which they lived was alien to their worldview. They had been compelled to inhabit structures designed for the practice of a way of life compatible with some alien worldview but not with their own.

Worldviews, on Walsh's understanding, are plausibility structures that answer four questions: Where are we? Who are we? What's wrong? What's the remedy? (Walsh, 2006, pp. 244–5). The environment in which we live, however, is not something that is neutral vis-à-vis a worldview. It communicates worldview: 'buildings, in their very form, symbolize a particular worldview embedded in a particular cultural story or myth' (Walsh, 2006, p. 248). Buildings are tokens of a type of technical construction. *It should follow that other types of technical construction, or even technologies in general, can also communicate worldview. Certain environments, such as algorithm-intense environments, could then be congenial to or incompatible with the sense of home, or the sense of any other vital notion, that emerges from any given worldview*.

Ideally, then, a worldview is that indispensable heuristic frame that should enable us to become attuned to the world we inhabit, shape a home in it, and fashion an ideal way of life. Circumstances can compel us to act in ways that are alien to our own worldview and that may or may not correspond to a way of life that is compatible with another worldview. Environments, both natural and artificial, can be congenial or alien to our worldview. Conversely, our worldview can be compatible or incompatible with the environment in which we find ourselves. On this view, the challenge of artificial intelligence cuts straight into fundamental notions: How we imagine home now and in the future (which, in turn, depends on how we understand human flourishing and our role in our habitats), who we are as those who are to be *at home* in it (humans, post-humans, machines?), and what the threats to home are and what the remedy is.

In also understanding worldviews in analogy with Kuhnian paradigms, the example of Ian Barbour (1996) is followed. Thomas Kuhn, in *The Structure of Scientific Revolutions*, uses the concept of 'paradigm' to explain how, in so-called normal times as opposed to revolutionary times, scientific communities progress in their work. A paradigm, according to Kuhn, stands for 'universally recognized scientific achievements that for a time provide model problems and solutions for a community of practitioners' (Kuhn, 1996, p. x). Scientific paradigms follow a sort of life cycle: they are formed, mature, and eventually reach the end of their usefulness, whereupon a revolutionary phase provides the space for the emergence of new cycles of new paradigms. In following the example of Barbour, it is posited that worldviews, in analogy with

scientific paradigms, develop in an organically cumulative way. In normal times, we use sets of mostly unexamined assumptions in order to conceptualise and add new understandings and knowledge of the world in which we live – that is, to add new pieces to our epistemological and normative puzzles. In times of crisis, one option – not the only option, nor necessarily the best option – is to examine worldview-embedded assumptions. This may be the first step in a reformation or an outright rejection (paradigm shift) of a worldview.

Were the misfortunes of the Grassy Narrows First Nation community primarily the result of a narrative collapse of their worldview, or the result of an environmental collapse caused by the subordination of their agency by a foreign power? We should be foolish to give any decisive answer before having examined those events in more depth. A good case could nevertheless be made that the worldviews of other peoples who have suffered similar subordinations, such as the Hebrews, appear to have been reinforced rather than having suffered narrative collapse. In our own times, many who inhabit societies that have not been subdued by foreign conquerors experience similar sensations – either dissolution, like the Grassy Narrows First Nation community reportedly did, or defiance, like the biblical Hebrews – vis-à-vis environments supposedly issuing from the native worldviews of their own societies. Explanations as to why such phenomena should occur are sought.

To summarise, the paradigmatic prism of worldview allows us to understand worldviews as something that can be either reformed or, on rare occasions, abandoned. Walsh helps us to understand just how difficult this can be in times of crisis. A worldview is not an abstract or neutral instrument that we can choose disinterestedly to use or disuse in our pursuit of higher ends. It is inhabited by us. It designates those higher ends. It seeks compatibility or communion with the features of the environment that can be made into home. It is spoken to by the environment. The narrative collapse of a worldview implies the loss of home.

If environments, technologies, and AI can affect worldview content, then how does such influence proceed? It is argued that it is through the practices of human agents. Between worldview and environment, we can posit general ways of life. In more narrow contexts, we can posit ‘ways of work’, ‘ways of amusement’, and so on. These categories represent ‘embodied practices’ that, to various extents, will conform to worldview ideals and/or to the requirements arising from more or less transitory environmental circumstances. As environmental circumstances change, new requirements pertaining, for instance, to ways of work can be revealed. If environmental circumstances change rapidly and radically, it is likely that requirements will also change rapidly and that change will be experienced as disruptive.

One of the key changes that is considered in this treatise concerns human collaborative contexts. It is assumed that daily practices, in and out of work, play the key middle role between circumstantial changes and worldview adaptation. In order to explore these dynamics, the concept of ‘multi-agent

system' is borrowed from the discipline of computer science. The concept of 'affordances', typically defined as 'an environment's or object's agent-relative possibilities for action', is also extensively used. The changing function of human agents in collaborative contexts – that is, multi-agent systems – is explored, from 'pure' contexts composed entirely of human agents to 'hybrid' contexts composed of human and other types of agent. As collaborative contexts for human agents become increasingly algorithm-intense, the roles and statuses available for humans change both in and out of work. Multi-agent systems and affordances are treated in part I.

'Modernity' is a multifaceted and flexible concept. In this treatise it is discussed first in part II, where the understanding being discussed applies to so-called late modern conditions. The concept is then treated in more depth in part III, where a generic modern narrative is reconstructed and the characteristics of modern worldviews are specified. The concept typically applies to narratives, worldviews, and conditions – all of which are treated as synchronic-diachronic structures that tend to undergo change in time and space. In this introduction, we briefly consider how modernity relates to human self-understanding.

A central worldview notion is the understanding of the human-being-in-the-world. At the centre of the inquiry undertaken here stands the human person understood as an agent – someone who acts in environments and who is able to use elements of environments as means in order to accomplish ends. Technologies can be understood as 'constructed means' that, once constructed, become part of environments in which humans can use them as means. From the point of view of modern worldviews, the human agent, when its higher potentials are developed, is assumed to be the supreme agent in this world. This, it is argued, is a key assumption in the philosophical anthropology of modernity. Frederick Olafson defines philosophical anthropology as the 'discipline within philosophy that seeks to unify the several empirical investigations of human nature in an effort to understand individuals as both creatures of their environment and creators of their own values' (Olafson, 1998). During the course of the treatise we examine how algorithmic technologies challenge core conceptions of the philosophical anthropology of modernity.

Immanuel Kant's categorical imperative is at the heart of Enlightenment culture, which, it could be posited, constitutes a key event in the development of modernity. In its second formulation we are enjoined never to use humanity merely as means to ends: 'Act in such a way as to treat humanity, whether in your own person or in that of anyone else, always as an end and never merely as a means' (Kant, 2017, p. 29). The imperative resonates in the Christian culture in which it was formulated. However, in the pre-Christian Greek and Roman pagan cultures it appears to have been perfectly natural to use at least unfree humans exclusively as means to ends. Kantian ethics could be understood as one of the pillars of Enlightenment humanism, enjoining human beings to assume a mature and responsible and eminently reasonable stance vis-

à-vis one another and the world. In that which concerns the non-human elements of environments, it would still be permissible to use them exclusively as means, provided that actions do not contravene any of the other formulations of Kant's categorical imperative. Indeed, as Charles Taylor describes the matter, another pillar of Enlightenment thinking, René Descartes, has been consequential in laying the foundations for the notion that the very purpose of matter in time and space is to be used as means for human ends (Taylor, 2018, pp. 130–136).

A key characteristic of a modern philosophical anthropology, it is posited in part III, is the notion, developed by Charles Taylor, of a 'buffered self'. If a 'porous self' signifies a self that is at the mercy constant environmental influences that originate from mysterious, opaque, and impenetrable forces, a buffered self describes the self that emerges in modernity. In the version most distanced from its porous predecessor, the buffered self can be described in cartesian dualistic terms – as a mind controlling matter in time and space. To the buffered self, the material environment is conceived as something that can be rendered increasingly transparent to scientific analysis and controlled by means of rational and technical intervention.

Nevertheless, we seem to be entering a new era. As increasingly autonomous algorithmic systems are being designed and implemented, technologies – means – can themselves be understood to acquire various degrees of agency. This implies that, in some instances, human beings are becoming constituent parts of the environments of algorithmic systems. Given the conceptual apparatus that is used in the treatise, this is neither a conclusion to be reached nor a prediction for the future; it is a description of what is happening. We see that humans can already be used as means by the technological 'means' that humans have constructed in order to accomplish overarching ends. In some instances, algorithmic systems, without human input, can themselves infer the instrumental usefulness of subgoals in relation to overarching goals, where the subgoals may imply manipulation of human beings. In such instances, human agents, who are able to consider notions such as Kant's categorical imperative, have simply been removed from the decision-making process.

The new era dawning upon us, it is argued, necessitates a re-evaluation or re-imagining of the modern worldviews that contribute to producing algorithm-intense conditions, and/or of humanity's ongoing technological reconstruction of its environments, and/or of the human agent.

Structure and material

The treatise is composed of three parts.

In part I, the reader is introduced first to basic explanations of how algorithmic systems can accomplish goals. The concepts of 'agent' and 'agency' are introduced and problematised. Three paradigmatic cases for human agents

in algorithm-intense environments are then presented and discussed. Finally, the key concepts of 'multi-agent systems' and 'affordances' are introduced.

The explanations of how algorithmic systems can accomplish goals are primarily drawn from works by computer scientists Melanie Mitchell (2020) and Stuart Russell (2020) that are intended for lay readers. These explanations should help readers who are unfamiliar with the subject to form a basic understanding of how AI functions. They also serve an additional purpose in that they represent the workings of algorithmic systems in analogy with simplified understandings of how humans learn, reason, and act. The analogous approach should enable us to grasp better how even currently existing algorithmic systems – supposedly far below the capacity of general intelligence – can interact and interfere with human agents, as humans and algorithmic systems are increasingly sharing the same arenas and environments.

The following paradigmatic cases for humans in future algorithm-intense environments are presented and then problematised further: a world with less work for humans; human–AI partnership; and increasing cognitive demands on humans. These cases are not meant to be understood as mutually exclusive predictions of the future, but rather as tendencies that, to varying degrees, may be likely to develop in algorithm-intense environments. Many thinkers have considered these or similar developments. Here, the case for a world with less work for humans is drawn from a book by Daniel Susskind (2020); the case for human–AI partnership is drawn from a book by Henry Kissinger, Eric Schmidt, and Daniel Huttenlocher (2021); and the case for increased cognitive demands on humans is drawn from work by Norbert Wiener (1990, 1999).

The final section of part I introduces two concepts that fulfil key explanatory functions in the rest of the treatise: multi-agent systems and affordances. In the presentation of multi-agent systems, which draws on a chapter written by Catrin Misselhorn (2015), we acquire conceptual tools that enable us to better understand the dynamics between human agents and phenomena that could be categorised as 'artificial agents'. 'Agency' and 'agent' may refer to the properties of any discrete acting entity, such as a person. The concept of multi-agent system will help us to understand phenomena that could be described in terms of transpersonal agency and transpersonal agents – phenomena that involve non-personal systemic entities that, in various respects, behave as if they were agents and/or tend to be perceived as agents. Here we also consider how the concept of 'autonomy' relates to the concept of 'agency'. Finally, 'affordance' is introduced in order to enable us to grasp the key agent-relative environmental mechanism by means of which the desires or goal-directedness of agents tend to have repercussions in the world that extends beyond the agents. This discussion draws on the pioneering work of James Gibson (1979) and on an article by Ann Taves, Egil Asprem, and Elliott Ihm (2018).

In part II, human agents are considered under various social and technical conditions. First, we become familiar with Shoshanna Zuboff's critical

evaluation of how algorithmic systems today are commonly applied to social arenas (Zuboff, 2019). Although her descriptions are generally maintained as valid, some of her explanations are further problematised. Drawing insights from various critical perspectives on modernity and technology, such as those of Ivan Illich (2021), Jacques Ellul (1990), and Zygmunt Bauman (2008), an understanding of human–technology interaction is formed. It is argued, with the aid of the concept of affordances, that material circumstances (environments, and sub-components of environments, such as technologies), given the bio-psychological or technical constitution of any given agent (human or non-human), and, in the case of human agents, given any particular worldview, will tend to elicit different kinds of response from different types of agent. It is argued that, although algorithmic technologies do indeed introduce some novel features in contexts of human–technology interaction, current forms of algorithmic technologies could also be understood in continuity with technical structures that preceded them.

Throughout part II, the paradigmatic cases introduced in part I are problematised relative to the so-called surveillance capitalist structures described by Zuboff, and relative to the critiques advanced by Illich, Ellul, and Bauman: a world with less work for humans, with human–AI partnership, and with increased cognitive demands on human agents could have many different implications for humans located on different levels of technical hierarchies.

In part III the prospects for modern worldviews in algorithm-intense environments are considered. First, the general notion of modernity is discussed and problematised. Drawing on thinkers such as Romano Guardini (2019), Zygmunt Bauman (2007), and Steve Pinker (2018), it is argued that modernity can exist in different versions. Guardini gives us an idea of how a historical ideal modernity could be conceived. We, however, live under late modern conditions or borderline conditions, in which concrete life conditions correspond less and less with the original modern ideals.

Ann Taves, Egil Asprem, and Elliott Ihm then help us to acquire a formal structure for worldview content, a structure that, through the concept of affordances, is also capable of integrating concrete environmental circumstances (Taves, Asprem and Ihm, 2018). The characteristics that would render a worldview modern are discussed and specified. Modern worldviews, it is argued, are animated by a modern mythological narrative. This narrative is inferred. The philosophical anthropology of modernity is then pieced together, using notions and insights from a disparate group of thinkers: James Frazer (1996), Karl Jaspers (1949), and Charles Taylor (2018). The philosophical anthropology of modernity evokes a human agent that, buffered from its environment, by means of the use of innate reason and by resorting to the stored knowledge of a modern, reasonable, and scientifically informed culture, should be in a supreme position among agents: able to control and dominate its environments.

This, it is then argued, will likely correspond less and less to the actual status of human agents in algorithm-intense environments. Although the actual status of human agents may vary in accordance with where they are located in technical hierarchies, and although phenomenological experience may often be at odds with 'actual status', algorithm-intense environments imply that something fundamental is about to change.

In the final section, we explore some possible adaptive measures that could contribute to restoring a sort of equilibrium between worldviews and environments. Modern worldviews, and the philosophical anthropology lodged within them, could adapt and change into something that we no longer recognise as modern. The human agent could be subject to reconstruction for the purpose of *becoming* able to master increasingly complex environments. The reconstructive endeavours rendering human life-environments algorithm-intense could slow down or even stop.

In the conclusion, the analyses, reasoning, and findings of all three parts are reviewed, discussed, and summarised.

Part I: From simple chess-playing systems to the role of humans *qua* agents in hybrid multi-agent systems

What is meant by artificial intelligence? Computer scientists, cognitive scientists, and philosophers can mean different things when they use the term. Their differences largely stem from how they understand 'intelligence'. Many argue that there is not yet any artificial intelligence, that the algorithmic technologies of today are not *really* intelligent. That is to say that the bar for when something becomes intelligent is set so high that no artificial system yet passes the test. The definition of artificial general intelligence given in the introduction represents one possible intelligence bar. The many instances of so-called narrow artificial intelligences that are already in full operation should not, according to some, be referred to as 'intelligent' at all.

Questions of what 'real' intelligence is or what might be required for the attainment of 'genuine' intelligence are not treated here. Instead we consider procedures by means of which algorithmic technologies, by design and subsequent training and/or interaction with environments, could be made to acquire properties that enable them to perform more and more autonomously and more and more efficiently in more and more domains. We look at how it is that algorithmic technologies today are made to replicate and go beyond human behavioural achievements, and, in turn, consider what this implies for humans who are embedded in increasingly algorithm-intense environments.

We consider three scenarios for humans in algorithm-intense environments that are discussed by prominent thinkers: *a world with less work for humans*; *human–AI partnership*; and *increasing cognitive demands on humans*. These scenarios are then used as paradigmatic cases in subsequent discussions and, in parts II and III, are further problematised. Part I begins with an explanation of how algorithmic technologies can be made to learn how to pursue various ends conventionally pursued by humans. It is followed by an account of how the concepts of 'agent' and 'agency' are used, and a discussion of how the role of humans *qua* agents in hybrid multi-agent systems is considered.

How algorithmic technologies can be made to learn to pursue ends

If by 'intelligence' we mean something like an 'ability to engage in instrumental reasoning', then even the simplest chess program can be understood to

possess intelligence. By 'instrumental reasoning' is meant the ability to reason hypothetically in order to achieve a goal.[2] Instrumental reasoning is integral to algorithmic technologies. Many of the more advanced algorithmic systems of today display an additional capacity that is commonly associated with intelligence: learning behaviour. In these respects, at least, contrary to what some argue, artificial intelligence does not seem to be a misnomer for current algorithmic technologies.

Here we primarily consider types of behaviour and ability that humans and algorithmic systems can have in common. Later, we also consider features that appear, so far, to be unique to humans. It is fitting, however, to begin with similarities. It is, it is argued, similarities that make the challenge to human identity – and many other challenges – acute, even at this early date.

Many argue that we have to wait for 'real' artificial intelligence, such as AGI and ASI, in order for truly revolutionary effects to kick in. Nick Bostrom (2014) and Max Tegmark (2018) worry that humans will lose control over fully autonomous artificial systems that attain intelligence levels that exceed human capabilities by far. Such systems, Bostrom and Tegmark argue, will represent an existential threat to humankind. There is no doubt that such developments would represent something like a turning point in or pinnacle of a revolutionary process. The focus of this treatise is nevertheless more on challenges presented by already existing systems and by the incremental improvement of such systems. These challenges involve the risk of loss of a different *sense* of control, namely, a gradually accelerating loss of the ability to rationally, in view of current norms and values, engineer and manage our technical environments. Such loss could be actualised long before the onset of AGI or ASI. The result of such a loss would not necessarily imply the end of humanity. It could result in radically novel life-environments that are conducive to radically novel ways of life structured in alignment with norms and values that most current readers would consider alien. In part III we consider more closely how ways of life and norms and values might adapt to new circumstances. From the point of view of many of the stakeholders who currently facilitate the development of algorithmic technologies, the adaptations considered in part III, if they were to be actualised, would no doubt be understood as unintended and undesired effects. The emphasis on an *already ongoing and continuous revolutionary process* is in line with the analyses that are undertaken in part II and III.

Let us now examine how algorithmic systems can be made to perform and accomplish tasks. The workings of such systems can be understood in analogy with how humans learn to perform activities and accomplish tasks. The

[2] Even a simple chess program is typically not restricted to issuing reactive commands – e.g., if *x*, then *y*. It is able to do so-called lookahead searches (which are treated in the pages that follow): it can posit *y*, and a number of possible countermoves to *y*, and consider – i.e., engage in instrumental reasoning – what the best countermoves to the countermoves might be in view of realising the goal of victory.

analogous approach is used partly as a pedagogical device in order to facilitate the understanding of algorithmic systems, and partly in order to direct attention to behavioural similarities between humans and algorithmic systems. Behavioural similarities do not of course entail similar ontological status. The cognitive-behavioural model used to describe human behaviour and learning is in common use in AI discourse, and it will also be familiar to readers of psychologist Daniel Kahneman.[3]

Admittedly, the human creature in its entirety is much more complex than, say, a chess-playing machine. But if we wish to teach a human how to play the game of chess well, we will have to abstract some very narrow relevant parts from the integrated whole. We must leave aside many factors that might be relevant to playing the game well – sleep patterns, hormonal levels, etc. – and instead focus on the rules of the game and, for instance, on how a decision tree can be applied to the game. In a similar vein, a cognitive scientist who wishes to understand how proficient chess players can make intuitive good moves may want to explore what happens on a neural level when the game is played, without at the same time having to consider every other aspect of the human being that might be relevant to the performance. It is in this sense that, in important respects, algorithmic technologies can be understood in analogy with human 'modes of doing things', that is, modes abstracted from the integrated agency of the whole human.

One purpose of proceeding by means of analogies is to make it easy to understand, on a non-technical level, how algorithmic technologies can be made to accomplish more and more complex objectives. The analogies can also serve to accentuate the challenge to human identity. In the introduction the challenge was considered from the point of view of an introspective human observer: If AI can do what we do, then who are we? We can turn it around: If AI systems merely correspond to something like human functions abstracted from the whole, then what are they? And if, one day, AI systems could be integrated into something like a fully capable and autonomous agency, then what would it be like? These are ontological questions. They are raised here as a means of hinting at the ambiguity of the new conditions that are likely to emerge in algorithm-intense environments – conditions in which these kinds of questions must be asked, but where clear answers are unlikely to be readily forthcoming.

Let us now proceed to consider some of the ways in which humans and machines could make decisions, learn, self-adjust, and perceive. For more thorough and technical explanations of algorithmic technologies, the reader is referred to the introductory works of Melanie Mitchell (2020) and Stuart Russell (2020).

[3] Kahneman's 'System 1' loosely corresponds to the mode here characterised as 'intuitive', and 'System 2' to the mode here characterised as 'deliberative' (Kahneman, 2013).

Algorithmic production of decisions or outputs

Algorithmic technologies that are referred to as AI can be understood as belonging to two general types. These two types mirror two modes of human decision-making. If we are asked to explain a decision that has been carefully deliberated, as, for instance, a move in a game of chess, our response might include an account of the deliberation. If we are proficient, we could add an account of a tactic for the context in which the move is made. We could even include an account of a general strategy for the game. Such deliberations, tactics, and strategies belong under the general category of methods that can be consciously deployed in order to achieve objectives – in this case, the overarching objective is to win a game. In well-defined games such as chess, there is logical rigour to how such methods can be applied.

A human player's practical use of such methods can be illustrated by how a decision tree is applied in combination with lookahead searches: The player visualises how the pieces will be distributed after prospective moves, then the distribution after the opponent's most likely countermove, and seeks to apply the heuristics inherent in his or her decision tree. Initially, we could imagine that the decision tree applied by a novice player is rather poor in heuristics, including only the rules of the game and the objective of winning. If these simple instructions could be used without any limitations, then there would perhaps be little need to study tactics or strategies in the game of chess: the best decisions, it would seem, could then simply be calculated, or logically deduced. But if one attempts to look ahead three or four moves in a game of chess and to apply one's rudimentary decision tree, one soon reaches the limits of one's capacity. Cognitive limitations can be compensated for in part by learning the meaning – the advantages and disadvantages – of various board positions and by studying winning tactics and strategies. The game of chess is then turned into an *art* of winning the game. There are rules and an objective, but also many defensive and offensive sub-objectives of tactical and/or strategic significance.

According to Mitchell, many of the most famous pioneers in AI research, such as Herbert Simon, Allen Newell, and others, endeavoured to construct intelligent systems modelled on formal and logical human decision-making (Mitchell, 2020, pp. 9–12). Artificial systems for decision-making were easiest to develop and apply in environments that were closed off from the multiple and varied complexities of the world-as-a-whole, in arenas in which rules and paths to goal-achievement could be clearly defined – arenas such as the chess-board. This is true to this day, which partially explains why algorithmic systems tend to excel in some areas in which it is difficult for humans to attain mastery, such as in the game of chess, while algorithmic systems still struggle to attain passable levels of performance in the achievements of many other tasks that humans routinely accomplish.

Consider a number of possible arenas, from noughts and crosses to driving a car on city and country roads. When rules and paths to victory are clearly defined in a limited and discrete arena, as in chess, in theory there may seem – at least for someone who is not a mathematician – to be no reason why an extraordinarily capable player should not be able to visualise all possible board constellations fifty or sixty moves ahead and, based on this, logically and infallibly deduce winning moves. In practice this decision-making method is prohibitively cost-intensive. In chess, owing to the number of possibilities that must be considered, it is more demanding than in noughts and crosses; in the game of Go it is more demanding still. Even if we admit that one *can* play chess or Go using this naïve lookahead procedure exclusively (it is unlikely to lead us to victory against an opponent above the level of beginner), this method for decision-making *must* be complemented by something else when more complex tasks, such as driving vehicles, are to be accomplished. We cannot simply instruct the self-driving car to visualise all possible states of the road one second ahead, two seconds ahead, etc., and then deduce the best action to take. In some important respects, even though many more humans are able to learn to drive well than to play chess well, driving is a far more complex activity.

Car-driving and board-gaming belong to two different domains, each defined by different spaces of action. The agent faced with the task of winning a game of chess will operate in a finite and discrete space of action with access to complete information, whereas the car-driver will operate in a continuous space of action with access to incomplete information in environments that are much less predictable. When we routinely perform continuous activities such as driving and walking, we appear to be functioning in a mode that can be described as *intuitive* in comparison with a more consciously deliberative and calculating mode. In the intuitive mode, the underlying cognitive states responsible for a decision or a course of action are not consciously deliberated. The more skilful a driver is, the less conscious thought is required for the activity of driving. The appropriate decisions or courses of actions are not reasoned, calculated, or deduced, but intuited. Intuitive and instantaneous decisions or courses of actions, to the extent that they are adequate to the tasks being undertaken, are likely to be the result of prior experience and training.[4] The experienced driver *knows* where to pay extra attention. We can infer that experienced drivers have acquired added layers of heuristics which enable them to drive skilfully based primarily on intuition. This is also true for the experienced chess player. In order to develop intuitive knowledge, both car drivers and chess players must practise. ‘Learnt’ and ‘acquired’ are the key

[4] In the case of animal agents, adequate actions can also be performed consequent to innate instinct. In human agents, innate instinct is perhaps the dominant mode for adequate actions in the case of the infant. The more human agents mature, the more adequate behaviour will be the result of learning and training.

concepts here, as opposed to 'reasoned', 'calculated', or 'deduced'. The more that learners practise, the more attuned their deliberative and intuitive heuristics, with reference to goals pursued, become. Deliberative heuristics can be schematised as decision-trees; intuitive heuristics are opaque to the conscious and deliberative faculty. Deliberative heuristics are cost-intensive, whereas intuitive heuristics tend to be applied instantaneously and effortlessly. The experienced chess player, then, does not need to rely solely on cognitively demanding deliberations: intuitive heuristics more and more attuned to the requirements of the game can be used in order to solve many problems; and cognitively demanding deliberations can then be deployed to the more important situations where potential gains outweigh costs.

The alternative to any type of heuristics is brute force. In the absence of heuristics, one has no alternative but to go through and check all possible options, and then consider the consequences of engaging each option, and so on, to the extent that it is possible. A heuristically poor chess player, whether it is a human or a computer, that only knows the rules of the game will be limited to lookahead searches that will be synonymous with brute force searches. The acquisition of useful heuristics, on the other hand, will enable the chess player to use cognitive resources with greater economy.

Deliberative methods are demanding not only for humans, but also for machines. Just as in the case with humans, machines can be made to proceed by two principal routes. A second approach to AI kicks in where the deliberative approach stumbles. In analogy with how humans can learn from experience, we can think of this second approach as an 'intuitive mode based on inferences from experience'.[5]

Melanie Mitchell distinguishes between 'symbolic' and 'subsymbolic' AI (Mitchell, 2020, pp. 9–24).The type of method that has been categorised as deliberative can be instantiated in machines by means of symbolic programming. Symbolic programming refers to instructions that are transparent and comprehensible to programmers. Human knowledge – or heuristics – about how to solve various problems can, then, be programmed into machines. Note, however, that in this case the heuristics are based on prior human experience, not on prior 'machine experience'. When systems programmed in this manner make decisions, programmers can usually fairly easily find out why decisions have been made. If different decisions are desired, programmers can fairly easily modify the instructions.

Pure symbolic systems typically rely on human expert knowledge in particular domains. Such knowledge, as for instance how to play chess, is then

[5] One could object that this alternative machine mode is just as much a matter of brute force as the one with which it is compared. Whereas the intuitive mode as characterised in humans is the result of some sort of *human* learning, the machine mode about to be explained typically relies on the systematic processing – brute force – of large data sets. Nevertheless, here we are primarily interested in the results of the process, which, as it turns out, add up to machines that are able to behave in ways that can be compared with the intuitive behaviour of humans.

fed into the system. According to Mitchell, by the mid-1980s researchers who endeavoured to construct robust symbolic systems had learnt something about both artificial systems and themselves. The systems turned out:

> to be *brittle*: that is, error-prone and often unable to generalize or adapt when presented with new situations. In analysing the limitations of these systems, researchers were discovering how much the human experts writing the rules actually rely on subconscious knowledge – what you and I might call common sense – in order to act intelligently. (Mitchell, 2020, p. 33)

Popular terms such as 'neural nets' and 'deep learning' are used to describe the second approach to AI. Mitchell refers to this approach as 'subsymbolic', in analogy with how symbolic language (English, propositional logic) emerges from the subsymbolic structure of the human brain. Subsymbolic systems, pioneered by the psychologist Frank Rosenblatt, mimic not the methods that humans deliberately and consciously use (symbolic), but the neural structures of the human brain (subsymbolic). 'Neural networks' and 'connectionist' are other common terms for this approach to AI. Developments and improvements of subsymbolic systems have played an important part in enabling current so-called learning machines to equal and even surpass human performance in more and more arenas.

A learning machine is a machine that can learn. In contrast to machines that do not learn, learning machines can increase performance through experience. On the basis of having played many games, a learning machine can become able to make good instantaneous chess moves in various positions. Here, decisions are made and objectives pursued not only on the basis of how the systems have been programmed, but also on the basis of how they have been trained.

Both symbolic and subsymbolic systems can be made to learn from data. Symbolic systems are fairly transparent, whereas subsymbolic systems tend to be opaque. This means that, if a symbolic system does not work as intended, it is usually fairly easy to locate the source of the error and to find a remedy. In subsymbolic systems, it can be very difficult, and even next to impossible, to find out how a decision is reached. Symbolic systems are cost-intensive. Subsymbolic systems can generally reach high levels of performance at far lower cost. In areas where transparency is important, symbolic systems will be preferred to the extent that the cost of using them is not too high. Moreover, the two approaches can be combined in overarching hybrid systems.

Error or undesirable performance in learning machines can be caused by how they are programmed, by how they have been trained, and by the data used for training. The hardware of the machines is, of course, also a relevant factor. Last, the interaction between the machine and the arena or environment in which it operates must also be considered. If there can be arenas that, like the game of chess, are eminently suitable for the application of learning

machines, then there can also be arenas and environments that – given the present state of algorithmic technologies, and given that we have an interest in the results – are much less suitable, or not suitable at all, for the application of learning machines.

Learning

When a child is born it does not speak for a long time. The newborn comes equipped with perceptive organs and, according to Alexa Modrell and Prasanna Tadi, a number of reflexes that include sucking and grasping reflexes (Modrell and Tadi, 2023). It also has an ability to communicate by means of crying. Through these initial means of engagement it is able to begin its practice of perceiving, reacting, and acting. As time goes by, out of the unordered chaos that is perceived initially, the child *somehow* learns to form useful categories: safe is distinguished from harmful, familiar from unfamiliar, sweet from sour, and so on. For more precise accounts of just how these patterns develop, some expertise in developmental psychology can be consulted.

A typical adult, when presented with a familiar situation such as waiting for and entering a bus, no longer needs to resort to sucking and grasping reflexes and crying in order to navigate the environment. The adult enters the bus equipped with multiple sets of cognitive heuristics that predispose him or her to interpret and respond to the situation in a pre-reflective and instantaneous way. The heuristics, presumably, are the results of previous experience, training, and education. As long as the heuristics are attuned to environmental circumstances, many everyday activities can be performed in this pre-reflective manner.

The adult has also learnt to home in on features in the environment that have proven relevant to the realisation of goals. Based on findings in neurodynamics and neuroanatomy, John Vervaeke and Leonardo Ferraro argue that relevance realisation fulfils a key function in general intelligence (Vervaeke and Ferraro, 2013, pp. 57–68). However, if the adult has learnt to focus on features that have proven to be relevant to the realisation of various goals, a consequence of such perceptual fine-tuning could be that the adult would likely be unable to perceive many aspects of reality that would still be accessible to the child. If the newborn has not yet developed heuristics that make useful sense of the world, the risk that the highly trained adult runs is to make *too much narrow sense* of the world – that is, to miss out on aspects of reality that have not been deemed to be relevant (but that might in fact be) to the realisation of important goals. The picture becomes even more complex if we admit that there may also be goals that are worth pursuing that we might not yet have deemed worthy of pursuit, and that we might also be better off not pursuing some of the goals that we have chosen and/or become accustomed to pursue.

How do we get to that highly ordered adult stage? Terms commonly used in contexts of machine learning are used here in order to give a tentative and partial answer to the question.[6] If we revert to the newborn, pre-equipped with perceptive organs and reflexes, the infant will be able to navigate its environment on its own to an extent, but, critically, under adult supervision. By touching all kinds of things, by putting objects into its mouth, the infant will learn what causes pain, what is soft, what tastes sweet, and so on. By means of a kind of reinforcement learning, the infant receives important feedback from the environment; on the basis of such feedback it can then, somehow, begin to form categories – safe and dangerous, and so on – in its mind. Adults can help by setting up safe and pedagogical conditions for the learning and by supplying critical feedback and information. Much of the actual learning will occur outside any formal teaching context set up by adults – that is, by means of explorative reinforcement learning and unsupervised formation of categories. Adults assist in suppling names to categories, such as the words 'safe' and 'dangerous'. In some instances, adults will actively teach the infant that a certain type of object belongs to a given category. Electrical sockets and sharp knives are too dangerous for the infant to explore on its own. So it will be taught that they belong to the category 'dangerous'.

Learning machines, in some respects, can be understood in analogy with at least some aspects of human learning. An algorithm in a learning machine is a behavioural instruction. 'Algorithm', in Mitchell's words, 'refers to a "recipe" of steps a computer can take in order to solve a particular problem' (Mitchell, 2020, p. 19). The newborn, supposedly, comes with 'behavioural instructions' too: to seek warmth, comfort and satiation; as previously stated, such behavioural instructions also include a number of reflexes – that is, recipes of steps that the infant can take in order to 'solve problems'. Integrated with its perceptive apparatus, sets of genetically coded instructions enable the infant to engage with its environment and, with the assistance of well-disposed adults, to achieve satiation in respect of being safe, warm, and fed. If something is amiss with any of the critical functions, the infant's engagement with its environment will be disturbed. We can understand algorithmic instructions in learning machines in analogy with the innate drives and engagement modes that motivate and enable the infant to engage with the environment.

The child, supposedly, gradually produces an inner picture of his or her world, or a world- and self-model, in which notions of purposeful agency can be imagined, and then tested in the real world. In humans, it is posited, the category of meaning is integrated into the complementary structure of worldview. Whereas worldviews, as conceptualised in the introduction,

[6] The purpose is not to reduce the complexity of the human infant, child, or adult to the function of machine learning. If we consult expertise in developmental psychology, we get fuller and more nuanced accounts of how humans learn. Human uniqueness notwithstanding, one of the points here is that, on some levels, there *are* similarities.

appear to be unique to humans, it is plausible to assume that world- and self-models are constructed by all advanced animal organisms.[7] An *adequate* world- and self-model enables an organism to engage with its environment in a fairly predictable and successful way. When algorithmic technologies construe simulations of their environments, they, too, are constructing something like world- and self-models. We could perhaps think more aptly of these as environment models. As learning machines engage with environments, their environment models can be fine-tuned, so that the configuration of the learning machine becomes more *adequate* for the purpose of achieving its goals.

Contemporary algorithmic setups are able to give direction to learning machines, to make them purposeful or agent-like in that they can be trained to play chess and drive vehicles, like humans do when they perform these functions, and to translate texts from one language to another, and – again mirroring the behaviour of human thinkers – even to produce answers with the appearance of originality when presented with questions. How does machine learning to such varied effects proceed? This is a complex technical subject, and it goes without saying that it cannot be done full justice here. Again, the reader who is interested to learn in greater depth about these things is advised to consult Mitchell (2020) or any other introductory work. Here, some key features that are especially relevant in considerations of potential algorithmic interference in spheres that concern human agency and autonomy are briefly described. These features are relevant in that they demonstrate how human agent processes can be mimicked without necessarily mimicking the ontological status of human beings. Given that ideal agent types represent a key notion in the category of worldviews, it is argued that the mimicking or, rather, emulation of human agent-like behaviour by other types of artificial agents is likely to challenge core structures in worldviews.[8]

The simplest version of a subsymbolic system is the perceptron, which is composed of a unit that produces outputs based on inputs. Perceptrons and compositions thereof, called 'neural nets', are composed of 'weights'. Contemporary neural nets have many interior units and millions of weights between the input and the output. A perceptron with an added 'layer' of 'simulated neurons' is called 'a multilayer neural network' (Mitchell, 2020, p. 23). Layers between output and input are often called (for some mysterious reason) 'hidden layers'; Mitchell points out that this simply means non-output units, and that a better name might be 'interior unit'. It is the number of interior units that defines the 'depth' of a system.

[7] More detailed understandings of world- and self-models and of the structure of worldviews are presented in part III.

[8] 'Mimicked' is used here in order to suggest a likely perception of the phenomena from the point of view of a human observer. Humans can learn by mimicking other humans. But 'mimicked' is not a good term for describing what occurs in learning machines. Learning machines do not mimic humans. Rather, they genuinely learn in a way that is proper to algorithmic systems.

Stuart Russell understands intelligence in terms of perception, objectives, and actions, so that 'an entity is intelligent to the extent that what it does is likely to achieve what it wants, given what it has perceived' (Russell, 2020, p. 14). A learning machine can be instructed by algorithms, on the basis of certain sequences of inputs (perception), and with a repertoire of certain steps (actions), to pursue certain ends (objectives). We have discussed chess and its well-defined arena and rules. In the task of driving a car, both arena and rules will typically be much less well-defined. The objective will be, say, to get from A to B. By means of perception of available driveways and other relevant factors, and a set of car-driving actions at its disposal, a vehicle-driving system may in theory be able to achieve its objective. However, in many respects the task is much more complex than the task of achieving victory in the game of chess. For instance, is safety to be emphasised in the achievement of the goal? Or speed? If the former, then the system might take forever to achieve the goal. If the latter, accidents will no doubt happen. And so on.

In technical jargon, that which is learnt by learning machines consists of 'statistical models' that 'fit' to given data. Mark Coeckelbergh explains how algorithms can identify patterns or rules in data and use those patterns or rules to explain some data and make predictions for future data. In learning machines this can be done autonomously by the system. Humans will always provide certain inputs that frame the system, such as presenting the system with a task. The system can then, without direct instruction from programmers, 'find rules or patterns that the programmer has not specified. […] Humans give feedback, but they do not feed it specific instructions or rules' (Coeckelbergh, 2020a, p. 84). Whereas adult humans with well-developed and fine-tuned heuristics tend to start with theory in order to make sense of the perceptions of their environments, learning machines, much as we can imagine the infant to perceive its world afresh, can in some instances start with data. Learning machines then adapt so that they fit to the data in a way that enables them to achieve their objectives.[9] Technically, there will still be inbuilt assumptions about the nature of the data processed by the system. Such inbuilt assumptions, however, do not resemble the structures that humans think of as theories. In order to grasp the difference, imagine the way in which a human infant perceives its environment, and then compare it with how a bee might perceive its environment: the same raw data, quite different perceptions of the data.

A 'reinforcement learning algorithm' is to the artificial system as reward and punishment is to the human learner: by instructing a system, for instance, to collect 'wins' or 'successful drives', its human handlers incentivise the

[9] Later we shall see that machines are capable not only of adapting themselves to data, but also of adapting data; notably, they resort to such adaptation in order to be able to predict future data patterns better. When human beings, or rather things abstracted from human beings, such as preferences and behaviour, constitute the data, we get algorithmic behavioural modification.

system to learn. With 'simple' tasks such as playing chess, the system can 'know' when victory is achieved without human feedback; with more complex tasks such as driving a vehicle safely, human feedback is typically necessary to instruct the system as to its success rate.

Machine learning, furthermore, can be 'supervised' or 'unsupervised'. These technical terms have nothing to do with how much humans are involved in the learning process. In order to illustrate the difference, Coeckelbergh uses the example of the task of dividing people into high and low security risk. In supervised learning, the algorithm is provided with a particular variable as 'the target for prediction' (Coeckelbergh, 2020a, p. 85). Possible categories are 'age', 'sex', 'beard', etc., and possible variables are 'years lived', 'male or female', and 'colour/shape'. The algorithm is then fed with data; it makes judgements about high or low security risk; and its human handlers provide feedback, such as 'correct' or 'incorrect'. If the system is well designed and trained on relevant data sets, performance will tend to improve with practice.

Unsupervised learning is more akin to how we could imagine that the infant, of its own accord, will structure the world cognitively by forming such categories as familiar and unfamiliar. In unsupervised learning, categories and particular variables are not provided to the algorithm. Instead, the algorithm is left to its own devices and given access to data. If the task is to divide people into low and high security risk, the algorithm selects its own variables and generates its own categories; these might seem arbitrary or even nonsensical to human experts, but they can nonetheless turn out to be statistically significant. Humans still provide feedback to the system with such terms as correct or incorrect. Sometimes the categories invented by the system make perfect sense; and in such cases this method can open the way to new knowledge (Coeckelbergh, 2020a, p. 86).

In the examples discussed above, the analogy of 'learning' in 'machine learning' has been emphasised. But, as in the case of 'intelligence' in 'artificial intelligence', although analogies in general can enable us to better see some common patterns in the items that are analogised, they may also frame a distorted or one-sided view of reality. In seeking fuller understandings of new phenomena – in this case, advanced algorithmic technologies – we should probably not stop at the first convenient analogies that happen to come our way. We should be open to the exploration of complementary analogies. Matteo Pasquinelli suggests that we think of learning machines or AI 'as an instrument of knowledge or logical magnification that *perceives* patterns that are beyond the reach of the human mind' (Pasquinelli, 2019, p. 4). Not only does it 'perceive', it also 'produces'. In the words of Pasquinelli: 'From the point of view of the statistical model, three modalities of operation of machine learning are given: 1) training, 2) classification, and 3) prediction. In more intuitive terms, these can be defined as: pattern abstraction, pattern recognition, and pattern generation' (Pasquinelli, 2019, p. 8).

We are now familiar with the analogies of intelligence and learning. Both are analogies between something that humans tend to have or do and machine behaviour. If, in our endeavour to understand the nature of algorithmic technologies, we move beyond analogies altogether, we enter the domain of computational statistics. We need analogies in order to grasp the cognitive and social implications of the technologies under consideration. Pasquinelli, however, suggests that more useful analogies could be construed: analogies not between what humans tend to have or do and algorithmic technologies, but between other technologies and algorithmic technologies. According to Pasquinelli, the concept of 'optical media' is a more heuristically useful analogous concept if the purpose is to understand what AI really does. Computer scientists, according Pasquinelli, customarily think of AI technologies as techniques of information compression. Information compression can enable us to see things we would not be able to see without it, but it also implies information loss. We can understand some algorithmic technologies in analogy with how we understand the microscope and the telescope rather than in analogy with how we understand the human brain: the instruments enable us to see things that we would not be able to see without them; but when looking through them only, we also lose information that we would most likely be able to access without them.

Here we can return to our comparisons between the non-trained and heuristically poor child and the well-trained and heuristically rich adult. In circumstances for which the adult has been trained, the adult's rich heuristics will be of much benefit, and the information loss will be of little practical consequence. When circumstances change abruptly, established heuristics can become useless. The under-specialised child may then be in a better position to adapt to the new circumstances. This pattern is borne out in the common understanding that the young are better at adapting to new circumstances. By virtue of their adaptability, a good case could be made that the young play a crucial role in enabling entire populations to adapt to new circumstances. Where changing circumstances cause the long-established heuristics of older generations to fail, the developing heuristics of the young, who are more attuned to new circumstances, enable gradual adaptations of entire populations.

The analogy of optical media provides a complement to the more established analogies of intelligence and learning. For the purposes here pursued, all three analogies are useful. As we shall see, depending on the level of abstraction we adopt when we interpret what algorithmic technologies do, algorithmic technologies can be interpreted differently. We shall see that it often makes sense to interpret them as extensions of individual agencies or collective organisational agencies; in some instances it also makes sense to interpret algorithmic technologies as agents that make decisions of their own accord. To illustrate: in the context of a large corporation, algorithmic technologies can boost in multiple ways (and, as we have previously intimated, at the same time limit or narrow) the perceptive and executive power of the executives

who run the corporation; at the same time, some algorithmic systems can, of their own accord, veer in unexpected directions, and in ways that suggest high degrees of both instrumental rationality and autonomy.

Adaptation

The infant, in analogy with how learning machines 'fit themselves to the data', can also be understood in a sense to fit itself to its environment. As the infant engages with its environment by exercising the moves at its disposal, feedback from the environment enables it to finetune its moves and to improve its performance. Likewise, a tennis player who learns the game, over time and on the basis of feedback, will finetune movements to produce an even stronger performance. In our complex world, as we saw in the example with the heuristically rich adult who suddenly becomes over-specialised, prior learning and training do not always entail cost-efficient rewards. In some circumstances, prior knowledge and skills may even be to a subject's disadvantage.

These days, work-related circumstances frequently change so that previously valuable knowledge and skills become irrelevant or obsolete. From a personal point of view, we may at times re-evaluate the very goals that acquired skills are meant to pursue. In such contexts we may find that, to paraphrase Jesus, we need to become less adult and more childlike;[10] it means that we need to bracket heuristics that have proven useful in the past in order, once again like a child, to look on things with new eyes. The experience of something like 'relevance failure' can awaken an awareness that we must go back and reconsider pre-held beliefs and convictions.

Learning machines typically have inbuilt mechanisms for dealing with discrepancies between inner heuristic configurations and environmental requirements. Through iterative trial-and-error processes, machines are able to construct and adapt heuristics so that they become more attuned to environmental requirements, thus enabling them to accomplish their goals more efficiently. When environmental circumstances change, they are able, to various extents, to readjust. Through what computer scientists call 'back-propagation' learning, machines can adjust the weights in their multiple units to fit the data better, and thereby improve output.

The 'deep' in 'deep learning', Mitchell informs us, refers not to the depth or sophistication of what is learnt, but to the '*depth in layers* of the networks being trained' (Mitchell, 2020, p. 72). In other words, it refers to the previously mentioned 'hidden layers' or 'interior units'. In the case of a self-driving vehicle system, operating with the objective to navigate from location to location, there can be millions and billions of interior units. Each unit has many

[10] Matthew 18:3: 'Truly I tell you, unless you change and become like little children, you will never enter the kingdom of heaven' (NIV). One possible interpretation is that the example of Jesus is so *unlike* anything to which his listeners have become accustomed that any would-be apprentice shall have to become *like* a child in order to learn anew.

weights that determine its output. Back-propagation mechanisms enable these systems, on the basis of feedback, to self-adjust weights in their units. Feedback informs the system about discrepancies between desired outputs and actual outputs.[11] We could say that learning occurs as the machine, by processing example data, reduces this discrepancy.

There are limits to the adaptive capacity of current algorithmic technologies.[12] If circumstances change too much, adequate adaptation will require human intervention. Furthermore, if we think of algorithmic technologies as being in the state of acquiring something like an agency, in analogy with how we may understand humans to have agency, then this agency will typically be very narrow in comparison with human agency. A self-driving vehicle system cannot suddenly 'change its mind', decide that driving vehicles is not a worthwhile activity after all, and instead choose an entirely different pursuit. It lacks that 'something' that would enable it, like a human, to choose to pursue an entirely different objective.[13]

This does not necessarily mean that the initial example of human re-evaluation of goals was misplaced or misleading. If it is the case that algorithmic technologies cannot enact anything like an existential re-evaluation of the main goals they are set to pursue, it is nevertheless the case that, on the way to accomplishing a main goal, there can be many subgoals. Some systems can at some stages deem it instrumentally useful to pursue some subgoals in order to accomplish a main goal efficiently. The instrumental value of subgoals, as opposed to main goals, *can* be re-evaluated by some algorithmic systems. New subgoals can be deemed worthy of exploration.

Knowledge and know-how ambiguities

Some of the difficulties encountered when learning machines are trained and used appear to mirror a common but often neglected human predicament. The analogies between learning machines and human beings suggest that there might be something inherently problematic with the acquisition of knowledge and know-how – that, in fact, more knowledge and know-how do not necessarily grant more advantages to the knower. When knowledge is attuned to

[11] Or, in the technically more precise words of a computer scientist: 'As its name implies, back-propagation is a way to take an error observed at the output units […] and to 'propagate' the blame for that error backwards […] so as to assign proper blame to each of the weights in the network. This allows back-propagation to determine how much to change each weight in order to reduce the error. *Learning* in neuronal networks simply consists in gradually modifying the weights on connections so that the output's error gets as close to 0 as possible on all training examples' (Mitchell, 2020, p. 31).

[12] We must always keep in mind that there are limits to everything, including the adaptive capacity of human beings.

[13] This, at least, is how humans tend to understand themselves. If we adopt the position that humans do not have free will, then the 'uniqueness' of human agency will tend to fade. The difference between us and a self-driving car would not then be that we can choose and the car cannot, but that we humans, somehow, have integrated many more goals that *can* be pursued, whereas the self-driving car could be understood as a simpler sort of monomaniacal agency.

critical tasks and to relevant environmental circumstances, of course, all other relevant circumstances being abstracted, it is advantageous to the possessor of knowledge. But as Joseph Weizenbaum once quipped in response to Herbert Simon's assertion, 'Knowledge is better than ignorance': 'Yes! But not at any price' (Weizenbaum, 2008, p. 42). Whether we should consider that more knowledge is indeed desirable will depend on the cost of acquiring it and maintaining it. It is not necessarily the case that a civilisation that knows how to build space rockets is necessarily better than a civilisation that has not acquired such knowledge; nor is it necessarily the case that a civilisation that knows how to construct artificial intelligence must necessarily be considered an improvement on the civilisational structure that preceded it.

There are at least two sides to the ambiguous status of knowledge. The first and most straightforward is simply the result of inherent systemic limitations: everything that is learnt implies a cost. As became obvious when we considered the example of chess-playing agents, both human and artificial players are in fact quite limited. Moreover, humans will typically not have the time to learn to do many things very well. If this is so, how do we, in the larger scheme of things, know which knowledge and know-how will be relevant to future contexts?

The second side to the ambiguous status of knowledge involves the quality or, rather, the attuning of things learnt. It will be useful here to think of both knowledge and know-how as heuristics. Let us consider the case of a type of heuristics that could be labelled 'ideological'. If one attaches this somewhat disparaging label to a person or a group of persons one may also observe that, although ideologues typically describe some genuinely existing patterns correctly, which is to say that ideologies do not typically propagate views of reality that are one hundred per cent inaccurate, ideologues also typically miss out on many important circumstantial aspects, variables, and pursuit-worthy goals. It will more often than not be the case that the person who makes the judgement that someone else is ideological does not consider him- or herself to be ideological. But we have just been informed that processes that imply information compression also imply loss of information. An ideology, in the pejorative sense, can be understood in analogy with how we understand the microscope and the telescope; but so too can all knowledge and know-how that we do not discard as ideological, including scientific theories. This dynamic of gains that always imply losses appears, then, to be all-pervasive. It becomes problematic to the extent that the value of gains does not outweigh the importance of losses.

Humans, in their agentic capacities, are *trained to see things in certain ways*. The craftsman, the scholar, and the scientist are trained in accordance with the requirements of their professions. When the training is successful, each practitioner becomes able to perceive and act in ways that, in turn, are perceived to be of value to other people and to social institutions. In societies that depend on high degrees of specialisation, specialisation becomes a

problem when specialised persons, owing to loss of information and hence a lack of width, miss out on things of general value. This can become a serious problem if specialised persons begin to miss out on things that are critical to basic health and survival.

By thinking of specialisation in terms of heuristics, we can compare specialisation with the pejorative notion of ideology, which, of course, is also a form of heuristic. In the case of ideologies, most people can name at least one or two examples that they consider entirely harmful. However, is it inconceivable that there might also be forms of specialisation, such as scientific and technical forms, that are – if not entirely, then at least on balance mostly – also harmful to human society? If one were to answer this question affirmatively, then one would, at least implicitly, make the judgement that entire ways of training perception and cognition can be harmful, or at least mostly harmful, to human society. In order to answer such a question concretely, one must of course first specify some 'good' or set of 'goods' that could be undermined. The answer will depend, to a large extent, on what one means by 'society'. The organisation of contemporary technically advanced societies depends, of course, on high degrees of specialisation. If one were to find that some form of specialisation on which one's society depended also happened to be very harmful – for example, that it severely hindered human flourishing – one would find oneself in an intriguing predicament.

If there is some merit to this way of reasoning about the general role of heuristics in human society – regardless of whether, in the eyes of the person making the judgement, the heuristics under consideration happen to be framed by structures that are pejoratively labelled as ideological or by scientific and technical methodological frameworks that are deemed to confer on heuristics the status of knowledge – then what are we to think of the proliferation of algorithmic technologies – algorithmic technologies that, based on *their* own heuristics, are achieving ever higher degrees of 'specialised' knowledge and know-how? Given the degree to which human specialisation has contributed, for good and ill, to changing our life-milieus, what can we expect that such machines will accomplish in our life-milieus? Moreover, as narrowly efficient algorithmic procedures begin to frame human culture, what information do we risk to lose?

So far, the differences in how infants and adults perceive their environments have been discussed. The rich cultural diversity of humankind reveals another dimension that must be considered. Since adults all over the world learn differently, there is a difference between adults too – depending on their sex, culture, professional training, temporal contexts, etc. – in how they perceive their environments. As algorithm-powered organisations such as Google attain global hegemony, do we also risk a loss of cultural diversity?

In the case of unsupervised learning, we saw that learning machines seem able to access patterns of reality that have not been accessible so far to humans. The flipside of this is that learning machines will rarely perceive the world in

any way that resembles human perception. In some instances, they are simply unable to perceive that which humans commonly perceive, such as smells. In other instances, they may perceive much more, although this 'more' may also imply a loss. In the future they might become able to perceive all that humans can perceive and far more. We could, then, end up with a society of human agents whose perceptual and cognitive features are fairly familiar to us, and clusters of very different entities that perceive and act in ways that will often be incomprehensible to us. This difference, as we see later, can yield both opportunities and risks. On Pasquinelli's analogy with the microscope and telescope, it can enable us to see new things; but we also risk missing out on important things *if* we allow ourselves to become too dependent on such technologies. On the learning agent analogy, it may enable automation of many functions that have not yet been subject to automation, which may increase administrative efficiency in all kinds of domains; but we also run the risk of constructing life-milieus and societies that are even more difficult to observe and understand than our already highly specialised technoscientific societies.

We have now reached an understanding of how algorithmic technologies work that is sufficient for the purposes pursued in the coming sections. The remainder of part I contains a presentation and discussion of how some prominent thinkers view the role of human agents in future algorithm-intense environments. These views are posited as three paradigmatic cases for humans in such environments. The paradigmatic cases then serve as points of reference in subsequent discussions.

Since the concepts 'agency' and 'agent' are central to all subsequent sections, the time has come to explain how the concepts are used. Then, after the presentation of the three paradigmatic cases, the concept of 'agent' is integrated into the concept of 'multi-agent system'. Part I ends with a brief discussion of yet another concept: 'affordance'. Affordance and multi-agent system are then used as theoretical tools in order to discuss the role of human agents in algorithm-intense environments.

Agency and agents

This section clarifies how the concepts 'agent' and 'agency' are used. This is done in order to consider human beings in the context of so-called hybrid multi-agent systems – systems within which both human and various types of artificial agents interact for the sake of higher purposes. First, some key examples of how the concepts have been used in philosophy of action are provided. Then it is specified how the concepts are used.

Markus Schlosser introduces his article 'Agency' with the following sentence: 'In very general terms, an agent is a being with the capacity to act, and "agency" denotes the exercise or manifestation of this capacity' (Schlosser, 2019). Later Schlosser adds that it is widely agreed 'that agency involves the

initiation of action by the agent' (Schlosser, 2019). This widely inclusive use of the concepts is eventually retained; but first we look briefly on some more specific understandings of agency that are also relevant to the purposes being pursued.

Schlosser informs us that the philosophy of action provides us with a standard conception and a standard theory of action: 'The former construes action in terms of intentionality, the latter explains the intentionality of action in terms of causation by the agent's mental states and events. From this we obtain a standard conception and a standard theory of agency' (Schlosser, 2019). The standard model, according to Schlosser, builds on a long philosophical tradition that includes Hume and Aristotle and the more contemporary analytic works of Gertrude Elizabeth Margaret Anscombe (1957) and Donald Davidson (1963). Actions, according to this conception, are to be understood as intentional performances that are performed for reasons. The standard theory, in turn, explains that 'something is an intentional action and done for reasons just in case it is caused by the right mental states and events in the right way. The right mental states and events are states and events that rationalize the action from the agent's point of view (such as desires, beliefs, and intentions)' (Schlosser, 2019). These claims have been questioned from many different angles, which in turn has produced an ongoing debate in the philosophy of action.

Some opponents of the standard conception argue that it is wrong to reduce action to intentionality motivated by reasons, and that:

> the exercise of agency may be entirely spontaneous, in the sense that an agent may initiate an action for no reason and without prior intent. On this view, reasons and intentions may have a strong and even decisive influence on how an agent acts. But agency has its source in the power to initiate, and the exercise of power cannot be reduced to the agent's being moved by reasons or intentions. (Schlosser, 2019)

This critique of the standard conception of agency amounts to an alternative conception of agency. Two names associated with the alternative conception are Carl Ginet (1990) and Timothy O'Connor (2000).

Moreover, the standard conception's focus on reasons seems especially problematic when we consider simpler organisms. The standard conception may be fairly adequate for describing a special type of agency – that is, the intentional agency of human beings. Assuming that simpler organisms have agency, we are invited to consider the possibility of agency without mental representations. Schlosser offers three claims that are raised in connection with this line of criticism:

> According to the first, there are non-human beings that are capable of agency and that do not possess representational mental states. Second, there are many instances of human agency that can and should be explained without the

> ascription of representational mental states. Third, all instances of agency can and should be explained without the ascription of representational mental states. (Schlosser, 2019)

This criticism raises the prospect of different types of agents expressing different types of agencies. On an intuitive level, it seems that the mere act of ascribing agency to any phenomenon implies that we also ascribe something equivalent to mental states to the agency. But when is it at all appropriate to ascribe mental states and agency to phenomena? Daniel Dennett proposes an instrumentalist view of agency (Dennett, 1987). Schlosser explains that, for Dennett:

> the question of when it is appropriate to ascribe mental states cannot be separated from the question of when it is appropriate to ascribe agency, and both questions are to be answered in terms of predictive success: it is appropriate to attribute mental states in the explanation of agency when doing so supports successful predictions of behaviour. (Schlosser, 2019)

If we imagine some prehistoric animistic context in which agency is attributed to trees, then it seems that, according to these instrumentalist precepts, this attribution would be justified to the extent that it enables successful predictions. On the other hand, 'most proponents of the standard theory presume some form of realism, according to which the ascription of mental states is appropriate only if the agent in question possesses the right internal states with the right representational contents' (Schlosser, 2019). We have here two views in apparent conflict concerning how one ought to proceed when ascribing agency to phenomena. On the one hand, we have the realist view, which includes the standard and the alternative conceptions; and on the other hand, we have the instrumentalist view.[14] In the context of this treatise, we could think of them as complementary rather than conflicting.

In algorithm-intense environments it can be difficult to analyse environments from the realist point of view – that is, in terms of discrete agents expressing the agency proper to each agent. Does an algorithmic chess-playing machine have agency? To the extent that it is able to initiate actions on some level of analysis, it could be understood to have some rudimentary type of agency, but in this case that agency would be restricted to a very narrowly defined arena. The same would be the case for current algorithmic assistants and social robots, but here the arenas would be less narrowly defined.[15] To the

[14] This very brief account does not do full justice to either view. The different options can have all sorts of ontological implications that cannot be treated in the context of this treatise. Notably, arguments for the instrumentalist stance may involve a materialist conception of nature that denies that phenomena such as mental states have the significance that realists ascribe to them.

[15] When their being understood as having some degree of agency is conditioned on their ability to initiate action *on some level of analysis*, then this should be understood to occur within, for instance, the context of an ongoing game of chess, or in the context of an ongoing human-assistant/social robot interaction. We learnt earlier that algorithmic systems can be able to

extent that algorithmic technologies emulate human social behaviour, and given the human perceptual apparatus, it may even make good predictive sense for humans to interpret them as agents expressing their own proper agency. However, on a different level of analysis, that which to humans appears to be an agent may turn out to be a mere instrument in the service of some higher and more integrated systemic agency. On the realist view, it would be awkward to interpret any contemporary algorithmic gadget or interface as an agent expressing *its own* agency.

If we adopt Dennett's instrumentalist stance, there is, on the other hand, no obstacle in principle to attributing agency to algorithmic systems. The criterion for when this is appropriate is tied to predictive success. From the point of view of a human being, without any doubt it would often be reasonable to adopt the instrumentalist stance. On some basic level, our survival and well-being depend on our ability to make fairly successful predictions of our environments. With respect to these matters, the question of which view to adopt is a pragmatic one. However, even if attribution of agency may contribute to yielding successful predictions on some specific level of analysis, attribution of agency might simultaneously dissimulate the real workings of larger contexts.[16] If, instead, we attempted to read algorithmic environments from the realist stance, we would be incentivised to widen our horizons. In algorithm-intense environments, it will often be the case that the seemingly discrete and identifiable *thing* – the algorithmic-assistant, the social robot, or the AI-partner – that behaves as if it were an agent is physically separate from the events that ultimately cause its behaviour. We live in the age of interconnected devices and interfaces – the Internet of Things. It may therefore make better sense to understand individual devices and interfaces as limbs or organs in larger bodies than as discrete agents.

Such an understanding, in turn, leads us to question the role of humans who interact with such systems. Is it always straightforwardly sensible to understand humans as agents expressing their own agency, or should we, when we interact with certain complex systems, also or even primarily be understood as limbs of larger organisms? A realist stance, while it may come with its disadvantages, prompts us to seek further for the source of a higher agency. Does the technically engendered gestalt that is perceived really have the power to participate in meaningful interaction? If not, where should we look for agency? Are we becoming immersed in environments partly composed of

initiate the pursuit of subgoals in order to achieve their main goals. If we restricted the level of analysis to the context of the machine and the goal pursued, the initiation of pursuit of subgoals could plausibly be understood as initiation of action, which is qualitatively different (or on a different order of complexity) from mere mechanical responses to environmental stimuli or input.

[16] Compare, for instance, a level of analysis that focuses on an agent's ability to navigate its environment so that it achieves some goals with a level of analysis that focuses on the power structures that define that same environment.

artificial agent-appearing phenomena that in reality are expressing some higher systemic or human agency? Such questions, as is demonstrated in part II, are not only of abstract academic import. Adopting a realist stance could be just as pragmatically justified as adopting an instrumentalist stance, in view of different kinds of pursuits.

If these notions seem somewhat vague at the moment, things will become clearer as the concept of multi-agent system is treated and the analyses in part II are undertaken.

***

It is now specified further how the concepts are used. The concept of 'agent' applied to 'human beings' emphasises the capacity of humans to engage in activities for the sake of higher purposes.[17] In the case of human beings, humans *qua* agents are themselves to be understood as initiators of acts. We can, in normal circumstances, fairly straightforwardly ascribe agency to human agents. The possibility for spontaneous action is not without interest, in that spontaneity can disrupt work for higher purposes and/or accidentally lead to new venues for higher purposes. However, in exploring how human agency and autonomy are likely to be affected in algorithm-intense environments, it is very much the capacity for humans to be agents for higher purposes – higher than, say, eating or sleeping – that is to be problematised.

The notion of different *types* of agents, including artificial agents, has already been evoked. Even prior to any talk of artificial agents, humans have observed and interacted with different types of agent. In many instances, humans have included other types of agent in collective endeavours for higher purposes. Old-style farm villages have used (and some still do use) dogs, oxen, and horses in the fulfilment of context-specific objectives. Horses have played a significant historical role in military affairs. Meanwhile, the human-independent case of the ant colony has frequently been used as an illustrative analogous case for how a well-disciplined military organisation or even a well-ordered polity ought to function. The human being *qua* agent, it seems, has always been the expression of a type of agency that seeks to use and to compare itself with other types of agency.

[17] The assumption is that human beings live in hierarchies of purpose. A peasant may learn the art of agriculture to become a valued member of peasant society. The art is then practised with all kinds of practical agriculture-related purposes in view, but also, perhaps, for the higher purpose of feeding a family. Purposeful hierarchies can be understood or perceived more or less vaguely or more or less precisely. The hierarchy of the peasant can fade out after the rearing of a family, in which case there is no higher purpose than the family. Or the family can be reared for the purpose of a country, or for the purpose of giving glory to God. In algorithmic systems, the highest purpose will tend to be precisely defined: it is the main goal that it has been algorithmically instructed to accomplish. If human beings are often wobbly and weak in their pursuit of their highest purposes – there are so many distractions that interfere – algorithmic systems will tend to behave like monomaniacs in their relentless pursuits.

In comparison with horses, oxen, dogs and, most of all, ants, the human type of agency appears to have by far the largest scope for spontaneous action and by far the largest capacity for inhabiting a variety of different agent-roles.[18] In organisations that require high degrees of order and discipline, spontaneity is often understood as problematic. This explains why the ant colony can be viewed as an ideal, as the 'wiggle room' for spontaneity among ants is generally believed to be close to zero. However, ant colonies are rarely used as ideals for modes of human organisation that are perceived to require high degrees of creativity or, indeed, wherever democratic decision-making is upheld as an ideal. Different types of organisation entail different sets of requirements of the agents that compose them. This is also the case with the hybrid multi-agent systems that are discussed later.

For the purpose of the analyses to be undertaken, a fairly generic model for human agency is assumed. On this assumption, humans act within prior geographic and cultural constraints. There is nevertheless also considerable scope within which we experience ourselves as able to think freely, imagine, and initiate actions. We also experience considerable scope for spontaneous thought and action, which we may interpret as 'accidental' when spontaneity incurs unfortunate consequences and as 'serendipitous' when it incurs fortunate consequences. We encounter more specific philosophical anthropologies in part III. The generic model for human agency nevertheless serves as a background assumption throughout the treatise.

Assuming such an understanding of human agency, we can also infer a few key ways in which human beings can enact agent-roles for the sake of higher purposes. We could plausibly suppose that ants in an ant colony have no choice whatsoever when they assume their respective roles in the colony. The cases of the dog, the ox, and the horse probably involve considerably more room for negotiation in the assignment of status and roles in the respective contexts in which they live their lives.

In the case of human beings, there are no doubt many ways in which humans can choose or be made to enact agent-roles. In traditional societies, by means of socialisation, enculturation, and other processes, humans have been made to assume the traditionally sanctioned roles that have animated any given culture.[19] Another historically prevalent method used to make humans

[18] Again, if one happens to be convinced that there is no free will, the phenomenon that humans tend to interpret as spontaneity may indeed be the result of deterministic processes.

[19] Bambi Chapin, Christine El Quardani, and Kathleen Barlow define and differentiate the processes of socialisation and enculturation in the following way: 'Socialization refers to the process through which people develop culturally patterned understandings, behaviors, values, and emotional orientations. The meaning of the term overlaps with "enculturation" (the process through which children first internalize culture), "acculturation" (the process through which people adopt new cultural models and ways of behaving), and "subject formation" (the process through which subjectivity is shaped)' (Chapin, El Ouardani and Barlow, 2016). The brief characterisation of traditional societies is loosely drawn from René Girard's wide-ranging

accept agent-roles has no doubt been coercion. In the case of the former, the agents who accept the roles generally stand to benefit from accepting them to varying degrees. In the case of the latter, there may be little or no benefit to those who are coerced into accepting their roles.

The very same processes continue to influence humans to accept roles to this day. Cultural norms have changed. With the onset of modernity, the way of spontaneity or free choice gained prominence. There are many philosophical underpinnings of the dawning notion of the autonomous individual (e.g., Immanuel Kant) being able to choose his or her own destiny (e.g., Friedrich Nietzsche). Although the notion of heroism is commonly associated specifically with Nietzsche, in modernity there is a more generally heroic understanding of the individual who liberates him- or herself from the restricting conventions of culture in order to become a free, autonomous, and therefore also 'self-made' reasoner and chooser.

Even though in contemporary contexts we may experience that we exercise our individual liberty relatively unrestrained in comparison with how we understand the conditions of our ancestors, such impressions may be misleading. This brings us to another way that is gaining prominence, especially in late modernity. In the context of scientifically informed and technology-intense modernity, contemporary societies have developed new means of incentivising human beings voluntarily to accept roles that have been prepared for them.[20] The new means have for a long time been restricted to conventional forms of education, publicity, and other so-called integrative propaganda methods.[21] Now, when we are on the verge of crossing over into algorithm-intense environments, we must consider the extent to which our life conditions, including the voluntary and involuntary acceptance of agent-roles, may increasingly be defined by manipulative techniques applied not by human individual or collective agents, but increasingly by algorithmic systems attaining ever higher degrees of autonomy. Algorithmic systems, in turn, will be deployed in the larger context of hybrid multi-agent organisations.

***

Now that we have acquired a basic understanding of how the concepts of agency and agent are used, we can begin to look at the three paradigmatic

discussion of prominent anthropologists' understanding of the role of sacrifice in *La violence et le sacré* (*Violence and the Sacred*) (Girard, 2010).

[20] Roles will be accepted through the same processes (socialisation, enculturation, etc.) as before, but here the emphasis lies on the new technical means developed in order to impose desired thought and behaviour.

[21] The concept of 'integrative propaganda' is drawn from Jacques Ellul (1990). Whereas 'subversive propaganda' is leveraged with the intent of undermining the cohesion of the targeted society, integrative propaganda can be understood as an umbrella concept for all techniques and means that are leveraged in order to make people feel meaningfully at home in the targeted society.

cases for human agents in algorithm-intense environments. All three cases represent predictive scenarios. Moreover, they imply different types of role for human agents.

Towards a world with less work and more leisure

The notion of AI supplanting human beings in most work-related domains is a commonly expressed fear. If this scenario were to be realised, then AI would of course have profound societal implications in work-related domains and beyond. Here 'work' is initially understood as salaried or paid tasks and responsibilities. Work is then contrasted with other purposeful activities belonging to the category of 'leisure'.

Are we on the verge of yet another wave of technology-caused unemployment? If so, will this wave prove different from previous instances of technology-caused unemployment? In this section we consider the likelihood that contemporary and future algorithmic technologies will have the effect of making human beings redundant. Some possible implications of human redundancy and some possible ways to deal with these implications are discussed. The challenges that are considered in connection with this scenario belong to three main categories: 1) loss of income, 2) loss of purpose and meaning, and 3) loss of *use* for human beings. The first challenge is only briefly considered in isolation from the others. This section includes a discussion primarily of the second challenge, while the third will be further problematised in parts II and III. As we proceed from part I to parts II to III, a variety of perspectives that, in some respects, interlink all these challenges is brought to our attention.

Technological unemployment is a phenomenon that predates modern automation technologies. Historically, new innovations have frequently lowered the demand for certain skills and materials. The innovation of petrochemical engineering almost eliminated the need for whale-hunting skills and for whale oil. At the same time, the innovation generated a need for new skill sets and new types of material. Systematic implementation of early industrial automation technologies also had the revolutionary effect of rendering entire skill sets obsolete. As more and more functions that could be routinised were automated, newly marginalised workers were encouraged to re-educate themselves in order to fill functions that could not yet be routinised and/or entirely new functions that emerged in the wake of the latest waves of innovation.

Similar structural shifts occurred before the modern age. When humans first abandoned their nomadic ways of life and founded permanent settlements, entire skill sets were no doubt also rendered obsolete, while new ones came into demand. In contrast to most contemporary outlooks, the new settlers were surely envisaging a new, stable, and long-lasting way of life for which newly acquired skill sets would prove to be of value from generation to generation. In our contemporary ways of life, in contrast, recurring disruptions and revolutions are not only expected but are often actively promoted. We

therefore do not necessarily expect that the skills of one generation will be relevant to the generations that succeed it. We must constantly adapt to new circumstances, innovate, and then adapt anew to changing circumstances. One could argue that the kind of revolutionary shift that was undertaken by prehistoric nomads-turning-settlers has been institutionalised in the late modern way of life. That is to say that we do not envisage *one* revolutionary shift, but a *rapid series of ongoing revolutionary shifts*.

Presently we are discovering that just about any human work-task can be automated. The tasks that still hunger for human workers are either ones where the incentive to automate is low because of technical difficulties in combination with the availability of cheap labour, as is the case with the tasks of cleaners, or ones that cannot yet be automated. The cognitive requirements for tasks that remain in the latter category tend to increase. As more and more tasks are automated, competition for those that remain becomes fierce. One risk is that vast segments of populations will find themselves permanently outside the workforce. They will then have become structurally redundant. Or so at least the narrative goes. But is it plausible?

The relationship between technological configurations and employment is without a doubt one that involves complex dynamics between multitudes of variables. To illustrate one kind of effect, Stuart Russell uses the image of the gradual development of painting technology: from brushes that are one hair wide to brushes that are one millimetre wide, to ten centimetres wide, to large rollers and spray guns, and finally to teams of house-painting robots that can be controlled by one technician. With the brush that is only one hair wide, no painters are going to be employed. In the following stages, up to and including the ten-centimetre-wide brushes, employment increases. With the invention of rollers and spray guns, the demand for painters begin to decrease. As the technology reaches a stage of quasi-automation, the demand is close to zero. 'Thus, the *direct* effects of technology work both ways: at first, by increasing productivity, technology can increase employment by reducing the price of an activity and thereby increasing demand; subsequently, further increases in technology mean that fewer and fewer humans are required' (Russell, 2020, p. 115).

So, if a technology is at the stage in which further improvement reduces the price of an activity, more demand for the activity can be generated by improving the technology, and thus the need for professional human workers will increase; but when efficiency is increased beyond a certain point, the demand for human workers begins to decrease, and at the stage of automation, demand becomes close to zero.

If the aim is always to take improvement further in view of increased efficiency, we could expect, given that technical resources appear to undergo constant improvement, that more and more functions will sooner or later be on the verge of automation, and that further improvements will lower the demand

for human workers.[22] Historically, industrial societies have been able to rely on the 'complementing effects' produced by innovation and new technologies: human engagement with new technologies has repeatedly tended to promote new developmental processes, as in the case with the painting brush described above. Optimists believe that this will remain the case. But there are reasons to be cautious when we consider highly automated and AI-intense environments.

Historically, Russell explains, the demand for new jobs has tended to arise in the wake of automations that are driven to their logical conclusion. First, in the case of the painter robots, someone would have to make the robots. Still, the rise in this kind of employment would be considerably lower than the decrease in demand for painters. But at the consumer end, it might now be considerably cheaper to have a house painted. Consumer savings could then be spent elsewhere, the financial aggregate of which could finance innovations; thus the developmental process of technology, from invention and gradual improvement to quasi-automation, can begin anew with new technologies (Russell, 2020, pp. 116–17).

Nevertheless, the availability of new areas that could undergo such developmental cycles may eventually decrease, according to Russell. Today a large number of functions that employ large quantities of people are up for automation: targeted functions include a wide range, from relatively practical tasks such as warehouse work and professional driving to tasks that are commonly thought of as white collar and relatively cerebral, such as tasks in the legal, medical, academic, and scientific professions. Tasks that we are accustomed to think of as 'creative work', such as video and music production, writing, and the fine arts, are increasingly being accomplished by algorithmic automation. Even the generation of computer code is subject to automation. According to Russell, 'almost anything that can be outsourced is a good candidate for automation, because outsourcing involves decomposing jobs into tasks that can be parcelled up and distributed in a decontextualized form' (Russell, 2020, p. 119).

Daniel Susskind highlights three features that enable the automation of tasks: '[I]f you come across a task where it is easy to define the goal, straightforward to tell whether that goal has been achieved, and there are lots of data for the machine to learn from, then that task can probably be automated'

[22] Is the aim always towards further improvement in view of efficiency? A good case could be made that the aim in contemporary late-capitalist societies is profit above all. One of the thinkers discussed in part II, Jacques Ellul, argues that the aim of increased efficiency has been embraced as a sort of dogma or axiom in technological societies. Of course, we need not accept any dogma at all in the present analysis, whether capitalist-Marxist or Ellulian; yet the intuitions that underlie these two frames of analysis could be combined. On the basis of empirical observations in the *long-durée*, it could be inferred that technological innovation and increased efficiency, even where no profit venue is obvious at first glance, tend to open up new profit venues in time. A capitalist argument for why 'increase efficiency in all domains' might be embraced – if not as dogma, then at least as a guiding principle – could thus be construed.

(Susskind, 2020, p. 77). On the other hand, the incentive to automate is lower when tasks are involved that are difficult to routinise and/or when tasks can be accomplished at a low cost by minimally paid human workers. Could we rely on current trends to be good indicators of future developments? Perhaps, perhaps not. A case could no doubt be made that we are presently experiencing a development in which the tasks available for humans tend to become either increasingly cognitively demanding or increasingly low-paid. If such trends were to continue to evolve, we would face a deepening of current extremes. As we see when we consider the other paradigmatic cases, this is far from certain.

Russell argues that it is probable that, within a few decades, all routine labour will be done more efficiently by machines. 'Since we ceased to be hunter-gatherers thousands of years ago, our societies have used most people as robots, performing repetitive manual and mental tasks, so it is perhaps not surprising that robots will soon take on these roles' (Russell, 2020, p. 119). For those who find themselves unable to compete for the few high-end jobs that are left, there might be no viable way left to earn a living (Russell, 2020, p. 120). That which is impossible to routinise today may of course be routine tomorrow.

Is this really the scenario we are facing? Earlier decreases in demand for labour have always been followed by new increases. As new technologies have rendered human labour superfluous in some tasks, demand has risen or even been created for the performance of other tasks. Susskind does not doubt that there will be an increase in demand for the performance of new tasks; but he questions whether this would solve the problem of human redundancy, and does this by problematising the requirements of the task-performing agent.

People who fear technological unemployment are often charged with committing the so-called 'lump of labour fallacy', which assumes that there is a constant lump of labour. When some individuals do too much work, or when the automation of tasks performed by human labourers occurs, the arguer concludes that there will be less work left for the totality of the workforce. In other words, the fallacy ignores complementing effects. In industrial and technological contexts, complementing effects have typically taken the form of new needs for specialised human workers to serve and maintain new technological infrastructures. Another type of complementing effect might result from technologically enabled savings that free up capital for new investments. So, there is no constant lump of labour; and when economies evolve and grow, the kind of labour that needs to be done tends to evolve and grow with it. Susskind argues that, over time, the lump of labour fallacy argument is likely to become a fallacy itself, which he calls the lump of labour fallacy fallacy (or LOLFF):

> It may be right that technological progress increases the overall demand for work. But it is wrong to think that human beings will necessarily be better placed to perform the tasks that are involved in meeting that demand. The lump

> of labour fallacy involves mistakenly assuming that the lump of work is fixed. But the LOLFF involves mistakenly assuming that that growth in the lump of work has to involve tasks that human beings – not machines – are best placed to perform.' (Susskind, 2020, p. 126)

In other words, re-engineered technical environments may produce contexts in which human beings no longer embody *the type of agent* best suited to meet new requirements. Perhaps there will be other *agentic systems* that will be better suited to meet the requirements of future work-related tasks.

Susskind argues not that the world of tomorrow will necessarily be a world with no work for humans, but that it will be a world with less work. He gives no precise prediction as to how long it would take to reach a state of redundancy for large numbers of humans. He is confident, however, that the machines of tomorrow will be more capable than those of today. He predicts, therefore, that more and more tasks will be taken over by machines (Susskind, 2020, p. 128). In the short term, the challenge is so-called frictional technological unemployment – that is, unemployment that could, over time, be remedied in various ways; but in the long term, Susskind argues that we must take the challenge of 'structural technological unemployment' seriously: there might simply not be 'enough demand for the work of human beings'[23] (Susskind, 2020, p. 129). *The reason is a potential waning of complementing forces as they bear on human agents, but not necessarily a waning of complementing forces as they bear on all other types of agent.*

Susskind's scenario seems plausible. Technologically advanced societies are, to varying extents, already preoccupied with this challenge. As the automations of new tasks proceeds, we can expect that the intensity of the challenge will also increase. At the very least, occupational structures to which we have grown accustomed will undergo profound change.

We now consider two of the challenges related to human redundancy. Let us begin with a brief consideration of the fairly familiar predicament of loss of income. Most people, at some stage of life, have experienced loss of income. If the scenario of Susskind is on target, we should expect that populations in the near or not-too-distant future will experience loss of income on a massive scale. Some argue that loss of income from unemployment could be remedied by a universal basic income. This would take care of the problem of poverty and inequality.

[23] Susskind describes 'frictional unemployment' as follows: '[T]here is still work to be done by human beings: the problem is that not all workers are able to reach out and take it up' (Susskind, 2020, 99). This inability to reach out is explained in turn by mismatches that could be remedied over time, such as 'skills mismatch', 'identity mismatch', and 'place mismatch'. 'Structural unemployment', on the other hand, implies that there is simply not enough work for humans to do, full stop. Structural technological unemployment, on this view, is a consequence of the weakening of complementing forces (Susskind, 2020, pp. 99–113).

Loss of income is in some respects the most acute challenge. If it is not dealt with, people will simply lose their place in the social order, which in our time includes only people with income or wealth. Some even appear to see loss of income as the major problem to be dealt with, understanding the cause of the loss in terms of the otherwise positive potential liberation from the need to work. The liberation, in turn, could be seen as an opportunity to increase opportunities for human flourishing. Russell tells us that people such as John Maynard Keynes and Bertrand Russell envisaged a world with less work, to the benefit of future generations (Russell, 2020, p. 120). There is in fact a long-established convention of viewing work as demeaning and burdensome, as a *must*. On this understanding, the need to work is a hinderance to finding genuine purpose and meaning in life. So, if we could only solve the problem of income, then everybody would be better off in a world with less work. But should this view be accepted? The remainder of this treatise offers many implicit answers to this question. From now on, loss of income is considered only to the extent that it relates to the other two challenges, loss of purpose and meaning and loss of use for human beings.

Let us then consider the question of purpose and meaning. We have seen that Russell makes the assumption that, ever since humans ceased to be hunter-gatherers, most societies have used most people as robots (Russell, 2020, p. 119). In many instances, humans have no doubt been used as robots. Yet, to make such a sweeping generalisation is perhaps also convenient in view of the need to legitimise automation. As we proceed through parts I, II, and III, this assumption is problematised. We consider the human being as a potential knower and practitioner who is endowed with a capacity for spontaneity. The agency-venues, including work-related tasks, that are available to humans in any given culture could be understood as representing frameworks within which humans, thus understood, could find purpose and meaning. By 'agency-venue' is meant an agent-specific opening for action and praxis in any given context. To the extent that agency-venues represent tasks that are necessary for group survival, they also represent means by which a plurality of human beings could be bound together in common purpose and meaning. Some degree of common purpose and meaning, one could plausibly assume, must represent a basic building-block of human society. When agency-venues expressing common purpose and meaning are habitually enacted, we can assume that they will also be encoded in worldviews. On this view, the very viability of human society may depend on the availability of critical work-tasks.

Russell's analogy between humans and robots may make some sense in contexts where some groups of humans have exploited other humans as manual slaves. In the context of social communities, however, it is just as plausible – or perhaps even more so – to understand work-related tasks as venues through which human beings, through their knowledge, skills, and courage may receive the practical valuation and dignity that are associated with being

useful to their communities. Now, in the context of automation, as some of the work-related agency-venues available in any given culture are being taken over by automation, other agency-venues may open. But what would happen if there were no community-enriching agency-venues available at all to most human agents? What would happen if humans became useless to other humans and to society? We delve deeper into these social dynamics in parts II and III.[24] For now, let us see what Susskind thinks about the loss of work-related purpose and meaning.

Whereas many individuals would no doubt be better off with less work, the loss of purposeful obligations on a grand scale could hardly be considered an unmitigated blessing. Susskind recognises the problem of meaning and purpose, to which he devotes the concluding chapter of his book *A World Without Work*. Technological unemployment would deprive people of significance and sense of purpose: 'In a world with less work, we will face a problem that has little to do with economics at all: how to find meaning in life when a major source of it disappears' (Susskind, 2020, p. 215).

Revisiting Sigmund Freud and Max Weber, Susskind brings to the discussion another long-established view of work (the former being work conceived as drudgery): at least in modern western culture, work may have structured important foundations for human community and even provided outlets for religious devotion (Susskind, 2020, p. 216). Susskind then delves into another aspect of the valuation of work: in a meritocracy, the perceived merit of those who achieve implies the lack of merit of those who do not achieve. In traditional class-based societies, on the other hand, there was no doubt a good chance that those who were born into great fortune might have been aware of their luck. Is there a similar awareness in the meritocratic order of the luck of being born into a good family, receiving a good education, and so on? It is no doubt the case that those who go on to accomplish great things often think that they merit their lot; and those who do not achieve great things think they may merit their lot as well. The meritocratic ethos, then, tends to make things even more difficult for those who are unemployed.

If this is a fair description of modern meritocratic societies, then how have these dynamics played out in other types of society? Susskind informs us that, on the whole, hunter-gatherer societies spend less time than we do on work-related tasks. Classical western culture, according to Susskind, appears to have despised work. One Greek word for 'work', 'ascholia', translates as 'non-leisure'. In Hesiod's *Work and Days*, work is understood as a punishment from

[24] Agency-venues tied to 'group survival' was mentioned as one potentially unifying factor. Later we see that many other goals could serve this unifying function, and that people – in sports, in religions, in politics, and in cultural endeavours – could unite in common purpose and meaning independent of mere survival.

the gods.[25] In the Biblical story about the fall of Adam and Eve, work is described as a negative consequence of the expulsion of Adam and Eve from Paradise. 'These stories should remind us that the link between work and meaning, no matter how Freud and Weber may extol it, may in fact not be so clear' (Susskind, 2020, p. 222).

Work can, of course, consist of tasks that vary in nature. Many kinds of work are rightly described as drudgery, and much work in today's semi-automated warehouses is no doubt hardly more meaningful than the degrading factory work considered by Karl Marx during early industrialisation. Susskind nevertheless argues that work today has been exalted, at least for many of us, as the means to find meaning and purpose:

> Like a drug, it provides some people with a pleasurable burst of purpose. But at the same time, it intoxicates and disorientates, distracting us from looking for meaning elsewhere. […] Work is so entrenched in our psyches, we have become so dependent on it, that there is often an instinctive resistance to contemplating a world with less of it, and an inability to articulate anything substantial when we actually do. (Susskind, 2020, p. 225)

Marx's assertion that religion was the opium of the people implied that religion was invented by humans in order to give life meaning; Susskind argues that work has now become the opium of the people. He quotes John Maynard Keynes in order to articulate the predicament: 'The worry, as Keynes put it, is that "there is no country and no people, I think, who can look forward to the age of leisure and of abundance without a dread. For we have been trained too long to strive and not to enjoy"' (Susskind, 2020, p. 225). Leisure, then, is a problem, for we no longer know how to be in leisure. Susskind asserts that, just as we have unemployment policies, we ought to consider having leisure policies. Such policies would imply revisiting education and imagining educational institutions that train students not for work but for human flourishing in a world with less work. We need to seriously consider policies that shape how people spend their leisure hours. Initiatives to shape how we spend our time already exist, not in official policies but, for example, in contexts of volunteer work. In a world with less work '[s]ocieties will need to think far more deliberately, comprehensively, and coherently about their leisure policy' (Susskind, 2020, p. 229).

If we entertain the notion that the effects of loss of work could be remedied by leisure policies, then we must also seek to clarify the difference between work and leisure. One could make this task simple by stipulating that work represents tasks that are necessary and/or salaried, and leisure represents all kinds of activities that do not fit in the category of work. But this is surely not

[25] A punishment to the effect that '[t]he gods keep the means of life concealed from human beings. Otherwise you would easily be able to work in just one day so as to have enough for a whole year even without working' (Hesiod *et al.*, 2006, p. 91).

what Susskind has in mind, nor does it correspond to the classical notion of leisure. Leisure is not supposed to represent every conceivable activity outside the frame of work. Susskind sees a possible model for how purposeful and meaningful leisure could be conceived in charitable undertakings. The kind of leisure enjoyed by figures such as Plato and Aristotle could be described as an ascetic mode of being in which certain difficult practices, such as geometry and philosophy, are learnt and practised for the purpose of attaining wisdom. In neither of these instances does leisure represent a stage in which the person in leisure ceases to be of use to his or her community; rather, it represents a stage in which the person in leisure continues to be of use, but on a different level.

If we seek to find a place for leisure in our own time, there may be no compelling reason why we should endeavour to impose standards held in classical antiquity. Perhaps, as Susskind suggests, the way in which charitable activities are undertaken today could provide a frame for how leisure could be conceived in the age to come. However, in classical antiquity there is a clue that the implications of work-related challenges might turn out to be rather different from how they have been described so far. Plato's Socrates, at the height of his leisurely and discursive activities, is frequently preoccupied with considerations of the so-called *techne*, which we can translate as art or professional skill. Each art, in the context in which it is practised, has its proper purpose and requisite skills; but each art is also situated in a larger social order, within which it can be understood to be serving a higher social purpose. Plato's dialogues can be interpreted as telling of a leisurely culture that does not outright despise work, as previously alleged, but one that values the knowledge and skills that pertain to the arts practised in the context of each profession, at least to the extent that the professions contribute to the well-being of the whole. Among other arts, those of medicine, carpentry, house-building, and piloting a ship are cited as paradigmatic examples of practices that must be understood as useful and valuable in the context of the function that each art fulfils in the larger whole. The society that emerges through such a reading of Plato is one founded on valued knowers and doers: from the shoemaker to the philosopher, every free member of such a society receives his or her lot of purpose, meaning, and dignity. Ideally, the entire society would be composed of skilled knowers and doers on every hierarchical level.

Susskind is primarily preoccupied with what humans are to do in contexts where there is less work. In many of the pages that follow we are preoccupied with the potential for a loss of critical knowledge and skills; this loss, it is argued, is inextricably intertwined with the loss of purpose and meaning, the loss of use for human beings, and the loss of cohesion and autonomy of traditional or conventional social structures.

In order to form a preliminary idea for how such loss could occur, we can return to an earlier quote from Susskind: '[I]f you come across a task where it is easy to define the goal, straightforward to tell whether that goal has been

achieved, and there are lots of data for the machine to learn from, then that task can probably be automated' (Susskind, 2020, p. 77). To this we could add that the word 'probably' most likely has a social significance that reaches beyond any probability estimate. Even if a task is not presently automated, awareness of probable future automations decreases the incentive for people to undergo demanding trainings that would enable them to perform the soon-to-be-automated tasks.[26] This could result in a lack of the professionals who would be needed to complete critical tasks, which in turn would increase the incentive to automate those tasks. As such dynamics take hold, there is a risk of 'knowledge deficits' that temporarily impede the functionality and/or progress of technologically advanced societies. We consider this in more detail in part II with the help of Ivan Illich and other thinkers.

In addition to the risk of harmful knowledge deficits, we must also consider the risk that automation will fall short of the desired results and/or produce other unforeseen results. These risks combined imply that the reconstruction of society as an algorithm-intense automation society could be likened to a wager on an irreversible process. If at some future stage we discovered that full-scale automation had generated unsatisfactory conditions, then critical human skills to sustain society without the automations already in place would likely no longer exist. Future generations may, then, not be empowered but rather disempowered by current innovations. If this were to turn out to be the case, then all hype and over-selling of new technologies would acquire a particularly nefarious patina: it would represent the gloss by which we are lured to surrender our 'native' powers.

It is perhaps premature to raise these matters at this early stage, as they are drawn from analyses that are undertaken later. Nevertheless, they are raised in order to make the challenge of a world with less work more relevant to the questions that are to be answered. How would algorithm-intense environments affect the exercise of human agency? If we primarily saw the disappearance of salaried work-tasks as the main challenge to be dealt with, then perhaps leisure policies could provide a remedy to the problem.[27] If the disappearance of salaried work-tasks also implied loss of incentives to learn critical skills, would it then also be conceivable that leisure policies could provide the incentive for individuals to undergo long and difficult training in order to learn

[26] The disincentive may strengthen to the extent that we are aware of the ongoing 24/7 algorithmic training that takes place all over the technological world in air-conditioned computer centres in order to produce artificial agent-like systems that are able to outperform us in every conceivable task. I, the author, am sometimes barely able to get up in the morning. What are my prospects in the face of such razor-sharp clockwork efficiency? Why should I undergo any difficult training at all?

[27] We should not expect, of course, that algorithmic technologies will only affect work-related tasks. It is perfectly possible that the impression will be generated that many tasks that could be imagined under the heading of 'leisurely activities' could also conceivably be more efficiently done by means of AI. If this were to turn out to be the case, what then would be the incentive to undergo burdensome training in order to learn critical skills?

corresponding levels of critical skills? This, it seems, would require the imposition of something like the ascetism inherent in the ways of leisure practised by Plato and Aristotle. In our days, such a moral framework imposed by the state could rightly be interpreted as utopian or totalitarian. Or would most skills simply cease to be critical, and thereby become superfluous? This, on the other hand, seems to imply a weakening of the human being, a debilitating process.

In conclusion, to the extent that the near future reflects current trends, we could expect more and more automation of work-tasks. In some instances, complementing effects may produce work for humans in new domains. The scenario described by Susskind is accepted as one plausible future development. In parts II and III we explore some of the potential wider implications of such developments as they pertain to human autonomy and worldviews.

A greater danger has been evoked, one that implies a loss of the incentive to acquire critical skills, which in turn may generate contexts in which humans not only become useless vis-à-vis a workforce, but also increasingly useless vis-à-vis one another. Some undesirable effects of such uselessness might be mitigated by the availability of AI – versions of algorithmic technologies that will know us better than we know ourselves, and that will be of use to us to a degree that we could not be in relation to one another. Such developments would not necessarily imply that future AIs would know us better than we tend to know ourselves at present; that they may, but owing to a potential lack of incentives to make an effort, the standard of our present general understanding may also deteriorate, with the effect that, in the future, a fairly dull AI may come to know us better than we then may know ourselves.

Techno-social dynamics may also evolve in other directions. Two more paradigmatic cases for an algorithm-intense future remain to be considered. Next we consider the notion of human–AI partnership.

Human–AI partnership

In *The Age of AI and Our Human Future*, Henry Kissinger, Eric Schmidt, and Daniel Huttenlocher attempt to find a role for humans in increasingly AI-intense environments. The authors see both great potential benefits and risks with the onset of what they call ‘the age of AI’. One of the properties of AIs that are discussed is their ability to perceive or access aspects of reality that humans cannot apprehend. According to the authors, this ‘portends progress towards the essence of things – progress that philosophers, theologians, and scientists have sought, with partial success, for millennia’ (Kissinger, Schmidt and Huttenlocher, 2021, p. 14). On the horizon of potential exploration are aspects of reality that no human has known until the present moment. We already experience that, through knowledge embedded in software, we are able to do all sorts of things that previous generations could not have even

imagined. But further embeddedness of knowledge in systems that become opaque to humans, the authors say, faces us with a dilemma:

> [w]hen a human-designed software program, carrying out an objective assigned by its programmers – correcting bugs in software or refining the mechanism of self-driving vehicles – learns and applies a model that no human recognizes or could understand, are we advancing toward more knowledge? Or is knowledge receding from us? (Kissinger *et al.*, 2021, p. 17)

The authors ask many important questions: 'Who operates and defines limits on these processes? What impact might they have on social norms and institutions? And who, if anyone, has access to what AI perceives?' (Kissinger *et al.*, 2021, p. 109).

In order to consider these and other questions asked by the authors, it could be useful first to consider briefly how the relationship between humans and technologies in general could be understood.[28] As long as we perceive humans as being in control over instrumental technologies, we could conceive of technologies as mere tools of autonomous human agents. Tools are then simply used as instruments by humans in order to accomplish goals. Conventional automation technologies could also be understood as controlled by human overseers. On this view, both cases enable human-controlled actions. One could say that technologies, viewed from this angle, extend the reach and power of human agency. However, there are many other ways of understanding technologies in general – for example, in analogy with prosthetic limbs and as extensions of the nervous system.[29] How appropriate is the purely instrumental understanding in relation to algorithmic technologies? What does it help us to see and what does it leave out?

If we were to deal with a technology that was only able to *access* aspects of reality unknown to us, then the findings, in order to become meaningful and useful, would have to be interpreted by humans and integrated into human-curated knowledge. Here we could think of Pasquinelli's analogy with optical media, for the telescope and the microscope are examples of such technologies. In the case of many algorithmic technologies, the analogy seems misleading. We are confronted with something of a different order: forms of automation that are able, to varying degrees, both to *perceive* and to *act*, and that, in more and more domains, exceed capacities that humans can attain.

That which AIs perceive is typically not merely extracted by humans and integrated into human-curated knowledge; instead it is integrated into artificial and/or human–artificial hybrid systems that can act with varying degrees of autonomy. We are confronted with systems that are novel in structure and that display a form of collective agency that, to a large extent, is animated by

[28] This is explored in more depth in part II.

[29] For the prosthetic understanding of technology, see Norbert Wiener (1990, p. 73-76); for technology as extension of the nervous system, see Marshall McLuhan (1994).

technological executive functions. Conventional automation is designed to act automatically in narrowly predetermined ways when human controllers give the requisite commands. In the case of advanced algorithmic technologies, an additional layer – or rather, multiple additional executive layers – are added between the task that is to be accomplished and any humans who are supposedly still in control of the technology. Moreover, in contrast to conventional automation technologies, the effects of current algorithmic technologies are becoming less and less narrow and more and more unpredictable.

According to Russell, '[t]he central concept of modern AI is the *intelligent agent* – something that perceives and acts' (Russell, 2020, p. 42). AI automation, then, begins to look like something akin to *automated* or *artificial agency*. If conventional automation can perform tasks – more speedily, more powerfully, and more efficiently – in ways that exceed the degrees that humans are able to attain by means of manual techniques, AI automation can perceive aspects of reality that no human agent can perceive, and, in addition to this, act with varying degrees of autonomy. Much of the revolutionary potential of algorithmic technologies seems to follow from the integration of this dual capacity in one and the same type of system.

AIs, like all technologies, do indeed gift their users with new powers. But they also redefine the playing field. Neil Postman likens the acceptance of new technologies to a Faustian bargain, for 'technology giveth and technology taketh away' (Postman, 1993, p. 5). What is it that AIs might take away?[30] Kissinger, *et al*. have already intimated that something of fundamental importance might begin to recede from us. Before we know it, in one field after another, artificial systems will be accomplishing tasks in ways that in various respects exceed human capabilities. As tasks and arenas operated efficiently by AIs change from 'winning games on chessboards' to 'supplanting functions of humans' in 'social arenas', it seems likely that knowledge will begin to recede from us. What knowledge, then, might be lost? Trivial knowledge? Or knowledge of how basic practices are to be done – practices that structure and maintain the environments in which we live, knowledge that pertains to core features of our life-milieus? These questions are explored from various angles in parts II and III. The remainder of this section focuses on the future role of humans as envisaged by the authors referred to below.

Kissinger *et al*. argue that we are on the verge of a tipping point, and that the age when human reason guided society may already be behind us: '[O]nce machines approximating human intelligence are regarded as key to producing better and faster results, reason alone may come to seem archaic. After

[30] 'What is it that AIs might take away?' is the simple, and perhaps even simplistic question. It nevertheless has its place within the context of analyses concerned with what AIs might *do* to us. The framework that will be elaborated in the treatise focuses more on the situatedness of human beings in larger wholes. To be sure, some individual AI systems may affect humanity inordinately. However, the more holistic revolution is actuated through increasing human embeddedness, via ubiquitous computing, in algorithm-intense environments.

defining an epoch, the exercise of individual human reason may find its significance altered' (Kissinger *et al.*, 2021, p. 207). This amounts to a weighty challenge to human self-understanding. For the first time there will be agent-like units and systems among us that, in many respects, will be able to do things better and more efficiently than humans. The authors fear the impact that this might have on humans:

> [s]ome may be tempted to treat AI's pronouncements as quasi-divine judgements. Such impulses, though misguided, do not lack sense. In a world in which an intelligence beyond one's comprehension or control draws conclusions that are useful but alien, is it foolish to defer to its judgements? Spurred by this logic, re-enchantment of the world may ensue, in which AIs are relied upon for oracular pronouncements to which some humans defer without question. (Kissinger *et al.*, 2021, p. 210)

Are we, then, moving towards a one-sided dependency in which humans will not be merely redundant but also mentally subject to a 'higher' artificial agent-like power? Not if the vision of these authors is fulfilled. Kissinger *et al*. see three primary options that are available to us. They reject deference to AI – that is, subjection. They also reject the confinement of AI. Instead, they envisage a constructive role for humans in partnership with AI.

We are accustomed to humans making decisions, and we are becoming accustomed to machines making decisions. The onset of the age of AI seems impossible to stop, for '[o]nce AI's performance outstrips that of humans for a given task, failing to apply that AI, at least as an adjunct to human efforts may appear increasingly as perverse or even negligent' (Kissinger *et al.*, 2021, p. 23). Somehow, in order to avoid deference and rejection, we must cultivate a mindset in which we become accustomed to humans and machines making decisions together. This, it seems, is already occurring:

> AI is also in the process of transforming machines – which, until now, have been our tools – into our partners. We will begin to give AI fewer specific instructions about how exactly to achieve the goals we assign it. Much more frequently, we will present AI with ambiguous goals and ask: "How, based on *your* conclusions, should we proceed?" (Kissinger *et al.*, 2021, p. 20)

The authors envisage a future in which all kinds of creative tasks, from writing songs to discovering medical treatments, are done in collaboration with machines. The revolutionary change has, in fact, already begun. The authors provide numerous examples. The shift from atlases to online navigation services shapes contexts in which 'the individual using such a service is not driving alone; instead he or she is part of a system in which human and machine intelligence are cooperating to guide an aggregation of people through their individual routes' (Kissinger *et al.*, 2021, p. 108). Individuals using such systems are 'entering into a form of human–machine dialogue that has never

before existed' (Kissinger *et al.*, 2021, p. 109). In return they get unprecedented conveniences and gain new capabilities. But such dialogues also 'have the capacity to shape human activity in ways that may not be clearly understood – or are even clearly definable or expressible – by human users' (Kissinger *et al.*, 2021, p. 109).

If the role for humans envisaged by the authors became a reality, some of the challenges discussed by Susskind could be overcome. There would still be important tasks for humans to do, and thus there might not be any experienced loss of purpose and meaning from a scarcity of work. Humans in an AI-intense civilisation could even experience a significant expansion of knowledge and power. If so, then this expansion would not be the fruit of a conventional relationship between human beings and technology; it would be the fruit of a relationship between human beings and a new type of partner that should be understood – if we follow these writers – not in analogy with previous technologies, but in analogy with previous kinds of human partners. Technical know-how would have achieved a leap enabling technological civilisation to replicate social relationships by means of artefacts.

Kissinger *et al.* mention online navigation systems as paradigmatic examples of budding partnerships. Interaction with certain search engines, AI-assistants, and chatbots could also be understood as instances of budding partnerships. In all of these examples we could see utility for the humans – partners or users – who interact with the technologies. But if we choose to view such interactions through the analogy of partnership, then we should take a moment to consider the purpose and significance of partnerships in more conventional contexts.

Partnerships can be formed for all kinds of purposes. The significance of a partnership can be evaluated with a view to its efficiency in accomplishing results that are in the interest of the partners. We could say that a collaboration that benefits the interests of all partners is significant independently of how the collaboration affects people who are not included in the partnership. Many would hesitate to label a collaboration that serves the interests of one collaborator but that does not at all serve the interests of other collaborators 'a partnership'. But how should we think of collaborative interactions where benefits and inconveniences are more difficult to discern? How, for instance, should we think of a relational interaction that benefits the important long-term interests of one and satisfies the short-term cravings of another?

In the case of online navigation systems, search engines, and chatbots, the human interactor is presented with energy- and time-saving shortcuts to realising goals. It is not suggested that there is anything wrong with this in and of itself. On the contrary, this is a fairly general feature of many technologies. However, if we were to construct environments in which just about all demanding tasks could be outsourced and there were no longer any need to make any effort to learn any skills, then we would certainly find ourselves in unknown territory. Nevertheless, if we consented to think of these kinds of

relationship as forms of partnership – relationships in which the human interactor routinely saves energy by outsourcing tasks that would otherwise require learning critical skills to machine-systems – then how should we evaluate the significance of such partnerships? Moreover, are there, somewhere in the technical hierarchies with which we mere users interact, other partnerships with which we could compare the significance of our purported partnerships? Would the envisaged kinds of partnership issue in an egalitarian society? These lines of inquiry are explored further in part II.

In conclusion, Kissinger, Schmidt and Huttenlocher alert us to a risk already mentioned in the previous section, that knowledge might recede from us. They warn us against relying on AI as an oracle-giver. They go further, and argue that the human reason-guided society may already be behind us. They reject the options of deference to or containment of AI, and instead attempt to pave a way towards partnership with AI.

If the best-case scenario of the authors came to fruition, the danger of a knowledge-deficit would be overcome, and instead we would experience an impressive expansion of knowledge. Human beings would not be superfluous, but instead would experience their agency and power as radically enhanced through human–AI partnerships.

To the paradigmatic case of a world with less work, we can now add the paradigmatic case of human–AI partnership.

Increasing cognitive demands on humans

The third paradigmatic case points to developments that, in many respects, run opposite to Susskind's prognostics. The case is constructed on the basis of a few automation-related challenges that Norbert Wiener discussed early in the history of automation. It is possible to come up with methods to cope with, and perhaps overcome, some or all of these challenges. One endeavour undertaken by Stuart Russell to overcome *one* of the challenges is also briefly discussed. Still, the potential overcoming of *that* challenge in particular would be likely to generate new challenges, which in turn, it seems, would make this third paradigmatic case even more pressing. The case puts the position of the human agent at the centre of an increasingly complex environment. The increasing complexity of the environment implies changing requirements for what would count as *adequately informed* and *cognitively adequate* human agents.

Norbert Wiener identified a series of moral and technical problems that arise as consequences of automation. The ones considered here are the timing problem, the goal-intention problem, and the mechanical slave problem.

The timing problem manifests when two systems interact with each other, and one system operates on a much slower time scale than the other system. If we take human individuals to be the standard system, timing problems can manifest in two directions. When humans endeavour to understand processes

that unfold on time scales that are much slower than the ones experienced by individual humans, such as 'political history' or 'the development of science', then there is a timing problem in one direction. The other direction becomes apparent when humans interact with automation machinery, which usually operates on time scales that are much faster than the ones that individual humans experience (Wiener, 1999, p. 88).

In order to make machines perform in a way that aligns with human wishes, the machines must be subject to critical feedback by humans. Owing to time-scale differences, relevant critique may be delayed to the extent that it ceases to be relevant when it is provided:

> By the very slowness of our human actions, our effective control of our machines may be nullified. By the time we are able to react to information conveyed by our senses and stop the car we are driving, it may already have run head on into the wall. (Wiener, 1999, p. 81)

Making machine performance *align with human wishes or general intentions* can have a more general meaning than mere *optimisation of narrowly efficient machine performance*. From the more encompassing viewpoint, the timing problem applies on individual, social, and ecological levels. In all cases, human feedback is needed in order to ensure that machine performance align with short-, medium-, and long-term interests. On the social and ecological levels, it is often discovered, in retrospect, that machines have been too narrowly efficient at the expense of important long-term interests. When machines that are able to accomplish certain objectives at enormous speed are abruptly introduced into an environment, other processes in the larger ecology proceed as before on slow non-machine time scales. It may take a long time before undesirable social and ecological disturbances in the larger context are discovered.

How does the timing problem bear on AI? Today there is much discussion and research devoted to understanding the harmful effects of screen usage in, among other activities, learning and socialisation.[31] As is typical in instantiations of the timing problem, we have already constructed life-milieus in which screens have become omnipresent. It is not, however, the screens in and of themselves that operate on a faster time scale than is usually the case in conventional human contexts. Screens are merely windows into algorithmically enabled information spheres, with the effect that huge quantities of information become instantaneously available to the screen user. That which the screen enables represents a new type of life condition, a technology-enabled novelty, one that positions a species that evolved in hunter-gatherer contexts in uncharted territory. Further ahead we could speculate about analogous unforeseen effects engendered by chatbot usage: how would learning and

[31] See, for instance, Oswald (2020) *et al.* and Ophir, Rosenberg, and Tikochinski (2021).

socialisation be affected by human interaction with chatbots that deliver instant outputs on request – outputs that in many instances may be so qualitatively complex that a human would need years of specialised studies and training in order to produce equivalent results? These examples illustrate that, although we may well be under the impression of having control in a narrow sense over technologies operating on fast time scales in the present, such technologies could affect important slow-moving processes such as learning and socialisation; and as long as we have not understood how the former affect the latter, we are in fact applying the former to unpredictable long-term effects.

We can speculate, and speculation is all there is to it, the technological optimist might retort. The technological optimist would of course have a point, because, given the timing problem, there is no way that we can *know* in advance or even apace with product development and product implementation how things will evolve. Speculations can nevertheless be more or less reasonable and informed.

Regardless of how any single type of technology in the end affects the society and ecology into which it is introduced, the timing problem should receive the attention of any society that considers technological innovation a highly pursuit-worthy aim. If we carry on to apply newer and newer algorithmic technologies in our societies before we have adequately understood the impact of older architectures, any insights that we may gain from past experiences and from the study of existing architectures might become irrelevant before they can be of good use. If this increasingly becomes the case, then this would be not only because of the timing problem *per se*, but also because of what appears to be a prevalent cultural aim in technically advanced societies, namely, to accelerate the innovation and reconstruction of our bio-technical life-milieus. For Wiener, on the contrary, speed is not at all a virtue: 'To be effective in warding off disastrous consequences, our understanding of our man-made machines should in general develop *pari passu* with the performance of the machine' (Wiener, 1999, p. 81).

If Wiener is correct, then the non-observance of his advice should tend to generate results that may be experienced as efficient in the short term but that are likely to be experienced as dysfunctional in the medium and/or long term. It has already been argued that, in societies that automate fast, the incentive for humans to learn certain skills is likely to decrease, which in turn might result in knowledge deficits that temporarily frustrate the functionality of technological societies. This, as we see in more detail in part II with the help of Ivan Illich, is in turn likely to increase the incentive to speed up the automation of additional functions in order to compensate for the lack of human skills. But if Wiener is right, we then run the risk of increasing overall dysfunction. If we proceed in fast-forward mode, should we expect humans to become less and less compatible with technical systems with respect to being able to provide timely relevant feedback? Perhaps automated AI agencies could compensate for human slowness? On the other hand, the indiscriminate application of

AIs may also increase overall medium- and long-term dysfunction. The timing problem informs us that we would most likely need to wait a long time before the full range of implications of any one kind of complex algorithmic technology, let alone whole ranges of algorithmic technologies, became manifest to us humans, who experience human time scales.

Wiener also called attention to the goal-intention problem. As long as machines operate in sealed and abstract contexts, it may be fairly unproblematic to give them goals. For instance, there is no immediate societal risk in giving a chess-playing machine the goal of victory. Once machines are applied in social environments, social risks must be taken into consideration. Wiener points out that the instructions that we tend to give, if taken literally, rarely correspond to our more precise intentions. If you ask a waiter, 'Please give me a cup of coffee!', you no doubt do not intend to instruct the waiter simply to grab a cup that has already been served to the table next to you and transfer it to your table. Yet that course of action may appear to the system concerned to be the most efficient goal-realising procedure. Formally vague but culturally adequate instructions generally function in social contexts where humans have learnt to integrate them into procedural hierarchies for how things ought to be done. They become less reliable in inter-cultural interaction. Machines initially lack all the sets of socialised norms that we tend to take for granted and that enable us to interpret instructions adequately. To illustrate this point, Wiener uses the example of giving a machine the goal of victory in a war game: '[I]f the rules for victory in a war game do not correspond to what we actually wish for our country, it is more than likely that such a machine may produce a policy which would win a nominal victory on points at the cost of every interest we have at heart, even that of national survival' (Wiener, 1999, p. 86).

Thus, to the timing problem, which arises from inter-systemic time scale differences, we must add the goal-intention problem, which bears on decisions that may be intra-systemically logical in view of the realisation of any goal given, but where the goal-as-it-has-been-literally-given does not correspond to the more general intentions of the human goal-giver. The novelty in the domain of human interaction, the machine, is unlike humans, even humans from unfamiliar cultures – it is unlike all biological organisms – in that it operates in a mode defined by formal logic. Presently, learning machines are achieving impressive results in replicating human social behaviour. It remains to be seen how far this replication can be taken. Perhaps machines will soon be able to understand humans on a level that equals humans' understanding of humans. If this were to become the case, then we could say that computer scientists would have found a technical solution that largely solved the goal-intention problem as it pertains specifically and narrowly to human–machine interactions. The goal-intention problem, however, is applicable on a much wider scale. In part III it is argued that something like the goal-intention problem permeates all kinds of processes involved in the construction of human

worlds. Another way to put this is that intentions never correspond precisely to effects.

The third problem raised by Wiener focuses our attention more squarely on the humans who interact with machines. We have learnt that Russell thinks that most humans, for most of history, have been used as robots to perform repetitive tasks. Many will no doubt be inclined to agree. To the extent that this is correct, the institution of slavery should be an especially apt embodiment of this tendency. In such contexts, automation can appear as an emancipatory blessing for oppressed workers. Wiener identifies a contradiction, a non-moral problem, in the context of human reliance on human slaves – one that would be equally relevant in the context of human reliance on intelligent machines:

> We wish our slave to be intelligent, to be able to assist us in the carrying out of our tasks. However, we also wish him to be subservient. Complete subservience and complete intelligence do not go together. How often in ancient times the clever Greek philosopher slave of a less intelligent Roman slaveholder must have dominated the actions of his master rather than obeyed his wishes! Similarly, if the machines become more and more efficient and operate at a higher and higher psychological level, the catastrophe foreseen by Butler of the dominance of the machine comes nearer and nearer.'[32] (Wiener, 1999, p. 86)

Kissinger *et al*. seek to avoid human subservience to AI. The analogy proposed by Wiener evokes a paradigmatic case of knowledge receding from the master, producing subservience, in turn, to the agent relied on. We should at least note that, in the novel forms of interaction discussed, the status of the human is threatened not only by potentially harmful acts undertaken by machines, but also by human *reliance* on machines. As humans increasingly rely on intelligent machines, as in the case of the Romans relying on educated Greek slaves, skills and knowledge about how things are done will likely begin to atrophy, as a result producing contexts not only of reliance but also of *dependence* on machines. This, it may be objected, has already been the case for a long, long time, from human reliance on windmills to contemporary cars. Still, it must be maintained that reliance on a type of machinery that could be understood as being in the process of attaining higher and higher degrees of agency and autonomy is of a different order.

An agent that has all things done for it by means of magical agencies may well be under the impression that it has little need for skill or knowledge beyond that which is required to give orders. If this were the case, we could all become prompting engineers. But if we follow this way, are we following a way of empowerment or of disempowerment? To paraphrase C. S. Lewis: If I

[32] The writer referred to is without doubt the British writer Samuel Butler. In an 1863 article titled 'Darwin among the Machines', Butler considers the possibility that machines could represent a type of mechanical life that in time may supplant humans as the dominant species.

manage to prompt an algorithmic system to present a good solution to a problem, this does not make me an intelligent or wise man. If, owing to the availability of algorithm-enabled shortcuts, I fail to develop critical knowledge and skills, then I put myself at the mercy of the technical systems that manifest as shortcuts to goal-realisations.[33] When such systems fail to perform adequately, or when such systems are deployed against our own interests, we will have little choice but to accept subservient positions vis-à-vis the systems, for we shall no longer know how to navigate this world except through their mediation. In relation to a machinery previously conceived of as tool or aid, we will now be in a relationship that Edward Ashford Lee labels 'obligate symbiosis': and to pull the plug is not an option (Lee, 2019, p. xiii).

Wiener presciently argues that contexts in which more and more tasks are accomplished for us by mechanical automation – or mechanical slaves – will require more, not less, effort from humans.

> [T]he future offers very little hope for those who expect that our new mechanical slaves will offer us a world in which we may rest from thinking. Help us they may, but at the cost of supreme demands upon our honesty and our intelligence. The world of the future will be an ever more demanding struggle against the limitations of our intelligence, not a comfortable hammock in which we can lie down to be waited upon by our robot slaves. (Wiener, 1990, p. 69)

If we wish to be able to maintain some degree of human autonomy in algorithm-intense environments, then, because of the timing problem, the goal-intention problem, and the mechanical slave problem, it seems that we must somehow exceed our cognitive limitations.

In considering Wiener's problems, we are led towards conclusions that appear to be opposite to those for which Susskind argues. Susskind admits the likelihood that there will be more work-tasks in the AI-intense contexts of tomorrow. Humans, however, will not be suited to perform most of the new work-tasks that will be created. Susskind's vision for the future implies a surrender, at least on the part of most humans, of human control and stewardship to algorithmic control. Susskind is cautiously optimistic about the prospects for human flourishing in such future contexts. Wiener warns us that such surrender may bring disastrous consequences. Contrary to Susskind, Wiener envisages contexts in which human cognitive capabilities must be continually raised in order to meet new and ever more demanding technical requirements. We appear to be confronted with two conflicting views. But perhaps the two could meet. Susskind's analysis is not strictly incompatible with that of Wiener. Susskind argues that the world of tomorrow will be a world with less work; he does not argue that the cognitive requirements of *all* of those who

[33] Lewis writes: 'If I pay you to carry me, I am not therefore myself a strong man' (Lewis, 2006, p. 54). The context from which the quote is taken is quoted in full and discussed in part III.

still work will be less. However, Wiener does not seem to think that it would be desirable for those who are out of work to surrender control to mechanical slaves. In part II we explore further the ambiguous interests that such mechanical slaves may serve.

Perhaps the challenges raised here could be overcome. The idea of flourishing human–AI partnerships hinges on their being overcome to a large extent. To the extent that they are not, human–AI partnerships risk turning into either dysfunction or relationships that fit the model of the Roman-Greek paradigm that Wiener criticises, or a combination of the two. To the extent that the goal-intention problem is not overcome, the relationship is likely to produce dysfunction. Meanwhile, it is difficult to see how the timing and mechanical slave problems could be 'overcome': by their very nature, future AIs will be able to operate on time scales and manifest knowledge and skills that exceed by far the capabilities of any human partner. The enabling of partnerships that human partners experience as genuinely functional and productive will require, at a minimum, much effort in adapting algorithmic systems to the way in which humans perceive, experience, and interpret themselves in their lifeworlds. It will very likely also require adaptation of humans to the way in which algorithmic systems will operate.

Next, we briefly consider how a computer scientist grapples with some of Wiener's problems.

User-friendly AI – overcoming the goal-intention problem?

Stuart Russell takes some of Wiener's apprehensions into account, emphasising the alterity of the new type of agency that is in the process of being constructed. Russell compares an analogy made by Wiener with his own 'King Midas problem'. In the former, Wiener understands humans and intelligent machines in analogy with how magic functions in Goethe's tale *The Sorcerer's Apprentice*, in which a broom is given the instruction to fetch water. In Russell's own illustration of the same problem, a legendary king in Greek mythology, King Midas, has his wish granted. In both cases, the humans involved ignore the need to provide multiple caveats along with their instructions. Disaster follows. The broom does not know how much water to bring or when to stop, and everything that King Midas touches, including food, drink, and family members, turns into gold. The analogies warn us that technologies will do what we instruct them to do, which will often turn out to be different from what we intend them to do. The risk is that machines will pursue incorrect objectives.

The problem is rendered more complex when we face systems that, like humans, are able to invent new goals that are instrumentally useful in the pursuit of their main goals. Russell refers to such goals as 'subgoals'. Suppose that we have managed to give a machine a fairly non-controversial objective, such as fetching coffee at opportune times and in adequate amounts. We have

satisfactorily addressed all of the pitfalls alluded to in *The Sorcerer's Apprentice* and the King Midas problem. Still, the danger is not over. If a machine:

> is sufficiently intelligent, it will certainly understand that it will fail in its objective if it is switched off before completing its mission. Thus, the objective of fetching coffee creates, as a necessary subgoal, the objective of disabling the off-switch. The same is true for curing cancer or calculating the digits of pi. There's really not a lot you can do once you're dead, so we can expect AI systems to act pre-emptively to preserve their own existence, given more or less *any* definite objective. (Russell, 2020, p. 141)

The example illustrates how Wiener's problems become more intricate in systems that are capable of ever-higher degrees of instrumentally intelligent initiatives. Russell argues that 'self-preservation' does not need to be built into the machine, because self-preservation is in itself instrumentally useful to the realisation of most of the original goals that we are likely to give to AIs. Moreover, there is no end to other subgoals that may be instrumentally useful in the realisation of all kinds of task. It would be exceedingly difficult to take all possible subgoals into account. Having access to money is important in society: 'Thus, an intelligent machine might want money, not because it's greedy but because money is useful for achieving all sorts of goals' (Russell, 2020, p. 141). If the goal is to find a cure for cancer: what is to stop it from engaging in a subroutine so that it induces 'multiple tumors of different kinds in every living human being so as to carry out medical trials of these compounds, this being the fastest way to find a cure' (Russell, 2020, p. 138)? In part II we see that the dynamics of the patterns exemplified above are, to some extent, already expressed in existing preference-satisfying algorithms.[34]

In the exemplified scenarios, then, we need to pay attention to potentially destructive subgoals. When a system is instructed to pursue a goal, new subgoals can be invented in order to achieve the goal. But keeping track of all potential dangers is impossible. Therefore, according to Russell, this 'standard model – whereby humans attempt to imbue machines with their own purposes – is destined to fail' (Russell, 2020, p. 137). Russell instead argues in favour of what he calls human-compatible and user-friendly AI.

How could we make machines that are useful in fulfilling our *intended* objectives, as opposed to the *narrow* instructions that we give them? Russell suggests three principles on the basis of which such machines may be engineered: 1) The machine's only objective is to maximise the realisation of human preferences; 2) the machine is initially uncertain about what those

[34] Such preference-satisfying algorithms are set to satisfy human preferences. In order to do so they must be able to predict human preferences. Human preferences become more predictable if the algorithm can make them more predictable: the subgoals consist of behavioural modification routines with the effect that human preferences and behaviour become more predictable.

preferences are; and 3) the ultimate source of information about human preferences is human behaviour (Russell, 2020, p. 173).

User-friendly AI is to be engineered so that it works to satisfy utilities (e.g., preferences) based on probabilities. Russell advocates building ignorance into probabilistic systems that are set to satisfy utilities.[35] In the words of Russell:

> Instead of a goal, then, we could use a utility function to describe the desirability of different outcomes or sequences of states. Often, the utility of a sequence of states is expressed as a sum of *rewards* for each of the states in the sequence. Given a purpose defined by a utility or reward function, the machine aims to produce behaviour that maximizes its expected utility or expected sum of rewards, averaged over the possible outcomes weighted by their probabilities. (Russell, 2020, pp. 52–53)

The second of the three principles that Russell advocates, *uncertainty*, is to provide a key feature in user-friendly AI. Inbuilt uncertainty, according to Russell, will motivate AIs to allow themselves to be shut off rather than obstinately to pursue given goals at any cost (Russell, 2020, p. 176). This, to allude to a notion previously used, might render AIs less monomaniacal. This feature would also no doubt differentiate future user-friendly AI from the preference-satisfying systems discussed in part II.

Utility-maximising probabilistic networks with inbuilt assumptions of uncertainty and incompleteness may constitute important steps in the engineering of environmental contexts that enable some form of the human–machine partnership envisaged by Kissinger *et al*. Russell and other researchers are aware, of course, that this solution would not solve everything. It might mitigate the risks of AI takeover and of human deference to AI. Owners and operators of AIs may still exercise deliberately nefarious influence over human populations. Human use of AIs may still produce unforeseen and undesirable effects on human preferences and decisions.

The third principle that Russell advocates is that the ultimate source of information about human preferences should be human behaviour. What does this tell us? Some forms of human behaviour, such as rape and murder, are classified as reprehensible by general and universal consensus. Such egregious types of behaviour could no doubt be eliminated from the sources of behavioural preferences from which user-friendly AI draws in order to help us satisfy ours. How about types of behaviour that cannot be classified as being outright and universally moral or immoral, but that may be problematic in the context of the life-experience of individual persons and/or in the context of specific cultures? One person's virtue may be another person's vice, just as that which is deemed a vice in one culture may be deemed a virtue in another.

[35] The logical and technical specifics involving the difference between Boolean and Bayesian systems for how this is to be done are too complex to motivate their inclusion in this treatise. The reader can consult Russell (2020, p. 48-51).

It is also the case that the behaviours of a person or a group of persons do not necessarily correspond to *higher* preferences. Behaviour may also reflect lower cravings. Surely we ought not to welcome the prospect of vice-reinforcing machines in our midst. But then that which one considers a vice, another might consider a virtue. The normative complexity of these dynamics is wide and deep. What potential gains and losses may we expect if we introduced the kind of machine that Russell labels 'user-friendly AI' in the midst of this normative complexity?

We are discovering that, although there may be feasible technical solutions to the problems raised by Wiener, the extent to which solutions are in fact decisive will rarely be immediately clear. 'Solutions' may turn out merely to recalibrate an already complex problem-landscape. And we may have to wait for a long time before we are in a position to make any general assessment of the result of innovative social experimentations.

The sheer complexity that needs to be dealt with in order to ensure that the AI-intense environments of tomorrow become hospitable for human beings appears to corroborate Wiener's intuition or hypothesis: the more we automate, the heavier will be the cognitive load on human beings; *this, at least, will be the case to the extent that we care about remaining in charge of things*.[36] This does not entail that automation will cease to proceed if the cognitive burden cannot be carried by automated society; if it does proceed notwithstanding, then, if we heed Wiener's warning, we should probably expect our life-milieus to become increasingly dysfunctional. We could hope that, somehow, AI would be able to carry this cognitive burden for us; but if we did, then we would risk getting drawn deeper into the kind of relationship categorised under the mechanical slave problem.

***

If we recognise that algorithmic technologies can be opaque and exhibit unpredictable behaviour, some of the preceding discussions require that we give our attention to the other side of the equation, to what is in many respects the even more opaque and unpredictable role of the human agent. It is possible to adopt a philosophical and abstract approach to AI, within which AI research could be understood primarily as a way to explore the ontological prerequisites for cognition, intelligence, and consciousness. In practice, however, most algorithmic technologies are conceived and designed in order to function under concrete circumstances and, ultimately, in human societies. When things go wrong, engineers routinely examine the technical systems involved in order to discover possible improvements. Frequently the human factor is also blamed. The remainder of part I deals with some aspects of the interaction

[36] And this, in turn, depends on the ideals we harbour of how human agents ought to exist in relation to their environments, which is the main subject of part III.

between human beings and technology. We will familiarise ourselves with a concept by means of which we could understand collaborative contexts between different – human and artificial – types of agent, the so-called multi-agent system. Finally, it is explained how the 'affordance' concept is used in the remainder of the treatise – that is, not only as an environmental agent-relative property that defines an agent's possibilities for action, but *as an agent-relative property that solicits and shapes behaviour*.

From simple to complex agents in multi-agent systems

Having considered the concepts of 'agency' and 'agent', and established that human beings represent a type of agency and agent, we now look at some possible features that 'artificial agents' have or could have. Assuming that humans in algorithm-intense environments will interact with algorithmic systems, this consideration becomes important when we seek to discern likely roles for human agents in algorithm-intense environments.

In discourses about AI it is common, explicitly or implicitly, to set up the human agent as a model criterion for successful artificial intelligence. Here, instead, we begin the consideration of artificial agents by looking to a considerably simpler biological model: the bacterium. Then we consider agentic properties that may apply to agents collaborating in multi-agent systems. Even now, before 'multi-agent system' has been defined, it may be appropriate to keep in mind that the complexity (or lack of it) of some specific individual agent does not determine the complexity of a multi-agent system as a whole.

Although we are primarily concerned with how algorithm-intense environments could affect humans *qua* agents, let us begin by considering a frightening scenario reminiscent of Nick Bostrom's famous paperclip scenario.[37] In contrast to Bostrom, who imagines his paperclip scenario as implemented by a superintelligence, the frightening scenario we are about to consider would be implemented by rather 'stupid' artificial agents.

Much has been written about the threats that may emerge in the context of a so-called intelligence explosion.[38] The obsolescence of human beings, transhumanist paths to salvation, and the annihilation of human civilisation are becoming familiar scenarios that are sketched in connection with a potential machine takeover. In academia and in popular culture alike, the risk is almost always imagined as consequent upon some ontologically brilliant form of artificial intelligence; rarely is it imagined to emerge from forms of artificial systems that would appear to humans as trivial, dull, or even stupid. How may

[37] Bostrom writes: 'This could result, to return to the earlier example, in a superintelligence whose top goal is the manufacturing of paperclips, with the consequence that it starts transforming first all of earth and then increasing portions of space into paperclip manufacturing facilities' (Bostrom, 2003).

[38] See, for instance, Nick Bostrom (2014), Ray Kurzweil (2018), and Max Tegmark (2018).

the introduction of very simple and, on the face of it, stupid but nonetheless goal-efficient artificial systems affect the ways of life of humans?

On the basis of how risks are structured in the biosphere, Ali Hossaini argues that the emergence of artificial agency might constitute the crucial turning point. But this emergence is understood as likely to occur *from below*, in analogy with how agency evolves in the biosphere, rather than as the result of a sudden takeover by an engineered superintelligence. 'Agent' is modelled here on the biological agent; it encompasses bacteria but not viruses. For Hossaini, 'agency' means 'the capacity to make independent, self-interested decisions' (Hossaini, 2019, p. 124). Hossaini takes for his model the basic unit of life, the cell. 'Organisms are intrinsically autonomous because their primary function is survival, and it is this imperative that produces hostility, docility and other behaviours associated with agency' (Hossaini, 2019, p. 127).

Hossaini's biological model leads him to emphasise threats that might arise from the instantiation of self-preserving and reproductive behaviour in artificial agents. Again, some of the least assuming biological agents – that is, bacteria – represent some of the most dangerous threats. Research in 'embodied cognition', Hossaini argues, reveal how machines, like biological agents, can be 'structurally coupled to their environments' (Hossaini, 2019, p. 127). Structural coupling does not require brilliant intelligences, as imagined by some futurists, but only the simple efficiency of bacteria. Machines that display patterns similar to those of biological agents could pose significant threats, even if they appear to be stupid, for 'biological adaptation operates in two directions. Over generations organisms adapt to their environment, but they also act to *adapt their environment*' (Hossaini, 2019, p. 128).

According to Hossaini, the prospects of ethical or moral AIs being developed are dubious, first because of humanity's conflicting beliefs, but more importantly because 'standards for ethical design miss a significant danger zone – they anthropomorphise rather than *biomorphise*. Dumb bacteria kill more people than smart bombs, and by focusing on intelligence rather than agency, we neglect the threat posed by biomorphic evolution' (Hossaini, 2019, p. 128). Hossaini leaves us with a frightening yet plausible-seeming scenario: What would happen if computer viruses began to mutate spontaneously? '[T]heir reproductive strategies could become dangerously unpredictable without a whit of intelligence' (Hossaini, 2019, p. 128).

Hossaini participates in discussions concerning major security threats, and argues that even very simple types of agent pose severe risks. If the public face of an AI apocalypse is represented by some vaguely superintelligent AI, then the public face of an AI utopia is somewhat similar. We are continually impressed by the latest cutting-edge chatbot because, when we interact with it, we get the impression, for a moment, of interacting with *one discrete agent* that approaches the ideal of human general intelligence. If it is a good idea to consider the risks associated with simpler types of artificial agent, then it is definitely also a good idea to consider the role of different types of agent,

including simple agents, in multi-agent systems. For that which appears to us as a discrete artificial agent can almost always also be described in terms of multi-agent system.

Catrin Misselhorn aims to provide her readers with the conceptual tools needed to understand different types of agent in multi-agent systems (MASs). She explains that:

> MAS[s] are systems composed of multiple interacting agents within an environment. There are biological, social as well as artificial multi-agent systems. One major advantage of MAS[s] is that they can solve problems that are difficult or impossible for an individual agent or a centralized system to solve. (Misselhorn, 2015, p. 3)

Imagine a comparatively simple cooperative body, consisting only of personal human agents, but agents who are nonetheless structured somewhat differently: a football team in which the individual members are of different cultures, speak different languages, and have learnt different play tactics. The coach's task is to train them so that they become able to cooperate and act as a unified team. We could include this team as an example of the category of MAS. Individual agents in a MAS can be much more differentiated than in the example given. In fact, the concept is used primarily in computer science as a way to describe how individual artificial and human agents can be integrated into systemic wholes. In the context of computer science, it falls to computer scientists to assume not only the role of the coach in relation to the football team, but also the role of the progenitors of the individual agents that compose the team. Misselhorn provides an overview of how MASs are conceived and how this relates to philosophical approaches to cooperation and collective agency.

In a MAS, agents cooperate in order to achieve results. The cooperation could be understood as collective agency, but, according to Misselhorn, 'the kinds of collective agency involved in different MAS[s] have so far not been sufficiently distinguished and investigated' (Misselhorn, 2015, p. 3).

In an artificial MAS, the agents that form the system will typically not have characteristics that define human agents. If the variety among biological and animal agents can be great, this can also be the case among artificial agents. In order to be able to differentiate between different types of agent, Misselhorn distinguishes between two dimensions of agency: 'autonomy' and 'intelligent behaviour'. Both autonomy and intelligent behaviour can vary in degree. To this Misselhorn adds that 'more demanding forms of intelligent behaviour are also coupled to more autonomy and vice versa' (Misselhorn, 2015, pp. 4–5).

In order to be able to conceptualise continua of autonomy and intelligence that can vary in degrees, it becomes important to grasp that which would count as minimally autonomous and minimally intelligent. Luciano Floridi and J. W. Sanders provide a definition that is adequate in the context of MASs:

> Autonomy means that the agent is able to change state without direct response to interaction: it can perform internal transitions to change its state. So the agent must have at least two states. This property imbues an agent with a certain degree of complexity and independence from its environment. (Floridi and Sanders, 2004, p. 357, quoted in Misselhorn, 2015, p. 5)

What counts as minimally intelligent? A general answer draws on the capacity for adaptivity and learning in the context of environmental interaction: 'An interactive agent takes input from its environment and brings about changes in the environment. The environment can be the *real world*, but it can also be a virtual environment like the Internet.[39] An interaction is minimally intelligent, if the agent's reaction to the input is appropriate with respect to the agent's survival' (Misselhorn, 2015, p. 5). More intelligent behaviour is associated with higher degrees of adaptivity and, therefore, is linked to the ability to learn. A more sophisticated agent, therefore, will be able to 'modify its reactions to improve its interaction with the environment such that they become more appropriate' (Misselhorn, 2015, p. 5) – more appropriate, that is, with a view to survival or self-maintenance and procreation or self-replication, as in the example with the bacteria, or with a view to any other goal.

In *Philosophy of Information*, Floridi adds an important caveat: definitions of agenthood only makes sense at prescribed levels of abstraction. At some levels of abstraction, even a rock could be understood as being interactive and agent-like. The same applies to certain software:

> If a piece of software that exhibits machine learning […] is studied at an LoA [level of abstraction] which registers its interactions with its environment, then the software will appear interactive, autonomous, and adaptive, i.e. to be an agent. However, if the program code is revealed, then the software is shown to be simply following rules and hence not to be adaptive. (Floridi, 2011, p. 3.4.1)

We cannot rule out the possibility that this may also be the case for humans. This means that it is possible that, on some level of abstraction, human beings will seem not to possess the features that would justify our attributing agency to human phenomena. This would not prove that human beings are not agents; it would merely indicate that, on certain levels of abstraction, there may be no compelling reasons to ascribe agency to human beings.

Floridi's philosophy of information, instead of reducing all phenomena to one level of abstraction, allows for emphasis on the dynamics between different levels of abstraction. That something is labelled as 'agent' on a given level of abstraction by no means implies that that something has obtained the degree of integrated agency that, on some other levels of abstraction, could be inferred from the behaviour of humans. It merely means that the 'something', on some specific level of abstraction, and with some degree of autonomy, is

[39] The 'real world', or part of the phenomenological experience of time and space through which humans have conventionally lived their lives.

able to display adaptive behaviour on the basis of environmental feedback and in view of achieving a goal.

The expression 'integrated agency' is used above because, considering the nature of MASs, it is not at all clear that human beings must necessarily be regarded as discrete agents. Human beings, composed of ordered structures of biological agents and complex neurological structures, could also be conceived of in terms of MASs. This is hardly a new insight. If we happen to be familiar with the cartesian trope of a mind controlling matter in time and space, this philosophical anthropology, conceiving of human beings as unified minds, has by no means been historically dominant among philosophers and theologians. Blaise Pascal, a contemporary of René Descartes, famously proclaimed that the heart has its reasons that reason cannot comprehend. A person's deliberative faculties may argue against the principles known by the heart, or even block the heart altogether and lend its services to principles alien to the heart. Here the philosophical anthropology is one according to which a person's faculties can be moored in the service of some first principles that are simply *known* by the heart; or they – a person's intelligence, deliberative faculties, desires and appetites – may begin to serve ends that are alien to the interests of the heart. In the first case we presumably approach that which we recognise as a wholesome person; in the latter we may eventually end up with a disjointed mess of a human being. We could also think of this in relation to integrated and disintegrated MASs, where the more integrated agencies begin to resemble that which we perceive as discrete unified agents.[40] If this is so, then this also has possible implications for how we can or ought to consider human beings interacting with artificial MASs. This is explored further in part II.

Other concepts used to differentiate different types of agent or stages of agency are *goal-directedness* and *intentionality*. Misselhorn explains that goal-directedness 'requires that agents try to achieve specific goals like getting the ripe fruit lying in the grass in front of them' (Misselhorn, 2015, p. 6).

[40] In *Pensées*, several of Pascal's fragments treat the aspect of the heart. See, for instance, fragment 101 (Pascal, 2022, p. 104). Modern psychology, pioneered by Sigmund Freud, traces the initiation of many acts to processes within the human being that are not fully conscious. Today Daniel Khaneman represents human cognition as 'System One' and 'System Two'. System One represents labour-saving cognitive mechanisms that could be described as intuitive, immediate and direct, whereas System Two represents labour-intense conscious thinking. Contrary to popular belief, Kahneman argues, our actions are predominantly influenced by the mode of System One (Kahneman, 2013). There is a long tradition among theologians of conceiving human beings in terms that can be translated to more-or-less disintegrated agencies; and this inclination can no doubt be traced as far back as the tradition that produced the biblical texts. For example, in the context of healing involving expulsion of a demon from a demon-possessed man, the following words are attributed to Jesus in Mathew 12:25-26: 'Every kingdom divided against itself is brought to desolation, and every city or house divided against itself will not stand'. Here, too, whether it concerns houses, cities, kingdoms, or human beings, health is understood in similar terms – that is, unification, for the sake of a purpose, of multiple parts. Using the vocabulary developed here, we could speak of it in terms of integrated agencies.

One can differentiate between *external* and *internal goals*. An external goal corresponds to something that is to be done in an arena; an internal goal concerns internal states that guide an agent's behaviour. Internal goals require mental states. The possession of mental states and internal goals 'leads to a new type of agents characterised by *intentionality* which makes for a qualitative difference in intelligent behaviour and autonomy' (Misselhorn, 2015, p. 6).

Intentionality is the capacity of something like a mind to be directed at things and to make inner representations of them. Intentionality is thought to be crucial for rational agency. Misselhorn cites two philosophical approaches to the explanation of rational actions. The first goes back to David Hume, and posits that intentional states are brought about by beliefs and desires. The other approach is represented by thinkers such as Michael Bratman (1984), and it posits that intentional action requires a 'specific kind of intentional state called "intention"' (Misselhorn, 2015, p. 6). The latter notion is argued from the point of view of the limitations of human agents: limited cognitive resources require that humans commit to a course of action; commitment requires planning; and planning requires sub-commitments. In other words: 'plans may include other plans: my plan to go for a drive may include a plan to find my car keys' (Misselhorn, 2015, p. 7).

Here a simple model for how intentional systems, such as human and artificial agents, could interact with environments is proposed. Intentionality presupposes the ability to have inner representations of exterior states of affairs. Inner representations can reflect the exterior states of things more or less accurately; but they can also suggest how the exterior states of things ought to be arranged. The 'ought' can be understood in two senses: in a moral sense, of course, but also in the sense of how things ought to be arranged in order to enable better the achievement of any goal – that is, in a sense of instrumental usefulness. *These two tensions – inner representation vs how things really are, and how things really are vs how they ought to be – provide dialectic grids that – somehow – enable minds and computational sets of algorithms to be directed at things or states of affairs in environments with a form of adaptive and dynamic intentionality.* This basic pattern is also integrated into the understanding of worldviews that is developed in part III.

An agent that operates in accordance with the model outlined above will not necessarily appear to be rational; it can even be completely dysfunctional; or, of course, it may appear to be irrational, but actually be very efficient in achieving its objectives. Misselhorn, nevertheless, cites two rationality constraints:

> First, intentions should be both internally coherent and coherent with the agent's other beliefs. Secondly, the agent should be able to monitor whether he or she was able to successfully realize his or her intention and to adapt his or her plans accordingly. […] Both kinds of rationality constraints imply a new

> form of autonomy: an intentional agent is not just able to control his or her behaviour, but must also have a certain amount of *control over his or her mental states*. One way to control one's mental states is by forming higher-order intentional states. (Misselhorn, 2015, p. 7)

Higher-order intentionality has intentional states as its object. For humans it means the ability, for instance, to form beliefs *about* intentional states. Intentional states can represent desires and beliefs. A person's first-order desire, combined with a first-order belief, may induce him or her to act – as we say – on an impulse; but then there may also be a second-order belief that tells him or her: *Don't you dare!* If the person in question is really conflicted, or just too philosophically minded, there may no doubt also be a third-order belief, or doubt, that suggests that perhaps the second-order belief is just the unfortunate result of a bad education. For a more extensive treatment of first order desires and second order volitions, the reader is referred to Harry Frankfurt who argues that these traits represent fundamental constituents of the concept of a person (Frankfurt, 1971).

Misselhorn asserts that higher-order intentional states are associated with the idea of free will and action. Naturally we therefore also identify with higher-order desires. They reflect that which people tend to think of as their true selves. According to this logic, people who do not manage to follow their second-order desire, such as drug addicts, fail to be that which they truly desire. They have either lost their scope of freedom or have failed to gain it in the first place (Misselhorn, 2015, p. 8). Here we have a model according to which agency can be accounted for by degrees. Higher-order intentionality qualifies a system for higher degrees of agency, which – according to Misselhorn – enables an agent to obtain higher degrees of autonomy. Higher degrees of agency and autonomy typically also imply qualitatively different types of agent.

By using the categories proposed by Misselhorn, it becomes possible to differentiate between different types of system. MASs, moreover, can exist in 'pure' or 'hybrid' forms. A pure MAS is composed of only one type of agent, whereas a hybrid MAS is composed of different types of agent. Pure systems can be classified on a scale that ranges from primitive to advanced:

> The most primitive form of pure MAS[s] are systems which only involve agents with basic autonomy and intelligence, but without intentional states. Pure systems composed from agents capable of goal-directed behaviour form a second category. Further there are pure systems with agents possessing first-order intentionality […] and finally the ones that just comprise persons. (Misselhorn, 2015, p. 10)

Hybrid MASs, on the other hand, involve at least two different types of agent.

If human persons are understood as the unit attaining the highest degree of agency, this in no way implies that pure MASs composed entirely of humans

are always highly efficient. On the contrary, the existence of incentives to automate functions implies that simple machines are often more efficient than more complex human agents: 'Agents that are cognitively less demanding at the non-intentional level are often computationally more efficient. Therefore, it is also worth investigating the forms of collective agency that can be achieved by cognitively simple agents' (Misselhorn, 2015, p. 11).

Owing to their innate constitution or programmed instructions, individual biological and artificial systems can have properties that enable them to act as unified agents. There are several types of regulatory systemic inter-individual mechanisms that enable collectives of individual agents to display systemic agentic behaviour. Misselhorn cites 'swarming' as a fairly simple coordinated behaviour that can be achieved by means of collision avoidance, velocity matching, and flock centring. Biological and animal agents can also fall into coordinated patterns through 'emergent coordination'. An example of this is when human agents adapt their walking speed to that of other pedestrians. By means of 'entrainment', a system composed of several agents can be synchronised and coordinated to achieve specific tasks. Object affordances, or the opportunities presented by an object, provide another venue for emergent coordination (Misselhorn, 2015, p. 12).

***

To think about agentic functions in larger systemic wholes is not new. The frame was applied earlier to a sports team, which we could label a pure MAS. It was also applied to an old-style farm village that included different types of animal agent, which we could label a hybrid MAS. All kinds of administrative management and bureaucratic modes of organisation also conceive of agentic functions in larger systemic wholes. The factor that makes MASs revolutionary is the *role* given to the artificial agent. Artificial agents are not simply included as new types of agent, as human cultures began long ago to include horses as a new type of agent. Their role is to be subject to continual improvement; their domain is subject to continual expansion. One way to measure the success of the innovative development of bio-artificial hybrid systems is the degree to which humans can be removed from the loop.

The concept of MASs is now adopted as a conceptual tool for the purpose of exploring the potential role of humans *qua* agents in algorithm-intense environments. Today, presumably, the type of MAS that is composed of the most advanced type of agents is pure MASs composed of human agents. If there is such a thing as a MAS ideology, then this ideology motivates computer scientists and other stakeholders to replicate all kinds of human goal-achievement capacities in purely artificial MASs. If such systems today happen to be composed of agents that are thought of as less advanced than human agents, the aim is the continual improvement of artificial hardware and software. The visionary goal – integrating the many and varied operational goals that are

specific to each project – is not only to be able to replicate human capacities and goal-achievements artificially, but also to move far beyond human capacities.

A MAS-like structure can, as in the case of a football team, be conceived of as a 'closed' type of system that operates in well-defined arenas. Or, it can be 'open' and operate in environments that are less well-defined, as a multinational corporation does in the environment of international affairs. In both cases, cooperation has the purpose of accomplishing a goal or set of goals.[41] Our focus should be wider still, in that we must take into consideration potential non-goal-related, or secondary, effects in societies that contain MAS-like structures. This potentially includes the effects on humans who are not *knowingly* participating in any MAS-structure as 'team-members'. From now on, stretching the meaning of the concept, the term MASs (MASs) is distinguished from multi-agent organisations, where 'systems' is understood to apply to structures of limited scale that are applied in fairly well-defined arenas with well-defined purposes, and where 'organisations' is understood to apply to larger and more open-ended structures such as corporations and state bureaucracies. In the remainder of the treatise, reverting to a less technical prose-style, the acronyms will not be used.

Possible complications in algorithm-intense hybrid multi-agent systems

What does all of this imply for the framing of human agency in algorithm-intense environments?

If we adopt the multi-agent system as a paradigm for action, we may discover that there are no clear limits to where a multi-agent system begins and ends. If we consider a phenomenon such as Google, the corporation of Google may be understood as being composed of multiple multi-agent systems; Google as a whole may be understood as a major multi-agent organisation that integrates many lesser multi-agent systems, which, in turn, could be understood as integrating even smaller multi-agent systems all the way down to the least conceivable agent. Where does Google begin and where does it end? It may not be enough to say that Google has global reach. According to the model proposed by Kissinger *et al*., all individuals who 'use' Google may also be understood as 'partnering' with Google. But partnering to what ends and to whose ultimate long-term benefits? Are human users or partners also being integrated into the collective agency of Google? These dynamics are explored further in part II.

From now on, large algorithm-intense structures, such as Google, are referred to as hybrid multi-agent organisations. 'Organisation' is used in order

[41] But it should be remembered that, even if there is only one main goal, the achievement of it may require the realisation of any number of additional goals that are instrumental to the main goal.

to emphasise a difference in size and scope, and the open-ended and continually evolving and adaptive nature of the compositional structure vis-à-vis the environments in which the structure is active. 'System' is used in order to emphasise a more neatly delimitated purpose and structure. The limit between the two should not be understood as binary. Typically, systems are applied by organisations for clear purposes in well-defined arenas.[42]

The situation of human agents in hybrid multi-agent systems and organisations is both similar and dissimilar to the situation of humans in the multi-agent systems and organisations that have integrated the agencies of horses, oxen, and dogs. The situations are similar in that both enrol types of agency and agent that are different from the human type of agency and agent in collaborative activities performed for the sake of higher purposes. They are dissimilar in how they envisage the medium- and long-term roles of the agents involved. The old-style farmers, presumably, envisaged a long-term viable way of life that implied stable patterns of collaboration for the sake of clearly understood goals. Stakeholders and users of algorithm-intense hybrid multi-agents systems envisage the continual development and improvement of the technologies that constitute our new collaborative servants or partners.[43] This dynamic implies the continual re-evaluation of the status of and need for different types of agent that happen to be collaborating in given hybrid multi-agent systems and organisations. Moreover, for many humans who collaborate with hybrid multi-agent systems and organisations on different levels of technical hierarchies, the significance of the higher goal or goals accomplished by the overall structure may be obscure.

Let us assume for now that, in accordance with the expectations of Susskind, the replacement of skilled human agents will continue as algorithmic technologies are developed and improved. How, then, should we envisage our future? Kissinger *et al*. look forward to a new form of relationship that they label 'human–AI partnership'. Humans will no doubt increasingly interact with algorithmic technologies. But whether we choose to label such interaction as 'use' or 'partnership', it will doubtless be the case that the character and significance of such interactions will vary. Some interactions may indeed approach a structure that reminds us of an old-fashioned, conventional kind of partnership. Other interactions will undoubtedly approach structures that could be described more idiomatically by other labels. Nevertheless, we

[42] Thus, we could speak of hybrid multi-agent systems being applied when we consider the prediction of behaviour and preferences on various platforms on the Internet, or when we consider air-defence teams, both operating in relatively well-defined arenas. Such multi-agent systems will be applied by multi-agent organisations – that is, corporations and state bureaucracies, the environments of which will be more open-ended.

[43] Interestingly, contemporary technically advanced societies are becoming more and more inclined to modify the genetic constitutions of conventional animal agents so as to 'improve' human use of them for various ends. Transhumanists envisage even more radical technological modifications of human beings. Are these mere instances of isolated trends? Or are they interconnected expressions of one and the same trend?

should not let the labels that interested parties use to describe interactive phenomena determine our understanding of them.

As we move forward, then, the conceptual tools discussed above are used in order to explore the character and significance of different kinds of human interaction with machines. The paradigm of the multi-agent system allows for unorthodox understandings of how agencies and agents are imbricated in societies. Agency, on the understanding of multi-agent systems, could in some respects be understood as transcending the physical limits of agents enacting their agency. It will not always be clear what drives the overall system: is the collective agency a mere effect of the added agencies of the discrete agents that compose it, or are the lesser agents somehow co-opted by some larger structurally instantiated agency? We could say that human agents, although endowed with agencies proper to them, have always also tended to enact agentic structures that are imposed from outside – for instance, by the authority of a king or the owners of a corporation. In these instances, the primary agencies involved are human-type agencies. But in every era humans have related to other types of agency. Regardless of whether they represent real phenomena, the agencies at work in folklore, mythologies, and religions have often been understood to be structured differently from human agency. As new and various types of artificial agents are embedded in our environments, are we providing spaces for agencies to which we humans are not accustomed and which we may not be intrinsically equipped to understand? What is it that rules over an agent's actions: the intrinsic qualities of the agent, or the structure in which the agent is enrolled? A 16th century king's court is very differently structured from the organisation of Google: the former, although steeped in religion, relies entirely on human agents, whereas the latter increasingly relies on artificial agents. Is the collective agency that can be perceived a product of the individual agents that compose the whole, or does it rather emerge from the very structure of the whole? The question, which is not answered within the framework of this treatise, is raised in order to emphasise the ambiguity of the type of interactions being explored.

Earlier it was argued that instantiations of algorithmic technologies may remove incentives for humans to learn critical knowledge and skills. If this in turn produced harmful knowledge deficits, the incentive to automate even more would tend to increase. This implies that human agents must continually adapt – or *be* adapted – to the new technical circumstances that are constructed. These dynamics are explored further in part II with the aid of Ivan Illich and Jacques Ellul.

A crucial way in which algorithmic technologies – and, for that matter, all kinds of technology – interfere with human ways of doing things is that they represent modifications of agent arenas. This is already implied in the example of incentive-removal: the value of human performance of established practices in a given arena is simply reduced. Agent arenas can be modified in all kinds of ways. Venues to perform established practices can disappear. Venues to

perform new practices can appear. A given arena defines opportunities and limitations to given types of agents. For humans, who vary in competences and acquired skills, some strata of a population may at some stage find that their arenas are being restricted, while other strata simultaneously experience an expansion of their arenas. In our times, entire categories of human-enacted agent-roles frequently and abruptly undergo devaluation in the professional arenas for which they are suited. Meanwhile, as we have seen, the value of other types of system increases. The dynamics of this can be better explained by incorporating the concept of affordances (to be treated in the next section) into our analysis.

Another way in which algorithmic technologies – and this is peculiar to algorithmic technologies – can be made to interfere with human ways of doing and understanding things is through deliberate conceptualised ambiguity. Algorithmic technologies can be made to simulate or emulate social behaviour, to appropriate human language, and to produce output that appears to humans to be rich in meaning. From the point of view of humans, this is likely to produce a space of agency ambiguity within which humans become less and less apt at differentiating between human agents and human-type agency on the one hand, and machine output on the other hand. Humans may then increasingly attribute human-type agency to machines.[44] Just as problematic, as we see further on, are contexts in which very real systemic and organisational dynamics are in operation – real in the sense that they shape the social worlds of humans – but in which humans fail to perceive either agency or consequential activity. This is typically the case when a human agent with a very limited perspective is enrolled in a very large multi-agent structure, the holistic nature of which eludes human comprehension.

The adoption of the paradigm of multi-agent systems, as we see in part II, is liable to shift our focus when we engage in risk assessment. On this view, it may not be so much the construction of discrete artificial agents – the mythical superintelligent AI – that should concern us. Understanding ourselves as interacting with hybrid multi-agent systems and organisations, certain potential complications should become clearer to us. These types of system do not simply replace us: they draw us in; they may even co-opt or usurp our agencies to the benefit of the autonomy of the system or organisation.[45] Such hybrid multi-agent systems, drawn to their extreme conclusion, could already be understood as artificial agencies that more or less integrate human agencies, to the extent that they have rendered most or all humans involved easily interchangeable, and in that they aim, in accordance with a MAS-ideology that could plausibly be inferred, to eliminate the need for the human in the loop.

[44] On the basis of an instrumental understanding of agency, this can of course be entirely justified; but awareness of the realist position should at least prompt us to ask whether such attribution is sound in view of more general human interests.
[45] How this could be done is described in part II.

This last detail differentiates the algorithm-intense and AI-directed systems of today from all previous industrial and administrative modes of organisation.

A priori, it may be very difficult for human agents to discern the potential stages where basic artificial agent-like units begin to be integrated on levels that, in various respects, equal or exceed the hypothesised integration that occurs in the human agent. This will be no less the case when human agents are themselves participating, as agents, in multi-agent systems. If simple units or systems – and even circles and triangles – can be perceived as fully-fledged agents, then how will the typical human agent – that is, *not* the trained computer scientist – be able qualitatively to navigate new complex and integrated spaces of ambiguity?[46] Are we moving towards contexts that are ripe with high perceptual ambiguity, contexts in which the risk of widening discrepancies between the interpretation of phenomena and the underlying realities is likely to increase?

From the point of view of human agents, there is already a prior propensity to attribute agency to phenomena that have no ontological features that should qualify them as agents. We are now busy adding to our environments systems that have ontological features that, on various levels of abstraction, could be understood to qualify them as agents to varying degrees. Innate human propensities and an increase in the performance and quantities of such systems will likely combine to produce new and confusing contexts. On the one hand, more and more artificial agents may be *perceived* – rightly or wrongly, especially if we assume realism – in the environments of tomorrow. On the other hand, it may also turn out to be the case that the most efficient artificial agencies are not perceived at all, that they elude human perception as they engage surreptitiously with their environments and shape the material and social milieus of tomorrow.

Affordances

Previously the terminology of environmental limitations and opportunities and of agents and agent arenas has been used. We could say that an agent's exploitation of environmental possibilities or opportunities enable the

[46] 'Circles and triangles' refers to the 1944 experiment by psychologists Fritz Heider and Mary-Ann Simmel, cited by Daniel Kahneman: 'They made a film, which lasts all of one minute and forty seconds, in which you see a large triangle, a small triangle, and a circle moving around a shape that looks like a schematic view of a house with an open door. Viewers see an aggressive large triangle bullying a smaller triangle, a terrified circle, the circle and the small triangle joining forces to defeat the bully; they also observe much interaction around a door and then an explosive finale. The perception of intention and emotion is irresistible; only people afflicted by autism do not experience it. All this is entirely in your mind, of course. Your mind is ready and eager to identify agents, assign them personality traits and specific intentions, and view their actions as expressing individual propensities. Here again, the evidence is that we are born prepared to make intentional attributions: infants under one year old identify bullies and victims, and expect a pursuer to follow the most direct path in attempting to catch whatever it is chasing' (Kahneman, 2013, p. 76).

realisation of agent-specific goals. The exploitation of environmental opportunities by a collective of agents could pave the way for organisational or societal structures. By accepting environmental limitations and by exploiting environmental opportunities – but usually with self-imposed cultural limitations – human societies have typically managed to bend environments into configurations more compatible with society-specific goals.

In *The Ecological Approach to Visual Perception*, James Gibson introduces the affordance concept in order to explain how agents navigate environments. Gibson considers specific environmental properties, such as ambient light, substances, and liquids, and observes that such basic properties mean different things to different agents. For Gibson's theory of affordances, the reader is invited to consult Gibson (1979, p. 127-43). Here, based primarily on Gibson's understanding, it is specified how the concept is used in the treatise.

Gibson alerts us to agent–environment properties that at first glance seem almost too obvious to merit attention. While they set physical limits to the movements of agents, natural environmental formations 'afford' different things to different agents. To a human a cave affords protection. A lake affords fishing. Below freezing temperature, a lake affords walking. A stick of a certain length affords the extension of the reach of the arm. A stone of a certain size affords being thrown like a projectile. In combination with wood assembled in certain ways, a body of liquid water affords transportation.

The agent-specific limitations of environments and objects are defined by that which they do not afford to specific agents. By adhering to environmental limitations and exploiting environmental affordances, humans have managed to use environmental elements in order to assemble artefacts. Artefacts are the building blocks of artificial environments. The artificial components of the environment of the Stone Age consisted of stone tools and artefacts. Artefacts and tools also come with their respective limitations and affordances.

We could think of affordances as properties located between agents and potential goal-realisations. The affordance itself is invariant, for it 'does not change as the need of the observer changes' (Gibson, 1979, p. 138). If an agent-specific affordance happens to manifest in an agent's environment, however, this does not imply that the agent automatically has the means to discover it unaided. In order for it to be apprehended, other environmental affordances may need to enter the equation. As Gibson expresses it: 'The central question for the theory of affordances is not whether they exist and are real but whether information is available in ambient light for perceiving them' (Gibson, 1979, p. 140). Ann Taves, Egil Asprem and Elliott Ihm restate a commonly shared understanding of affordances in the following terms: 'An affordance is a possibility for action provided to an organism by things and creatures in its environmental niche, given the organism's particular sensorimotor, perceptual, and cognitive abilities' (Taves e*t al.*, 2018, p. 208).

If we consider human-specific affordances, especially in contemporary technically advanced environments, we need to pause and consider what the

definition cited above implies. How should we understand the status of 'possibilities for action' presented to human agents? The word 'possibility' could suggest that affordances are mere neutral possibilities or that affordances are simply means under the control of rational human agents. The first thing to consider, then, is that, for humans, the category of 'cognitive abilities' will depend on a complex range of other factors, including genes, nutrition, and other physical circumstances, but also cultural circumstances that will frame knowledge, know-how, and inclinations. In order for any human to be able to grasp *all* the affordances presented by a modern computer, enormous amounts of computer-specific knowledge and know-how must first have been acquired. This dynamic could be understood in relation to cultural variables that frame how humans perceive affordances. In addition to this, we must consider species- or agent-specific features that may confer on the status of certain possibilities a quality that makes them more than mere neutral possibilities. Although technically correct, the term 'possibility for action' could mislead us in how we perceive certain phenomena. From a psychological[47] point of view, some affordances could also be understood as 'invitations'.

Consider the example of a laptop computer. If a human agent is able to throw a laptop like a projectile, then the laptop, at all times, affords throwing. But a laptop does not – other than in some exceptional circumstances – *invite* throwing. In most circumstances, throwing is not a cost-beneficial use of a laptop, because it is expensive to assemble and it affords – indeed, *invites* – users to do many other things that they tend to value. A reasonable human's judgement of what a laptop affords takes such things into account – that is, it evaluates. But it does not evaluate from the standpoint of some 'objectively reasonable' point of view; it evaluates on the basis of innate urges, ideological persuasions, survivalist desires, hedonistic impulses, and so on; it evaluates, in brief, on the basis of the goals that the human is convinced are pursuit-worthy. Affordance landscapes, then, are anything but neutral settings; with reference to the ends pursued by agents – human, animal, or artificial – an affordance landscape is instrumentally imbricated with values.

Although affordances represent possibilities, the concept is nevertheless used here with the understanding that they also often tend to represent invitations. For a more or less integrated human agent, there are many levels at which an invitation could be received, from invitations to satisfy hedonistic cravings to invitations that pertain to the realisation of the highest and loftiest ideals. Here, again, we could choose to understand human beings as a type of multi-agent system. We nevertheless immediately intuit a crucial difference between the human person – whether conceived as a unified discrete agent or as a multi-agent system – and the artificial multi-agent systems that are in ascendancy; in the latter case, unless we could imagine AIs that suffer from

[47] 'Psychological' is used here as an umbrella term incorporating vast ranges of 'heuristic' points of view: philosophical, ideological, survivalist, hedonistic, etc.

the same kind of addictions and afflictions that humans frequently tend to suffer, the kind of affective-rational tension that typically drives human cognition and behaviour would probably be absent. Algorithmic systems will be able to work methodically and persistently, 24/7, in pursuits of fixed ends, without being interrupted by instrumentally irrational distractions. In other words, they can be made into superbly efficient monomaniacs. By comparison, humans are prone to being diverted into inefficient and sometimes self-destructive pursuits through the beckoning invitations of uncountable affordances.

The angle that is being developed implies that human agents come with inclinations that can be deeply rooted genetically, culturally, and in the life-experiences of individual subject. To such agents, environments, objects, and technologies do not represent sets of objective opportunities for inclination-neutral goal-realisations; instead, they represent invitations to realise the most urgently-felt goals of agents, including the goals of the addict. For human agents, the urge to realise some goals may be mitigated by cultural and moral considerations. *It follows that the affordances of one and the same thing, although technically invariant, may nevertheless present different kinds of invitation to one and the same agent or society, as the agent or society undergoes constitutional change or as the worldview that dominates an agent or a society undergoes structural change.*

Affordances are, somehow, embedded in or imbricated with the factual circumstances that exist between an agent and any potential environment-mediated goal-realisation. How are affordances discovered and evaluated? In a general sense, we could posit that the engagement of agents with environments – with, for instance, caves, sticks, and stones – produces contexts within which affordances can be *revealed* or *appear* to agents. As new key affordances are revealed to human agents, new modes to realise old goals can be made possible. Sometimes discoveries of new affordances can even render previously inconceivable goals conceivable. Reconstruction of affordance landscapes, from, say, a landscape centred on caves, sticks, and stones to a landscape centred on algorithmic architecture, should not be understood as a practical reorganisation that enables better realisations of perennial human goals. Instead, we should expect that reconstructions of affordance landscapes will also tend to induce those who inhabit them to pursue novel goals even as some long-lived traditional goals may be abandoned.

Gibson's inquiry could help us to understand how meaning-making is tied to environmental affordances. Environmental affordances include, of course, affordances of objects and tools, but also, crucially, affordances of other animals and other persons. The latter, affordances of persons and human populations, is of special interest in part II, where we seek to understand what humans may afford to artificial agents. In the context of human societies, affordances of human beings have a fundamental function. In Gibson's understanding, '[w]hat other persons afford, comprises the whole realm of social significance for human beings' (Gibson, 1979, p. 128). Moreover, according to Gibson the

apprehension of all kinds of environmental affordances plays a crucial role in socialisation, and hence in meaning-making, for affordances are 'the invariants that enable two children to perceive the common affordance of the solid shape despite the different perspectives, the affordance of a toy, for example. Only when a child perceives the value of things for others as well as for herself does she begin to be socialized' (Gibson, 1979, p. 141). In the previous analogies between infants and learning machines, it was suggested that innate features in human beings filled some similar function to that filled by algorithmic behavioural recipes in learning machines. Owing to the expression of innate features, infants and young children can be expected to learn and behave in accordance with predictable patterns. This is also largely the case with adult humans, with the added layer of acquired socialisation and culture.

For the socialised agent, some given set of affordances will, instead of presenting the agent with a set of neutral possibilities, often invite the agent to act in specific ways. If a plurality of agents of the latter kind is involved, some given set of affordances, typically a set to which they are accustomed, could issue in the expression of cultural and societal structures. We are presented with a dynamic in which environmental features are methodically and systematically exploited, sometimes passionately, in order to realise agent-specific goals. Some similar dynamic appears to be operational in the context of artificial agents and their environments. Algorithmic systems are typically instructed to realise some goal. They will not be neutral with reference to the goal that they are instructed to pursue. If algorithmic systems and the environments in which they operate are sufficiently complex, the fact that they are instructed to achieve a goal may provide them with dispositions that are similar to a human's complex set of inclinations. For algorithmic systems this could involve the pursuit of many and varied subgoals that are deemed to be instrumentally useful in the pursuit of a main goal. Typically, algorithmic systems will be inclined to exploit systematically the very affordances that are deemed instrumentally most expedient in the realisation of the main goal. Presumably they will do this without disruptive interferences triggered by such human weaknesses as addictions and inclinations to engage in frivolous diversions in general.[48] In the case of both human societies and algorithmic

[48] The question of the role of 'frivolous pursuits' in human cognition probably cannot be separated from the nature of the overarching goal that a human can be reasonably understood to be pursuing. Whereas the overarching goal of algorithmic systems can be precisely defined, in practice, if not necessarily in theory, this is hardly the case with human beings. Humans may articulate well-defined goals, such as serving God, pleasure, or country; but what, in practice, do any of these goals imply? Usually, the articulated service of such goals comes with an understanding of what is good for the human agent. This understanding may imply that such service is the one true way to happiness, justice, and/or harmony; but, again, if we attempt to give such concepts precise definitions, we will soon run into difficulties. Human means-goal hierarchies are, it seems, permeated with vagueness and uncertainty. This inherent structural vagueness differentiates them markedly from algorithmic means-goal hierarchies.

technologies, we could expect systematic efforts to reconstruct environments to configurations that better suit agent-specific goals.

***

Values that are prioritised in any given society can receive their status on the basis of culture-specific value hierarchies sanctioned by tradition, religion, and/or law. Such value hierarchies inform us how we ought to engage with environments. If this is so, then societies are, in part, bound together by value hierarchies. However, if the affordance provides the conceptual link between the values or goals of an agent and the environment that the agent inhabits, then the reconstruction of affordance landscapes must have an impact on the feasibility of the agentic enactment of different values. If this is the case, then fundamental values, values on which human cultures base their ways of life, could be suddenly undermined by means of technical reconstructions, and the formation of other values that are more attuned to an emerging affordance landscape could be rendered conceivable; and this would be quite regardless of whether cultures continue to adhere formally to value hierarchies that are constructed on the basis of the undermined values. Affordances are intertwined or imbricated with values. If this is so, then societies are also practically bound together on the basis of affordance landscapes – that is, on the basis of practical goal-realisation viabilities.

Conclusions Part I

The evolving capacity of previous, present, and future forms of advanced algorithmic systems could be understood in terms of:

1) Increasing degrees of autonomy, or expanding scopes for autonomous action.
2) Increasing abilities to replicate and move beyond human behavioural output.
3) Increasing abilities to learn things that humans can learn and things that humans cannot learn.

With regard to plausible scenarios for humans in algorithm-intense environments, we have become familiar with the following three paradigmatic cases:

1) A world of leisure with little work for humans.
2) Human–AI partnership.
3) Increasing cognitive demands on humans.

These paradigmatic cases are problematised in the coming analyses.

The following conceptual tools are used extensively in coming analyses:

1) Multi-agent systems.
2) Affordances.

Part II: Agency and autonomy in algorithm-intense environments

So far, one might get the impression that a case is being made that differentiates neatly between pre-modern communities, such as old-style peasants, and algorithm-intense environments, and that it is being argued that the latter will have profound effects upon how human agency and autonomy are expressed. In this second part we consider a description of and a critical perspective on how some current algorithmic technologies are applied in social arenas. We also consider some critical perspectives on earlier phases of industrial civilisation. A more nuanced case is constructed. It is argued that, although there are some features that are unique to algorithmic technologies in how they might affect the expression of human agency and autonomy, the expression of human agency and autonomy have already been subject to revolutionary change during earlier phases of modernisation and industrialisation. The so-called AI-revolution, then, could be understood to be less of a revolution in and of itself, and more of a culmination of an already ongoing revolutionary process.

In part I we considered various ways in which algorithmic technologies might affect humans *qua* agents in their professional capacities. It was argued that, as artificial agents supplant humans in work-related tasks, humans might also be affected *qua* knowers and doers, in that they lose the incentive to learn critical knowledge and skills. Here we begin by considering applications of algorithmic technologies not to specific work-related tasks, but to less niched social arenas and environments, where they affect humans not exclusively *qua* workers but also *qua* consumers and *qua* socialising agents. Shoshanna Zuboff's description and criticism of the kind of structures that she labels 'surveillance capitalism' is used to initiate the inquiry. Some elements of Zuboff's critical analysis are critiqued: by interpreting the algorithmic structures she describes through the conceptual lenses of multi-agent systems and affordances, and by drawing on insights from earlier thinkers on earlier phases of technological civilisation, we look at the structures in question from different angles.

Ivan Illich, as we shall see, is much concerned with the potential effects of tool use on humans. Illich seeks to explain how tool use can either boost or undermine human autonomy and conviviality, and argues that we ought to strive for convivial societies. Jacques Ellul highlights the 'technical

phenomenon' – something akin to a mindset or a worldview – that is structurally intent on improving the conditions for the evolution of technique in order that technique may reach the ends suggested by its interior logic. Zygmunt Bauman shows us that many of the administrative and/or manipulative techniques applied by automated algorithmic systems are already in full swing in pre-algorithmic bureaucratic modernity. Norbert Wiener, as we have learnt, saw at an early stage of the history of automation some of the problems that would likely emerge as a consequence of human cognitive limitations in combination with the progressive automation of human society. None of these thinkers lived to see the so-called age of AI. Yet, from their respective angles, they foresaw many of its problems. Together, the analyses of Illich, Ellul, Bauman, and Wiener, along with the theory of affordances and the multi-agent concept, help us to evaluate dynamics that define the extents of and limitations to human agency and autonomy in environments constituted by different iterations of technologies.

Before we move on to consider Zuboff and surveillance capitalism, some of the items previously considered are recapitulated and further problematised.

Through the theory of affordances we can see the entirety of human history in respect of reconstructive affordance-landscape ventures. In the introduction to this treatise, the notion of home-making was introduced. A people's conception of home, according to Brian Walsh, will depend on the worldview inhabited by the people. Notions of home will, we must now add, also depend on the affordance landscape available to a people. In their origin, all human notions of home must in some sense be drawn from the respective affordance landscapes that they inhabit. Some affordances are reorganised into new structures with new affordances, and these structures are given specific meanings; and this, in turn, makes them home. Thus, we could read the history of homes – from those of the non-settled hunter-gatherers, to those of the early homesteaders, to the households of urbanised populations, and now to the evolving concept of the smart home – as, in part, reconstructive affordance ventures.[49]

Ideally a home can represent a symbolic, material, and social base for a way of life. Where humans make themselves symbolically, materially, and socially rooted, there we can say that they make themselves a home. A home, however, need not be sharply delineated: it can be a physical building or a vast and vaguely defined milieu. A home, furthermore, need not be an actually stable milieu: it could be structurally unstable; it could be threatened from outside by physical threats and novel ideas. A yearning for a home nevertheless provides the motivation to render any intended home into a relatively stable milieu.[50] What the analyses of Zuboff, Illich, Ellul, and Bauman have in

[49] A name is yet to be invented for the kind of person who will inhabit a smart home.

[50] Relatively stable, that is, in relation to the cultural environment that surrounds home. The cultural environment, in turn, would ideally be rendered relatively stable in relation to the peripheral environment surrounding it.

common is that they illuminate processes that are disruptive of the status quo of human societies. From the point of view of progressive modernity, this need not necessarily be interpreted as bad. But we need to think of what continuous disruption might portend for the notion of home. How will home-making be practised in algorithm-intense environments? Widening our horizons from home to civilisation, will algorithm-intense environments continue to be constantly disruptive, or will they issue in some new relatively stable and enduring condition? These speculative questions cannot be convincingly answered in this treatise, but they nevertheless serve to illuminate the condition of human agents in affordance landscapes that are being technologically reconstructed at a rapid pace.

When we consider more specific reconstructions of affordance landscapes, reconstructions that aim to make them more algorithm-intense, a somewhat novel dimension is added to the condition of human agents. In many instances, agentic aspects of the human person – a logical mode of proceeding, or a conflicted human psyche, or a human ideology, or a general lack of wisdom – might be *virtually transferred into environments*, not, however, in the form of buildings and monuments (cf. Walsh, 2006), but *qua* automated and quasi-autonomous agentic structures that will then proceed to act upon the human agents inhabiting the same space.

We saw earlier that Stuart Russell envisages a user-friendly AI that, instead of being given the assignment of achieving goals determined in advance, will be designed to satisfy human preferences. This endeavour is undertaken as a way to overcome the goal-intention problem and, more generally, the imposition of any AI-induced alien logic on human society. AI is to be friendly, not alien. But this way of proceeding accentuates other problems. In order to satisfy human preferences, behavioural preferences must somehow be observed. But if giving AIs predetermined goals implies the risks that we considered in part I, instructing them to satisfy human preferences would open up other sets of problems. Russell acknowledges some of these problems (Russell, 2020, p. 235). Humans are often uncertain about what they prefer. They can be in error. He acknowledges that, within a subject, an 'experiencing self' and a 'remembering self' can be in conflict. The experiencing self is hedonically in touch with its environment. The remembering self takes charge when important decisions are to be made. But the remembering self is not to be mistaken for a wise steward: it frequently misremembers and does not necessarily represent any higher wisdom or morals.[51] Last, human preferences seem to change over time in historical populations and in living individuals. This raises the question of the extent to which humans *have* preferences in the first place. How to understand the nature and the function of human preferences is a highly complex affair. What is the meaning of preferences? In order to reach any adequate understanding, we – *and* the machines with which we are concerned – need to

[51] Here Russell draws on Kahneman (2013).

understand more about humans. Since such understandings can be very difficult for humans to achieve, it seems to follow, problematically, that algorithmic systems would need to know as much about us as we know about ourselves. But this is not all. Russell takes for granted that intelligent machines, if they were to satisfy human preferences, would also have an effect on human preferences:

> Why might an intelligent machine deliberately set out to modify the preferences of humans? The answer is quite simple: to make the preferences easier to satisfy. […] One response might be to say that machines must treat human preferences as sacrosanct: nothing can be allowed to change the human's preferences. Unfortunately, this is completely impossible. The very existence of a useful robot aide is likely to have an effect on human preferences. (Russell, 2020, p. 243)

Then, in the next paragraph, Russell suggests that one possible way forward would be for machines to learn about human meta-preferences – 'that is, preferences about what kind of preference change processes might be acceptable or unacceptable'. The word 'processes' in 'preference change processes' implies something like a cognitive structure by means of which a preference could be implemented better. Russell gives the example of a person who wants to be able to eat less cake in the future, but who for the moment does not possess the means to implement this behaviour. On the face of it, that person already has the preference to eat less cake; but somehow, the preference to eat cake is the preference that is implemented.[52] In the rosiest of scenarios, presumably, intelligent machines would be able to help humans to become better versions of themselves:

> The idea that there are acceptable routes to preference modification seems related to the idea that there are acceptable methods of behavior modification whereby, for example, an employer engineers the choice situation so that people make "better" choices about saving for retirement. Often this can be done by manipulating the "non-rational" factors that influence choice, rather than by restricting choices or taxing "bad" choices. *Nudge*, a book by economist Richard Thaler and legal scholar Cass Sunstein, lays out a wide range of supposedly acceptable methods and opportunities to "influence people's behavior in order to make their lives longer, healthier, and better". (Russell, 2020, p. 244)

Russell is nevertheless wary about temptations to create a better world. He suggests that a better way forward might be to construct cognitive aides that 'highlight the longer-term consequences of decisions and teach people to recognise the seeds of those consequences in the present' (Russell, 2020, p. 244). By this stage it seems that it would not be enough for algorithmic systems

[52] Another way to put it is that old habits predominate over the subject's will, or, morally, that entrenched vices predominate over desired virtues.

merely to know as much about us as we know about ourselves; rather, the more they know about us, the better they would serve their intended purposes, and, therefore, something approaching omniscience would be desirable. The desirability of such omniscience, in turn, could work as an incentive to engineer social environments so that environmental machine omniscience becomes more conceivable. Admittedly, the costs of preference engineering, as we see later, may turn out to be high.

How to design technologies that satisfactorily and safely fulfil this or that requirement, and whether or not this could be done, is to some extent a technical matter. However, if the design of technologies depends on knowledge about human psychology and interactions, and if, furthermore, the implementation of technologies affects human psychology and interactions, then it cannot be reduced to a technical matter.

We have learnt that, if AIs are designed in order to satisfy preferences, then AIs can also set out deliberately to change our preferences. This could be done as part of the accomplishment of an instrumentally useful subgoal – for example, to make our preferences more predictable. This can seem frightening, but Russell rightly points out that behavioural modification techniques are not exclusive to AI, and that they are often – whether rightly or wrongly – considered acceptable (Russell, 2020, p. 244).

It has to be acknowledged that the aim of making human preferences or behaviour more predictable is a perennial trait in human society. What are cultures if not social moulds in which human cognition and behaviour are made more predictable? When the cultural aim reaches its target, the results, presumably, are social cohesion, fairly predictable behaviours, and a general sense of shared purpose and meaning. But if one of the functions of enculturation is to produce some forms of culturally shared understanding and behaviour, it is not at all certain that this would be the purpose for which algorithmic preference modification would be undertaken. As we shall see, a very common purpose for such modification at present is monetary profit. Other purposes could be to align a targeted population with politically correct views (integrative propaganda) or to undermine the cohesion of a targeted group (subversive propaganda). This should prompt us to question what might happen if algorithmic behavioural modification were deployed on a massive scale.

Many consider techniques used by public-relations and publicity professionals acceptable. Nudge theory is currently being adopted by state and corporate actors in order to influence human behaviour.[53] If we deem such techniques acceptable, then what is so special about behavioural modification effected by means of AI? It is, of course, always ethically problematic; but the same could be said about many of the techniques used by human agents in information campaigns. However, it is also techno-socially problematic in a sense that the application of non-AI techniques perhaps is not: for in non-AI

[53] See Thaler and Sunstein (2009).

instances, it is generally possible to infer, if not always to identify clearly, a human agent that could, in theory, be held responsible for how techniques are applied. There is usually an identifiable legal agent – individual or organisational – that can be identified as the emitter of a message. As we have learnt, algorithmic technologies can be made to autonomously execute procedures that are opaque to the programmers, owners, and users of the technologies. As we shall see, where algorithmic technologies are involved, it is often unclear who the end-user of the technology is. The AI economy is in some respects more complex and more opaque than any previous mode of organisation. AI systems are also liable vastly to increase the power of users. And if it is often unclear who the end-users of AIs are, then we must of course ask the question: Whose power is likely to be increased?

Algorithmic technologies must always be imbricated on some level with human society. Imbrication would probably constitute a major feature for algorithmic technologies that are intended to serve human preferences. The construction of AIs that, for instance, fulfil the wishes of Russell therefore represents a technical problem only to a certain extent. Academically, it is an interdisciplinary matrix of a problem involving, among other academic disciplines, computer science, ethics, cognitive science, and philosophical and cultural anthropology. In what follows, the challenge of AI is treated largely as a challenge that is rooted in radically novel affordance landscapes. Any given AI may present some affordance or set of affordances to any human agent; and, conversely, any human agent or any aggregate of human agents may present some affordance or set of affordances to any given AI. If, however, we accept that affordances do not present merely neutral opportunities to human agents, but rather invitations to goal-realisations, then, depending on how they are designed, AIs will invite human agents to realise whatever goals they desire to achieve. The complexity is deeper still, in that it will often be the case that human individuals and/or human populations present affordances to artificial agents and hybrid multi-agent systems and organisations. In such instances, artificial agents and hybrid multi-agent structures may, by modifying environmental variables, such as the preferences in human populations, effect circumstances that are more favourable to the achievement of *their own* goals. If humans present affordances to artificial agents and multi-agent systems and organisations, then it seems to follow that, as humans use or interact with such structures, human users or interactors will likely present new kinds of affordances to the humans who control the artificial and hybrid structures.

If this is difficult to take in at the present, it should become clearer in the sections that follow. So far it is difficult to offer any informed judgement on how the expression of individual human agency and autonomy is likely to be affected by human embedding in AI-intense environments. As we move on, it is assumed that historical and current trends in automation and algorithmic technologies do, in some respects, provide fairly reliable indications of future

developments. This may eventually turn out to be wrong, but it is accepted nevertheless as a working hypothesis.

We turn now to a critical analysis of some prominent applications of current algorithmic technologies. Embedded in, imbricated with, and rooted deep within the social milieu inhabited by contemporary humans, these technologies, it seems, have unprecedented potential to shape the expression of human agency in novel ways and to transform structures of human autonomy.

An example of an algorithm-intense ecology – surveillance capitalism

At present, algorithmic technologies contribute to increasing the efficiency in many work-related domains. Earlier we considered humans primarily *qua* workers, and arrived at three broadly plausible futures: a world in which humans are superfluous *qua* workers and must increasingly find meaning and purpose in leisure, in human–AI partnership, and in conditions of increasing complexity under which the cognitive demands on humans will increase. However, algorithmic technologies do not only affect humans *qua* workers. Here, we consider humans in algorithm-intense environments in a more general sense, including humans *qua* consumers and *qua* socialising agents. Increasingly, algorithmic systems define our lives both at and outside of work. The sophistication and reach of algorithmic systems continually increase. At present, algorithmic systems are ensconced in products that we use in our homes, in our pockets, and around our wrists. Since such systems now reach our most intimate and private spheres, it is vital that we consider *what they do* and *whose interests they serve*. At one level, the answer is easy: it is the kind of answer that is implied in any sales pitch. Smartphones and smart-watches, it must be acknowledged, have multiple affordances that are useful to their buyers. As we see when we consider Zuboff's analysis below, the algorithmic systems ensconced in such products often do not serve only the interests of consumers; they also serve the interests of other stakeholders.

Before we begin to look at Shoshanna Zuboff's understanding of surveillance capitalism, the position from which she makes her criticism should be clarified. A critical analysis of a phenomenon labelled 'surveillance capitalism' might produce the impression that the analyst analyses the phenomenon from an anti-capitalist angle. This would be incorrect in the case of Zuboff, who distinguishes between capitalism and surveillance capitalism. The analysis to be undertaken, in turn, distinguishes between Zuboff's descriptions of algorithm-intense environments on the one hand, and her reasoning about surveillance capitalism on the other.

One possible motivation for getting people to interact with machines has already been discussed at length. Learning machines learn by interacting with their environments, and people can be constituents of the environments of learning machines. Zubuff describes a structure under which there is a peculiar convergence of interests to get people to interact with machines. On Zuboff's

account, a combination between a certain ideology and the availability of algorithmic technologies issues in what she labels 'instrumentarian power', a kind of power that is centred not on Orwell's Big Brother, but on a 'Big Other'.[54]

If services made available for free should give us pause, this is often also the case with services that we are invited to purchase. Zuboff provides us with the example of a networked thermostat that comes with a 'privacy policy' and a 'terms-of-service agreement' and an 'end-user licence agreement'. If the thermostat is to function properly, the end-user must accept all the terms. The main purpose of the thermostat is not altogether clear. The so-called 'end-user' – that is, the homeowner – might not be its only or even its most important end-user. In this case, the supplier uses the device to harvest information about the surroundings of the thermostat – that is, about the home of the homeowner. Then:

> sensitive household and personal information are shared with other smart devices, unnamed personnel, and third parties for the purpose of predictive analyses and sales to other unspecified parties. Nest [the supplier] takes little responsibility for the security of the information it collects and none for how the other companies in its ecosystem will put those data to use. (Zuboff, 2019, p. 7)

The ecology described under the category of surveillance capitalism is one in which the deeper functions of things often differ markedly from their purported functions. Behind phenomena that promise convenience there are deeper structures of instrumentarian power at work in the service of a Big Other. Big Other, Zuboff argues, is a more suitable analogy than the often-used Big Brother to characterise the agencies that control instrumentarian power under the system of surveillance capitalism. The new totalitarian is not at all like a classical totalitarian stereotype. Big Other is initially indifferent to the content of our thoughts and our forms of behaviour: it merely harvests information about us in order to use it in predictive analyses. Although Big Other is morally indifferent to the content of our thoughts, instrumentarian power nevertheless implies a systematic shaping of our thoughts and behaviours. Instrumentarian behavioural modification, we shall see, is executed not for moral but for technical reasons.

Although Zuboff works from the assumption that ideology frames the use of technology, her story of how instrumentarian power came to be – one is almost inclined to suggest that it emerged – is telling. Companies such as Google, Zuboff tells us, were originally interested in gathering information in order to improve their services. Behavioural data was then stored on servers,

[54] These dramatic terms, although they are also used to describe the structures analysed, must be understood as key concepts of Zuboff's normative interpretation of the phenomena described.

and in time this produced a storage of so-called behavioural surplus data. At first this data was considered to be garbage, because nobody knew what to do with it. Then, as more potent technologies were designed, it was discovered that the data could be 'fed into processes known as "machine intelligence," and fabricated into prediction products that anticipates what you will do now, soon, and later' (Zuboff, 2019, p. 8). Prediction products can then be traded. Zuboff characterises this trade as a new kind of market – 'behavioural futures markets'.

An ecology in which affordances play a key role by no means undermines the significance of structures such as ideas, ideologies, and worldviews in the shaping of events. For now, let us note how a use for things that previously had no use appeared simply to be discovered as an affordance landscape was in the process of being redesigned.

Under surveillance capitalism, Zuboff argues, human experience is claimed as raw material to be processed into behavioural data and then assembled into prediction products. The fantasy of being able to predict the future is no doubt as old as the human species. If ways to arrive at such predictions were to be alleged to have been invented, it could be expected that this would provoke some rivalry and competitive dynamics – all the more so as long as the ability to *really* predict would remain a distant goal. Such dynamics seem to be mirrored in machines. In part I, we learnt that algorithms set to predict human behaviour may, in accomplishing instrumentally useful subgoals, begin to manipulate or shape human behaviour, so as to be able to predict it better. If we now add the monetary value of behavioural prediction in behavioural futures markets, then there is also a market incentive to shape human behaviour so that it become more predictable. We can add in passing that other types of organisation – governmental, military, etc. – will have other non-market-related incentives to engage in similar practices.[55] Zuboff calls our attention to a reorientation from knowledge to power: 'it is no longer enough to automate information flows *about us*; the goal is to *automate us*' (Zuboff, 2019, p. 8).

Surveillance capitalism, then, is a system predicated on predictions of human behaviours and opinions. Algorithmic statistical analysis provides ways to give predictions that may be considered adequate in some contexts, but they do not yet add up to a machine that predicts the future. Instead, predictability can be improved by simply making human behaviour and opinion more predictable. As intimated in the introduction to part II, one perennial function of all culture is no doubt to mould human cognition and behaviour so that, in some senses, cognition and behaviour become more predictable. If this is

[55] Perhaps another way to put this is to say that behavioural futures markets afford money-making. Algorithmic technologies that extract and store data afford better predictive analysis and behavioural modification. Algorithmic technologies that engage in predictive analysis and behavioural modification processes afford not only the construction of behavioural futures markets but also many other things, including things (more efficient propaganda, policing, etc.) that could be lumped together under the category of 'social control techniques'.

correct, then, at least in some respects, one becomes more predictable simply by inhabiting a culture. Companies in control of powerful algorithmic systems, presumably, want to influence us from outside the local context that we inhabit. They want us, then, to become part of their 'culture'. Surveillance capitalist structures can be understood in analogy and in continuity with some previous forms of governmental and corporate organisations that operate from a supracultural (e.g., global) point of view. Up to the present moment, governments, corporations, and other forms of organisation have been equipped with all sorts of conventional propaganda techniques. These techniques have also been used to shape human behaviour and opinion – that is, to make inhabitants of local cultures conform to governmental and corporate interests. One novelty in the age of AI consists in the algorithmic automation of both processes of prediction and processes of behavioural modification. Another novelty, one that Zuboff understands as a precondition for the viability of surveillance capitalism, enables and reinforces the potential reach and impact of the former: humans are becoming increasingly embedded in environments characterised by ubiquitous computing.[56] Computational devices are now well established in our homes, in our pockets, and on our wrists. The next level could be physical implants.

Instrumentarian power, according to Zuboff, does not exclusively nor even primarily refer to the technological means that are wielded in order to predict and modify human behaviour. Surveillance capitalism is a system under which:

> the means of production are subordinated to an increasingly complex and comprehensive "means of behavioral modification." […] Instrumentarian power shapes human behaviour toward others' ends. Instead of armaments and armies, it works its will through the automated medium of an increasingly ubiquitous computational architecture of "smart" networked devices, things, and spaces. (Zuboff, 2019, p. 8)

The novelty is not that corporations use instruments or technologies in order to pursue their ends. The novelty lies in the use of automated algorithmic architectures that, in turn, automate human behaviour so that humans become instruments in the pursuit of alien ends. Zuboff understands this as a form of totalitarianism. But it is a totalitarianism with a new face, or perhaps a non-face; because it is utterly Other. Big Brother has a face. Power asymmetries taken into account, in Orwell's dystopia the subjects of Big Brother at least vaguely know the score, what is expected of them. They too become instruments in the pursuit of alien totalitarian ends, albeit unwilling or willing instruments. Big Other is trickier.

[56] The concept 'ubiquitous computing', Zuboff informs us, was popularised by Mark Weiser (1991). Today, residents of technological societies are experiencing its coming into full fruition with the proliferation of so-called smart technologies.

In surveillance capitalism, products and services are rarely what they seem. They are not bought and sold on conventional value exchange markets that establish balanced producer–customer reciprocities. Instead, products and services that are bought often have implicit purposes that are markedly different from their apparent purposes. Products and services are no longer simply sold to be used by those who purchase them. Products and services sold to consumers will also continue to be used by the organisations that supply them to consumers and/or by other corporate customers of such organisations. Products and services are designed, of course, to present some value for those to whom they are marketed, but underneath they are 'hooks' that lure buyers into extractive operations:

> in which our personal experiences are scraped and packaged as the means to others' ends. We are not surveillance capitalism's "customers." Although the saying tells us: "If it's free, then you're the product," that is also incorrect. We are the sources of surveillance capitalism's crucial surplus: the objects of a technologically advanced and an increasingly inescapable raw-material-extraction operation. Surveillance capitalism's actual customers are the enterprises that trade in its markets for future behavior. (Zuboff, 2019, p. 10)

It seems likely that the introduction of powerful algorithmic means for prediction and behavioural modification will have profound effects on the ecology as a whole. If economic competitiveness dictates that they be used, then production, commerce, and consumption might simply be subordinated to an important extent to the means of behavioural prediction and control. Zuboff argues that these structures impose unprecedented asymmetries in knowledge and power:

> Surveillance capitalism knows everything *about us*, whereas their operations are designed to be unknowable *to us*. They accumulate vast domains of new knowledge *from us*, but not *for us*. They predict our futures for the sake of others' gain, not ours. (Zuboff, 2019, p. 11)

Objecting to Zuboff's analysis, one could argue that the structures that she labels 'surveillance capitalism' primarily represent a leap in consumer power. No one is forced to use them. If we worry that algorithmic systems influence human behaviour and opinion, then we must also acknowledge that human behaviour and opinion have been under the constant influence of cultural, religious, and political structures since time immemorial. Why should we be extra worried when the influence is exercised by algorithmic systems?

It is difficult, nevertheless, to avoid the disturbing realisation that we are sold products and services that pose as things that they are not – or, rather, that are many different things simultaneously, some of which are typically concealed from consumers. The constant mining of information and the subsequent treatment of that information by algorithmic systems issue in situations

in which algorithmic systems, or their owners and the customers of the owners, can, in some sense, 'know' more about us than we know about ourselves.[57] This 'knowledge' can be shared with multiple algorithmic systems that can then leverage it against us in the interests of intra- and inter-systemic efficiency and in the interests of the owners of the systems. If one of the functions of culture, as previously intimated, is to mould conformity of behaviour, agents – individual and collective – that operate algorithmic systems can purse all kinds of objectives. Regardless of how we view this difference morally, we should inquire into the kind of interests that can be served by behavioural manipulation by algorithmic systems in the service of hybrid multi-agent organisations. If, for instance, profit towers as an overarching goal, then the influence exercised by algorithmic systems over human populations will likely have a very different function from that of conventional cultures. If, on the other hand, algorithmic systems are wielded by some political corpus, then the function could, in some instances, be somewhat similar to the one attributed to conventional cultures.

Ideally, trade will tend towards a degree of balance that benefits both seller and buyer. This is certainly the case in the transactions that Zuboff describes and criticises. What stands out is that the transaction in surveillance capitalist structures implies so much more than a purchase. It is the beginning of an extractive relationship, within which the supplier will learn more and more about the buyer. Another way to put this is that the buyer is generally sold what amounts to short-term conveniences, and that such short-term convenience may come at the expense of long-term dependence and subordination. This, again, may not be a novelty in and of itself; but combined with the increasing ambiguous nature of products and services being sold, and combined with the increasing knowledge that can be leveraged against consumers, it nevertheless portends a radical change in the human condition.

If we also look on this through the paradigm of multi-agent structures, we can discern an additional aspect that should worry us. On some levels, the interaction between consumers and algorithmic products could be understood both as the conventional use of products and, in more innovative jargon, as budding partnerships. However, if Zuboff is correct, and if we take a more holistic view, we could understand the evolving structures as the absorption of human beings into hybrid multi-agent organisations. Why should this be cause for concern? Humans have collaborated in multi-agent organisations since time immemorial. Here, again, the ambiguous and deceptive nature of the game makes much of the difference. To the extent that the primary purpose of a product or service is not that for which the consumer uses it, humans will not primarily be absorbed *qua* workers or autonomous agents – notions that carry connotations of dignity and/or practical importance – but *qua* sources

[57] The word 'know' is here used in the sense of 'being able to predict future data'.

for raw material, as Zuboff argues, or, perhaps even more alarmingly, *qua* unwitting agents in the service of purposes unknown.

The impersonal indifference with which algorithmic systems appear to be applied in the contexts of the structures described by Zuboff evokes the notion of 'structural coupling'. Although the structural coupling between biological agents and environments, as Hossaini describes the matter, is hardly identical to the processes that drive the formations of the kind of structures that Zuboff criticises, it is nevertheless possible to understand part of the dynamic between algorithmic systems and human beings in terms that are similar to that of structural coupling, in which initial 'affordance-matches' turn into structural interdependence and new techno-social modes of being. A detached and abstract observation of the structures that are involved could at least issue in such a description. From another point of view, algorithmic systems in so-called surveillance capitalist structures are not at all applied with indifference but with utmost deliberation. From this perspective, algorithmic systems are understood to be under the control of powerful human agents who are in the position to reap enormous profits through them. From this point of view, we are primarily inclined to consider the interests of the humans that are involved. The structures can be considered from different angles. If we think of instrumentarian power as one possible expression of multi-agent structures, then, in a structure in which instrumentarian power is exercised, the intentions of humans who use such systems at the lower ends of technical hierarchies may have little in common with the goals pursued by the multi-agent structure as a whole; by using such structures, humans, then, could become instruments in the pursuit of alien ends.[58]

From the angle of us who are encouraged to participate in the economy from below, having been lured into the algorithmic grid through the hooks of convenience marketed to us, we now find it impossible to extract ourselves from an increasingly 'sticky' web. The Internet first becomes a convenient mediation between us and our social environment; then it becomes indispensable. The Internet expands to the Internet of Things: it now finds itself in our pocket, embedded in our fridge, wrapped around our wrist as a watch – first as convenience; then it becomes indispensable. We increasingly embed ourselves deeper in systems of ubiquitous computing, first as convenience; then it becomes indispensable. The sensors of ubiquitous computing, having permeated our life-milieus, are now waiting for the logical next step: to embed themselves in our brains and limbs. This too, if and when it happens, would no doubt first be marketed and perceived as convenience; then it would become indispensable. To be sure, that something becomes indispensable does not imply that it automatically ceases to be convenient; but whether or not that something continued to be experienced as convenient, it would remain indispensable.

[58] Give that a thought the next time you relax in your self-driving car.

In light of this, what are we to make of the paradigmatic cases presented in part I? When introducing part II, it was asserted that algorithmic technologies would affect us at and outside of work. The algorithmic systems that operated in the structures discussed by Zuboff, however, put its 'users' *to use*. But the use to which users are put differs from conventional understandings of both work and leisure. Value is nevertheless added as a consequence of users' interaction with such systems – both to the experience of the users, who get to accomplish desired goals, and to the wealth of the operators of the systems.

These types of interaction are far removed from anything that we could understand as salaried work. Is it still plausible to label them as leisure or partnership? Yes, if we zoom in exclusively on the humans who use the systems. At present, it only makes sense to apply the concepts 'leisure' and 'partnership' to humans. It does not make sense to say that any contemporary chatbot or a multi-agent system is at leisure, or that it engages in any meaningful way in partnership with human users. Let us nevertheless accept that it could be plausible to understand the *relationships* between humans and algorithmic systems as including leisure and partnership. What happens if we zoom out from the 'user' in leisure or the human 'partner'? If we zoom out far enough, we will apprehend that there are other 'users' or human 'partners' at higher levels of technical hierarchies. The activities of such human agents, however, could in no conceivable way be described as leisurely. Such human agents would be operators, owners, and stakeholders of macro-systemic structures, the lower-level interfaces of which would be accessible to 'users' or 'partners' at the lower ends of technical hierarchies; and operators, owners, and stakeholders would use such macro-systems deliberately for work-related purposes. To the extent that operators use such systems in extractive and manipulative operations – extracting value from users at the lower ends, manipulating opinions and behaviours in order to make users serve the ends of the operators at the higher ends – it would become quite misleading to characterise the relationships between low-end users and algorithmic systems as partnership. To the extent that such dynamics play out, such relationships would be better characterised as extraction and stealthy manipulation. What is this, if not an instantiation of the mechanical-slave problem described by Wiener? Even if actual circumstances turn out to be comfortable for users at the lower end, could such a status be said to be desirable or praiseworthy?

How is it that human use of algorithmic technologies produces the socio-technical structures described above? Surveillance capitalism, according to Zuboff, is a 'logic in action' and not a 'technology'. It is unimaginable outside of a digital milieu, yet the digital milieu can be shaped into many different forms. It is in the interests of surveillance capitalists to make surveillance capitalism seem inevitable – that is, entirely the result of technology. On the other hand, Zuboff argues that the factors that frame surveillance capitalism are ideology and the market: 'It is capitalism that assigns the price tag of subjugation and helplessness, not the technology' (Zuboff, 2019, p. 15).

Within the analytic frame of this treatise, it is taken for granted that phenomena categorised as ideology and, for that matter, as culture and religion, contribute to shape human life-milieus. However, is the function of technology to be understood as something strictly neutral, something *passive* that is formed and put to use by other *active* agents? Are digital milieus and algorithmic technologies mere prerequisites for instrumentarian power? Or do the previous sections *afford* a different understanding? The answer hinges on the following question: Would the affordances presented by digital milieus and algorithmic technologies awaken, to a statistically significant extent, certain temptations in human beings regardless of ideology? What are the more human-salient affordances that are revealed as these technologies – or affordance landscapes – are developed? How are we – who share the innate proclivities and vulnerabilities specific to human agents – *invited* to use them?

Mythologies, fairy tales, and religious narratives frequently tell us about moral vulnerabilities that appear to be inherent in human beings. Psychologists can express the same pattern in less moral language. The gist of such insights is that human beings have general propensities to respond to certain temptations. The concern, then, is the extent to which the structures critiqued by Zuboff should be thought of primarily as the product of ideology, or the extent to which something else, something more *basic*, is fundamentally involved. The technologies that enable instrumentarian power present power-boosting affordances to the human agents who control them. Let us consider it in analogy with something like J. R. R. Tolkien's rings of power.[59] Is it likely that temptation to power is entirely or mostly rooted in ideology or, to use the concept that is central in part III, to worldview?[60] Perhaps ideology could exacerbate or mitigate or otherwise affect the outcomes of temptations: if a surveillance capitalist, a socialist, or an Amish were to wield control over the same set of technologies, the shapes of events would be likely to turn out somewhat differently. Nevertheless, this does not mean that technology is neutral.

Indeed, communities belonging to the Amish and other similar denominations typically justify regulation of their members' use of technology with arguments based on the understanding that certain technologies are likely to undermine the communal living that their religious cultures uphold as ideal.[61] This means that they doubt that, *even though they, in their own understanding, have the right 'ideology'*, they themselves would be able to use certain technologies for good ends. The justification for limiting technology use implies a general understanding, namely, that the use of certain technologies can undermine certain ways of life and, by further implication, certain ideologies or

[59] The reader will have to read a few pages before the analogy can be grasped.
[60] The word 'ideology' is used here with the understanding that 'ideology' represents structures that belong in more encompassing worldviews.
[61] Assertions about the Amish are drawn from Patrick Deneen (2018, p. 105-7, 189-90, 191).

belief systems. Presumably there are other belief systems and ways of life than those of the Amish that fit better with the technologies that their communities reject. If they are right, then it is not quite the case that specific technologies have the power to determine specific ways of life and belief systems; but the human use of or interaction with specific technologies would certainly increase the likelihood that certain types of ways of life and belief systems would emerge and/or gain traction. If this is true, then we could say that affordances play a key role in shaping ideologies and other kinds of worldview structure.

Zuboff's explanation of the transformation of the aware home into the smart home implies, if not the primacy, then at least the overarching importance of ideology (Zuboff, 2019, pp. 15; 5). However, the example referred to could equally be interpreted to support the understanding that technology shapes human behaviour. The aware home was an earlier and, according to Zuboff, more benign version of that which later became known as the 'smart home'.[62] The aware home was 'meant to be a "living laboratory" for the study of "ubiquitous computing"' (Zuboff, 2019, p. 5). That it was conceived at least partly as a way to develop technology is significant. Although it was predicated on a set of technological innovations, the aware home was nonetheless conceived of in terms reminiscent of conventional understandings of 'home'; notably, it was imagined 'as the private sanctuary of those who dwell within its walls' (Zuboff, 2019, p. 6). The aware home was conceived as a closed loop controlled by the homeowner, and all information relevant to the operation was to be stored on the homeowner's personal computers. The aware home, furthermore, was conceived as an empowerment of the homeowner. Zuboff asks the question: How did we get from the aware home, which was intended to empower the home owner, to the smart home, in which information is extracted from the home in order to empower Google?

Ideology could indeed explain the shift to some extent. Zuboff informs us that the two concepts were developed by different companies, and makes much of the ideology of the developers of the smart home, namely, Google. Another explanation is offered if we consider the revelation of technical affordances. The innovators of both the aware and the smart home initially, perhaps, adopted simply by default some conventional or, rather, modern western notions about 'home'; but as the affordances of new technologies were revealed in the process, new venues for efficient goal-realisation emerged. According to this alternative explanation, we do not need to posit that the company Google was governed by a peculiar ideology in contradistinction to the company Georgia Tech, the developer of the aware home. This alternative instead explains the evolution as Google's more efficient exploitation of

[62] Zuboff informs us that the concept of the aware home was developed by Georgia Tech, while the concept of the smart home was developed by Google (Zuboff, 2019, pp. 5–6).

technical affordances in a pre-existing market economy driven by the search for profit.

In a critical review of Zuboff's work, Evgeny Mozorov argues that the phenomena that Zuboff categorises as surveillance capitalism merely represent new iterations of patterns that are consistent with plain old capitalism. The surveillance phenomena described by Zuboff do indeed represent something entirely new; but why, as Zuboff argues, should all means suddenly be subordinated to the means of behavioural prediction? In order to explain why this should occur, Zuboff posits a peculiar ideology. Mozorov offers instead a more parsimonious explanation, in that such subordination will occur when it is profitable; it will not occur when it is not profitable. The so-called surveillance surplus economy would, then, represent merely one additional profit venue in plain old capitalism (Mozorov, 2019).

The present analysis seeks to push the explanation beyond even the conceptual limits of capitalism. If we choose to understand the phenomena in terms of new and additional profit-venues in plain old capitalism, then we can also choose to understand the matter in terms of additions of new venues for empowerment in an even more general anthropological condition. 'Empowerment' signifies something rather more inclusive than the accumulation of wealth and satisfaction of greed. Let us posit that 'power' stands for an agent's ability to control aspects of its environment so that the environmental conditions benefit some of the agent's aims. Empowerment would then also include the acquisition of functional and efficient technique – in other words, *empowering means* – in a broad sense. Humans have endeavoured in all times to make good artefacts.[63] As we have seen, if AI is to be able to achieve all that Russell envisages, then algorithmic systems will have to know a lot about us. It just so happens that the digital milieu and algorithmic technologies *afford* the perfection of technologies by means of the methodical extraction of information from human users. A similar process of discovery seems to be at play in Zuboff's account of how a use for behavioural surplus information was 'discovered'. Put differently: as the affordance landscape of technological life-milieus was being redesigned, new affordances of behavioural surplus information were progressively revealed.

Mozorov points out that Zuboff appears to be against surveillance only when it is tied to making profit, not when it is used for product development. But the line between the two is not always clear. Nor is it obvious that surveillance would be less harmful when done for the purpose of product development than when it is done for the purpose of profit. This becomes clear if we consider technologies, not as mere potential means on the one hand to

[63] In traditional societies, such endeavours typically issue in well-crafted tools designed for culturally specific purposes. In contemporary technically advanced societies, a prevalent aim is to stretch the limits of science and technology, not for some culturally limited and specific set of purposes, but for the more general purpose of innovation and progress.

consumer satisfaction and on the other hand to corporate profit, but as means to power in a more general sense; in the case of market-based relations, means of empowering consumers, yes, but also means of empowering the institutions that produce and exercise control over hybrid multi-agent systems and organisations.[64] This would be a trivial observation if we did not take into account the evolving power asymmetries between agents at the lower and the higher ends of technical hierarchies.

In the introduction we learnt about Walsh's understanding of worldview and home-making. Walsh explains a Native American tribe's cultural disintegration by positing that their worldview and, notably, their sense of home were undermined by the modern western 'homes' to which they were relocated. An alternative explanation of the evolution from the aware home to the smart home can be constructed on the basis of Walsh's understanding: the innovators of an algorithmic-intense home start out by taking for granted their own presuppositions of what a home represents. As they undertake technological reconstructions of the milieu that is to frame the homes of tomorrow, old ideals become inconvenient in the face of evolving inter-systemic complexities. Unsurprisingly, if we happen to be in a context in which 'soft' values (e.g., values that pertain to our cultural notions of home) are considered to be relative, whereas technological efficiency has objective value, new conceptions of home begin to be formed – conceptions that better fit new and evolving inter-systemic structures. Is it a stretch to suggest that ideological, philosophical, or ethical reasoning may begin to emerge *ex post facto* as a justification for the impressive technological feats that are under construction? It is not being argued that prior worldview has no importance, that technology alone determines social outcomes. It is suggested, nevertheless, that environments and their technological components exert much more influence over human thinking and behaviour than is generally believed. This should be the case especially in contexts in which there are no longer any obvious God-sanctioned moral absolutes to which it is understood that customs and technologies must be adapted. And if, in such contexts, the development and refinement of new technologies is understood as an indisputable good, then it might make sense to adapt prior values, purposes, and customs to evolving technical structures.

Ideology and religious and ethical beliefs will indeed affect how individual humans and human communities respond to certain kinds of temptations, such as greed. But in the cases considered here, systems that afford the satisfaction of greed and other impulses are imbricated in affordance landscapes that are highly complex. Zuboff makes a rather neat distinction between user data that is used for product improvement and behavioural surplus data that is sold on behavioural futures markets. But where does product improvement begin and end? If Russell's vision is to be realised, then it seems that predictive analysis

[64] When algorithmic systems are deployed by institutions that do not have any goal of making profit, notably state institutions, we will likely be able to observe similar dynamics.

would have to be deployed on a grand scale, not for the purpose of money-making but for the purpose of product development. In a general sense, the overall functionality of complex interconnected technical systems will depend on accurate predictive analysis. From the point of view of systemic functionality, the predictive analysis and behavioural modification of instrumentarian power could be understood as a form of product improvement. Unless we have an acute dislike of profit in particular, it seems that it should be equally problematic if such measures were to be undertaken not for the purpose of profit but for the purpose of product development – unless, that is, we valued product development – technological progress – as a supreme good in and of itself, a good for which privacy and other conventionally held modern values could be sacrificed. By availing ourselves of smart technologies, we are consenting to participate in an economy animated by inter-systemic connectivity, and in that context it is not at all clear that the sale of surplus data could not be conceived in some way as product improvement in regard to the systemic whole. This is not to say that the systemic whole could not be improved by other means than those used in the context of surveillance capitalism.

Admittedly, differences in outlook matter: Microsoft and Apple may indeed have evolved differently on the basis of different outlooks; but the difference between Microsoft and Apple is not as great as the difference between both of them and any pre-computational company. Digital technologies add up to new affordance landscapes. Affordance landscapes make certain logics possible. Certain logics, aptly exploited, enable great power. Power and wealth often go together. Indeed, since here we understand power as an agent's ability to control aspects of its environment so that environmental conditions benefit some of the agent's aims, wealth is also a means of power. The algorithmic technologies considered by Zuboff are components of a wider digital affordance landscape. Zuboff's analysis shows us that the affordances of these technologies channel overwhelming power to the human agents who manage to exploit them aptly. It would hardly be surprising if, in order to justify the acquisition of power, previously held ideals and values were adapted or abandoned.

As algorithmic architectures are being erected and new affordances are being revealed, the attainment of overwhelming power becomes not only a possibility, but certainly also a temptation. The apt exploitation of new sets of affordances presented by the latest cutting-edge algorithmic systems produces empowerment; the failure to exploit them disempowers. The desirability of the prize would incentivise humans, regardless of their previous allegiances, to exploit efficiently the affordances available to them. Moreover, the availability of powerful tools – instrumentarian power, indeed – would also shape the way in which their potential users perceive their environments or affordance landscapes and the way in which they understand and envision their world. To a man with a hammer, as the saying goes, everything looks like a nail. Throughout the rest of this treatise we grapple with these elusive

dynamics – how tools and technologies contribute to shaping the worldviews, ways of life, and even mythological narratives of the agents who use them.

In the case of the previously mentioned Amish, who voluntarily refrain from the use of many technologies that are less powerful than the ones under discussion, they appear to be able to find the motivation in their communities not to use many specific technologies that are available to the wider population. Thereby they refrain from exploiting many affordances that are available to them and that, if they were to exploit them, would enable a more efficient achieving of many goals. But the Amish organise their societies around premodern principles that are conducive to a local and communal way of life. In a practical sense, their communities organise around the achievement of certain goals *as* communities. To the degree that any technology enables the individual achievement of such goals, it is understood potentially to undermine the need for community. On this basis, it becomes both meaningful and instrumentally rational for the Amish to refrain voluntarily from and/or limit the use of technologies that are understood to undermine their way of life. Who among us in the modern world are like the Amish?[65] It is more likely the case that, in the absence of firm moral commitments to any established way of life, as in the case of the Amish, contemporary humans, regardless of their opinions, will lack compelling reasons *not* to give in to the temptation to use legally permitted means in order to empower themselves, or, in morally less charged terms, in order to make life more convenient.

Another way in which technology will tend to shape society has already been alluded to. In the case of the surveillance capitalist economy, the very training of learning machines is facilitated if ordinary humans can be incentivised, in various respects, to interact with the systems of which learning machines are part. For developers of AI there is therefore an incentive, from the point of view of product development, to get people 'hooked'. As the developmental course of a learning machine involves the increase of its capacity to predict its environment, there is also an incentive to train AIs to predict their environments. And since social environments are often difficult to navigate,

[65] It is possible to argue from various ethical angles about whether the Amish way of life is worthy of praise or censure. That which is of immediate relevance to our discussion is the Amish commitment to local community and, from the point of view of this commitment, their rather distinctive understanding of technologies as either undermining of community or as potential means to realise community. When a community reasons that a given technology could function as a way to strengthen community, the technology is readily admitted. For further insights into the way of the Amish, see Deneen (2018, p. 105-7, 189-90, 191). Technically advanced societies, it seems, operate the other way around: as new technologies are introduced and implemented in society, the society and its ways of life tend to be adapted to the new circumstances. This adaptability is often described as a virtue. From the analysis in this section, we can now add that it may be the case that it is not only society and ways of life that tend to change as result of the application of new technologies: a case could be made that ideologies and other belief-systems, too, will be progressively adapted to new circumstances. In some instances, entirely new belief-systems may emerge as old ones fall into abeyance. Such a case is made in the sections that follow.

and since prediction becomes easier to the extent that behaviours become more uniform, there is also an incentive to modify human behaviour. These incentives make sense from a strictly technical point of view: if we desire more than anything else to build these technical systems and to make them function efficiently, it appears to be instrumentally rational to proceed in this way. A society's ability to argue against this way of proceeding hinges on the extent to which the society accepts any technology-subordinating values as objectively valid and/or binding. Amish communities appear to recognise precisely such values – values pertaining to their notions of community- and home-making. Many people in technically advanced societies, it could be plausibly posited, merely have opinions on such matters. Some may have very strong opinions; but strong feelings do not convert opinions into values that are considered objectively valid and binding. Therefore, the structures of homes, ways of life, and worldviews of general populations will be more malleable than those of the Amish.[66]

If we add the monetary aspect that is tied to behavioural futures markets, then incentives become even stronger to get people hooked. Yet salient affordances will be perceived by human agents and collective agents such as governmental and military organisations even without this monetary inducement. The algorithmic structures under discussion represent a control apparatus of overwhelming power. Religion and mythology often question the extent to which it is good for *any* human or *any* human organisation to be presented with the means to overwhelming power. The analogy with Tolkien's rings of power tells us that the temptation of power transcends creeds. The danger warned against is not so much that temptation is liable further to empower characters that are already understood as bad, but that temptation is liable to seduce characters that are understood as good and to turn them into characters that are understood as bad. This, of course, is from the point of view of the moral teacher – in this case Tolkien and the Roman Catholic Church of which he was a member. But the tales also tell us that structures empowered, for instance, by technologies, which in this analogy are symbolised by the rings, tend to produce their own self-justifying belief-systems. In the end, we may get societies in which there would be very few individual humans who understood themselves as self-serving exploiters; instead they could conveniently embed themselves in narratives in which they dutifully fulfilled functions that served some greater good or just end. In many tales this aspect is

[66] Even in cases where opinions become held like dogmas, the contemporary social structure in which they are formed, compared with a traditional dogmatic context such as the Roman Catholic Church, is likely to change any dogma-like conviction into a relative and malleable opinion. The dogma of the political left will be different from the dogma of the political right, and the mutual awareness that they are both branches of the same *modern* tree will undermine any pretence at genuine dogmatic status. The analysis of Jacques Ellul, which is considered in the pages that follow, indicates that there may nevertheless be some genuinely held dogma on which most modern parties agree: the supreme desirability of technological progress.

partly or completely concealed under the grotesque masks that are projected on to the agents of evil. This could in some instances be a narrative necessity for the genre. But in the actual worlds inhabited by human beings, it would be rather difficult for that which we may represent as evil to seduce humans to do its bidding if it manifested, in looks and manners, as raving orcs. Instead, its agents will often be veiled by good aesthetics. Even Sauron, we are told, used to appear in seductive forms. It is the fruits of evil that are explicitly distorted, stunted, and ugly. And, as many a biblical saying has it, by its fruits we shall know the tree.

We can understand the phenomenon labelled 'surveillance capitalism' as one that rests on three legs: one leg could be represented by ideologies and worldviews; another could be represented by anthropological, pre-cognitive, and innate proclivities; and the third would be strictly technical. But none of these, it appears, could be neatly separated from the others. To explain it on the exclusive basis of any one of these may provide us with simpler explanations. Sometimes simpler explanations are preferable. But in pursuing simpler explanations, we risk missing the full force behind the convergent dynamics that drive events. In this case, we would risk missing the full force of the challenge to human agency and autonomy.

Towards a holistic understanding of agency, autonomy, and environments

People holding a fairly wide variety of conventional views and ideals would no doubt tend to find it easy to assent to at least one of Zuboff's key assertions. There does indeed seem to be something parasitic about the algorithmic structures described vis-à-vis the humans who interact with them. However, this only makes sense if we are able to describe some ideal pattern or order that is undermined by means of surveillance capitalist systems. The ideal could pertain to either a cultural order or a sense of a more basic human nature or condition. If the ideal pertains to a cultural order, then the parasitic function would consist in the extraction of value from the host – not only a human individual or population in this case, but a specific cultural way of life – to the detriment of the host and to the advantage of the parasite.

Zuboff appears to intuit that something more fundamental than a culture or a way of life is at stake. The hosts, in this case aggregates of human individuals, risk losing something far more important than their respective ways of life and cultures, namely, a sort of natural space for human agency, and therefore also for human autonomy and liberty. Beyond a certain point, humans hooked into the eco-systems defined by surveillance capitalist structures will simply have no room for navigating outside the limits set by algorithms. On this view, basic liberty is at stake, and one-sided dependence or even slavery is at hand.

The emerging algorithmic ecology, then, threatens not only conventional ways of life but also a basic human condition. If we happen to value any of

the traits that would potentially be undermined, then we will see the parasitic functions. But what if we do not hold any firm values? Or what if we happen to value the potentials of new and evolving affordance landscapes more than old customs and old ideals?

Another way to look at the phenomena described and critiqued by Zuboff is in respect of structural and/or evolutionary change. Let us say that we do not hold any firm convictions that rank one way of life above others, but instead are open to opt for practices that, given changing circumstances, we would find useful in a pragmatic sense. Furthermore, suppose that we are of the view that humans have not had much liberty to begin with, that humans have always depended on multi-agent organisations with which they have interacted in more or less integral and more or less peripheral capacities. The emerging new systems discussed by Zuboff could then be conceived of as being an evolutionary change or leap, laying the groundwork for new ways of being or new ways of life. Having denied that there is such a thing as a basic liberty that could be threatened in any meaningful way by surveillance capitalist structures, the only factor that would remain at stake would be a way of life. And if we chose not to value firmly any way of life above other ways of life, then the asymmetric dependences feared by Zuboff could be technically and neutrally characterised in terms of emerging obligate symbiosis in co-evolutionary processes[67] or as budding partnerships that would be liable to propel humanity beyond the limits of its present horizons. Moreover, one could argue that the new ways of life potentially enabled by the algorithmic economy are only at the initial stage of an evolutionary process, and that surveillance capitalism actually represents hope in comparison with more conventional ways of life that have been defined by too many limitations and injustices. Loss of privacy and ubiquitous surveillance represent part of the price that must be paid initially; but it is a price that is worth paying and in time – who knows? – even that price might be mitigated in the ongoing evolutionary processes that carry us to new heights.[68]

If we were to agree with this evolutionary perspective, should we then understand new forms of interaction between humans and machines as yet another iteration of some timeless pattern? In what follows it is argued that, in important respects, there is a continuity between prior forms of organisation and the new patterns of interaction that are emerging in algorithm-intense environments, but that there is also a discontinuity. In general terms, we have a new context within which humans depend on human-run organisations, and within which the human-run organisations to some extent depend on populations in order to get the humans who are able to run them. This, so far, appears

[67] See Edward Lee's *The Coevolution* (Lee, 2019, p. xiii).

[68] Some have argued, in a similar vein, that humans in rich late-capitalist societies are in a position to reap the benefits of the sacrifices made by miners, workers, and sailors in the earlier phases of industrialisation. It is widely understood that progress requires sacrifice.

to be yet another iteration of a perennial human form of organisation; except, that is, for one new feature, namely, the evolving *bio-technical agentic hybridity* that defines these new forms of organisations.

Before we pay closer attention to the bio-technical agentic hybridity, which is indeed new, peculiar, and potentially revolutionary, these new structures are put into a context of social and technical continuity. We have already come across some understandings of autonomy in the previous chapters. If the goal of an AGI is ever achieved, it would then, according to such understandings, imply a system with very high degrees of autonomy. As autonomy is conventionally understood in computer science, this would designate a technology that in a general way could think and act intelligently and independently of human input. This conception of autonomy mirrors more conventional philosophical understandings, according to which human *individuals* are autonomous to the degree that, by means of their reasoning faculties and independently of customs and cultural prejudices, they can constitute a law unto themselves. Such conceptions of autonomy help us to see certain cognitive, social, and technical dynamics from an individualistic angle. Possibly, such conceptions of autonomy also blind us to other dynamics.

In the 1950s, long before the maturation of algorithmic technologies, Jacques Ellul argued that *technique* was becoming autonomous. The easiest way to explain what Ellul meant is first to pay attention to Ivan Illich, a contemporary of Ellul. Illich was concerned with the weakening of human autonomy in industrial economies. We could think of bio-technical hybridity applied to multi-agent systems as something entirely novel that is on the verge of ushering us into a futuristic brave new world. However, we shall see that the conception of autonomy assumed by these two 20th century thinkers already implies, in some respects, a degree of hybridity that problematises the notion of individual autonomy – hybridity, that is, between the individual and the social and, importantly, between the organic and the technical. Once the dynamics explained by Illich are understood, it becomes easier to understand how the technological society described by Ellul could emerge, for the two thinkers describe two mutually reinforcing processes. The logic in this technological society dictates ever more automation of multi-agent systems. As much as AI could be viewed as either a potential solution to or an aggravation of Norbert Wiener's problems, it also represents a possible end-point in the processes described by Illich and Ellul.

We now consider in some depth how things, technologies, and environments could affect or frame human beings. The understanding that is developed may at first glance seem to run squarely against systems of thought that stem from the western Enlightenment tradition. The environment in which humans act and the tools and technologies that they use would, on this view, have profound and often surprising effects on human cognition and ideas. In one sense this really should be compatible with much Enlightenment thinking, for many Enlightenment thinkers recognise the effects of religions, cultures,

and customs. To the extent that humans are encultured in pre-Enlightenment contexts, from the point of view of the Enlightenment, they are liable to accept all manner of superstitions and misconceptions. The aim of the Enlightenment was, by means of reason, to free humanity from all such error so as to arrive at culture-independent and universally valid understandings and knowledge. This aim does not imply that we must abolish religion, culture, and customs; it implies that we must make such institutional frameworks, to the extent that we keep them, enlightened – that is to say that we create an enlightened culture.

It is argued here that environments and technologies, far from being neutral in regard to the human agents who use them, have repercussions for human users that are analogous to how Enlightenment thinkers tended to understand the effects of parochial cultures on the humans encultured in them. This should, in some sense, turn on its head an image that has become a key modern ideal of the human agent, namely, that of a mind controlling matter in time and space.

Thinkers of a typically modern strand tend to understand technologies as something 'neutral'. By 'neutral' is meant something that is merely instrumental and that therefore, given a certain understanding of human nature, can be controlled and used rationally by a human mind. Still, everybody recognises that technologies are often used with unfortunate results. The neutrality of technology is maintained by the additional argument that technologies can be 'misused'. The fault, it is often asserted, is not with technology but with the users. Zuboff assigns part of the fault to ideology, which she argues frames the owners of algorithmic systems, from her perspective, to misuse them. All such understandings imply that, in order not to misuse technologies, human users must in one way or another become 'enlightened' users.[69] For a long time there has been hardly any recognition in the western intellectual tradition that different sets of tools and technologies might frame humans to think and act in different ways.

Neat separations between, on the one hand, the categories of technologies, techniques, and milieus, and on the other hand, ideologies and other worldview-related patterns, do not accurately reflect the human condition. The understanding of affordances as *behaviour-soliciting* posited earlier suggests that a more circular and complex model ought to be adopted.

Lewis Mumford was one of the first thinkers to pay greater attention to the way in which technologies shape culture and human behaviour. He famously discussed some of the great technological ironies, such as the clock. The clock was invented by pious monks in order to structure the prayer life of their

[69] In this sentence 'enlightened' does not necessarily refer to the western Enlightenment tradition, but rather to whatever 'philosophy' or policy that the person making such arguments considers reasonable. It is unlikely that any 'enlightened' way to use technologies could be specified that could not in its turn be criticised, from a different standpoint, as ideological or parochial misuse.

monastery; it ended up enabling regimentation and industrialisation (Mumford and Winner, 2010, pp. 12–18). Did industrialists misuse the invention of the clock? Did the inventors of the clock themselves misuse the clock? The questions are absurd. The invention of the clock afforded goal-realisations that, during the time that the clock was implemented in society, were simply expedient. In a similar vein, the development of algorithmic technologies produces affordances that enable the exploitation of information to ends that, in contemporary contexts, are simply expedient – expedient with reference to certain ideologies, yes, no doubt; but also expedient relative to making technical systems more efficient (i.e., general empowerment), and, possibly, relative to innate and experientially acquired human appetites and desires.

Once it is understood and accepted that technologies, and, for that matter, other environmental properties, have the potential to shape human perception and behaviour, we will have begun to distance ourselves somewhat from the philosophical anthropology propagated by Enlightenment philosophy. We could still think of human beings as supremely rational among the agents that inhabit Planet Earth, but it seems that now they are not as sovereign as they were once imagined to be. To use Charles Taylor's terminology, the hallmark of the modern human self-conception, the so-called 'buffered self', begins to yield; once again we are becoming porous.[70] Presently, the invisible world of spirits that was once imagined to impinge upon a porous human self is perhaps a far-fetched idea. Nevertheless, we begin to understand that we are not quite the controllers of matter in time and space that we imagined ourselves to be: our interventions into our habitats and the very artefacts that we manufacture and release into our environments have repercussions for the kinds of ways of life and worldviews that remain and/or become viable. This is because environmental interventions and the proliferation of new artefacts change our affordance landscapes in a way and at a pace that we cannot fully control. New sets of technologies may unexpectedly undermine some ways of life and favour others. This means that technologies are *not* neutral with respect to things that humans value.

Ivan Illich and the loss of human autonomy in industrial economies

Ivan Illich is much concerned with the loss of human autonomy in industrial economies. The very structure of industrial economies, he argues, undermines human autonomy. More specifically, Illich wants us to pay very close attention to the tools that we use.

In *Tools for Conviviality*, Illich envisages an alternative to industrial modernity: a convivial modernity. A 'convivial society', according to Illich, will be structured to an important extent around 'convivial tools'. One of the root causes of many of the social and ecological ills of modern societies, Illich

[70] For the buffered self and the porous self, see Taylor (2018). These concepts are discussed in more detail in part III.

argues, is that they are in the process of becoming more and more structured around 'industrial tools'. Industrial tools and structures typically produce important gains in efficiency in narrowly specific domains, but, unfortunately, at the cost of conviviality. A convivial modernity would not necessarily be a society utterly bereft of industrial structures, but one that limits the use and reach of industrial modes of production and industrial tools, leaving important spaces for conviviality to flourish.

In order to make machines work for humans, according to Illich's understanding, a schooling of humans in the service of machines is required. This paradigm, owing to its corrosive effects on conviviality, has failed. 'People need new tools to work with rather than tools that "work" for them. They need technology to make the most of the energy and imagination each has, rather than more well-programmed energy slaves' (Illich, 2021, p. 10). Illich's ideally shaped society, it seems, is the opposite of the society that Susskind envisages. It would also diverge from the ones envisaged by Wiener, Kissinger, Schmidt, and Huttenlocher, for it would be one of 'higher social effectiveness with lower industrial efficiency' (Illich, 2021, p. 20). For Illich, '[n]ew understanding of nature can now be applied to our tools either for the purpose of propelling us into a hyperindustrial age of electronic cybernetics or to help us develop a wide range of truly modern and yet convivial tools' (Illich, 2021, p. 34).

What, then, does Illich mean by 'conviviality'? The concept is used as 'an intrinsic ethical value'; it designates 'the opposite of industrial productivity', namely, contexts where 'individual freedom' is 'realised in personal interdependence'. Illich argues that 'in any society, as conviviality is reduced below a certain level, no amount of industrial productivity can effectively satisfy the needs it creates among society's members' (Illich, 2021, p. 11). What needs are created by industrial productivity and the erosion of conviviality? One of Illich's key explanatory concepts is 'radical monopoly'. Radical monopolies are formed in conjunction with industrial growth and the erosion of conviviality. In order to illustrate the typical process in which radical monopolies emerge, we could consider some anecdotal evidence from my father's birth village.

If conviviality, contrary to radical monopolies and industrial production, is often expressed in pre-modern ways of life, this does not of course mean that all pre-modern social structures were convivial. Both bureaucracy and slavery precede modernity, and neither of them could be understood as convivial. Illich also recognises that pre-modern societies often imply limits to human autonomy. Nevertheless, the family in which my father was raised and the wider village context in which it was situated could be considered rich in conviviality. It could be understood as an old-fashioned hybrid multi-agent system, one that included human and other animal agents. My father was born in 1931, and started to work on the farm at the age of six. Life was hard, but hardships notwithstanding, he and his siblings learnt a number of generally useful skills:

how to grow vegetables, how to take care of animals, how to produce food, how to make clothes, how to make tools, how to build things, etc. They even bred their own means of transportation – horses. Not all family members had exactly the same generally useful skills, but each could compensate for what was lacking by relying on other family members. Meanwhile, animal agents, such as oxen and dogs, were trained to contribute to the farm with their agencies. Rarer skills, pertaining to such things as shoemaking, could be found in the wider circle of the village. The family, portrayed in the photograph on the front cover of this treatise, was self-sufficient in everything that they needed with the exception of salt, sugar, and cloth, which they bought at the local market. This type of social structure could be understood to express a kind of relatively self-sufficient collective and local autonomy – a good example of a pre-modern structure that, to use Illich's words, could generate limited 'individual freedom' that is 'realised in personal interdependence'.

The changes that Illich tries to explain in his various writings are all connected to the mystery of how people like my father, who learnt how to do so many generally useful things, ended up parenting children like me, who instead learnt how to get things. For Illich, autonomy is never considered separate from the convivial nature of human beings. Autonomy does not begin and end in the human individual, but is something that human individuals can live out in contexts of conviviality. The concept of autonomy, based both on Illich's understanding of conviviality and on computer scientists' understanding of multi-agent systems, can now be stretched to acquire a more collective or holistic dimension. The family described above could be understood as a social body with a fairly high degree of autonomy. The family, in turn, is embedded in a village that may have an even higher degree of autonomy. Through the agents – human and animal – that constitute them, such organisations are able to organise the environments in which they are located so that the organisations become able to preserve and strengthen themselves (or to resist entropy). *The degree to which a system or organisation is able to do so can be said to correspond to the degree to which it is autonomous relative to its environment*. This conception of autonomy, then, is based on any system's or organisation's *viability in respect of survival and/or other goal-directed environmental interventions*. The autonomy of pre-modern organisations depended on the knowledge and skills inherent in the human agents who constituted them. The very generally useful skills of the human agents who animated such organisations, as in the case of the family described above, made humans useful in other similar pre-modern contexts, of which there were many.

It has already been argued that the increasing availability of varieties of goal-accomplishing algorithmic systems would likely remove the incentives for humans to learn difficult skills. This notion is drawn from Illich, who in fact describes this very dynamic in the context of industrial societies that predate the maturation of algorithmic systems. The dynamic involves the

emergence of radical monopolies.[71] As societies industrialise, industrially mass-produced goods that are released into pre-modern contexts can, in the vocabulary developed in part I, become incentive-removers. Skills that were once considered critical for survival might then cease to be transmitted to the next generation. In industrial economies, we learn how to get hold of what we need and, increasingly, what we want (which may have nothing to do with what we need) from venues associated with centralised industrial institutions. The type of practically skilled human agent who until recently was of great value then ceases to be of value. Villagers cease to need one another to the extent that they used to; family members cease to need one another to the extent they used to. Critical skills cease to be transmitted. The autonomy of the family and the village – timeless contexts for conviviality – is undermined.

Instead of the potential individual freedom realised in personal interdependence, new contexts are formed within which individuals – free or unfree; that point could be discussed – become dependent on industrial institutions. Illich uses the concept of radical monopoly to explain the change:

> By "radical monopoly" I mean the dominance of one type of product rather than the dominance of one brand. I speak of radical monopoly when one industrial production process exercises an exclusive control over the satisfaction of a pressing need, and excludes non-industrial activities from competition. (Illich, 2021, p. 52)

The erosion of the pre-modern peasant village way of life could be understood to be linked to the emergence of radical monopolies.

> The establishment of radical monopoly happens when people give up their native ability to do what they can do for themselves and for each other, in exchange for something "better" that can be done for them only by a major tool. Radical monopoly reflects the industrial institutionalization of values. (Illich, 2021, p. 54)

As with 'conviviality', Illich appears to understand 'industrialisation' as an ethical value. The surrender of 'native abilities', Illich explains, is `but the beginning in a self-reinforcing industrialisation loop. Industrial planners look upon society as it is at the moment of the planning. They take levels of knowledge, skill sets, and social cohesion as givens. However, as the activity of the industrial mode of operation undermines all of these 'givens', industrial production generates the demand for more and new kinds of industrial production. How? Families and villages that cease to be composed of members with broad practical skill sets, who are bound together reciprocally and with high degrees of cohesion, will also cease to do many things that were

[71] One of the factors that potentially aggravate the present algorithm-intense condition relative to earlier industrial conditions, the lump of labour fallacy fallacy, is lucidly explained by Susskind in part I.

previously taken for granted. In time, some of the activities that cease to be practised may be revealed to have been of critical importance not only for social cohesion and well-being but also for industrial production. Such disruption of established patterns can cause social and productive dysfunction. The industrial logic, unaware that it has significantly contributed to the cause of such problems, dictates that these kinds of problem too must be industrially solved. And even more industrial production will undermine even more convivial skill sets, which will generate a demand for even more and new kinds of industrial production.

High degrees of conviviality, Illich argues, are reached within certain limits of scale. There is no inherent structural need to expand beyond those limits. On the contrary, expansion beyond ideal limits will often even be counterproductive, in that the expanded structure risks losing its cohesion. Industrial structures, on the other hand, appear on Illich's understanding to have no natural limits of scale. Once industrial structures begin to expand, Illich argues, they produce the conditions under which they will be required to expand further in order merely to survive. Illich argues that '[a] society committed to high levels of shared learning and critical personal intercourse must set pedagogical limits on industrial growth' and that '[w]hen an enterprise grows beyond a certain point on this scale, it first frustrates the end for which it was originally designed, and then rapidly becomes a threat to society itself' (Illich, 2021, p. x). 'Society itself' here must be understood as some kind of structure that predates industrial structures. Alternatively, of course, one could argue that society undergoes structural change, that this need not be bad, and that it can even be good. Before we discuss this further, let us briefly consider the individual.

In industrial economies, the nature of critical skills changes. In order to be of value, individuals now need to learn skills that are useful in the contexts of industrial institutions. Skills that used to be generally useful give way to narrowly useful specialised skills. The low end of the autonomy of the village and the family is reached when individuals no longer need these institutions. But it is a mistake to assume that autonomy is simply usurped by individuals from family and village structures. On the contrary, individuals who cultivate the requisite specialist skills that the industrial economy demands become even more dependent on industrial systems than they ever were on family and village structures. They now no longer possess such generally useful skills as those of growing crop and producing food. In time, they lose the securities connected with being embedded in extended family and village structures. They have indeed changed the contexts in which they exercise their agency, but they have not necessarily become more autonomous *qua* individuals. Whereas much Enlightenment-inspired thinking considers modern contexts as more propitious for the exercise of individual autonomy, Illich argues that the modern structures that have *de facto* become entrenched in society are represented by the industrial way of production. Illich argues that such structures,

on the contrary, undermine human autonomy. The more that the industrialisation process proceeds, the more human autonomy will be undermined.

The present analysis may strike some readers as controversial. Why should we see the industrialisation of pre-industrial societies in such a bad light? Some skills will inevitably be lost, this is true, but other skills will be gained. Illich understands many such new skills as the result of schooling in the service of machines; but, by acquiring new skills, humans can perhaps learn to flourish in new industrialised environments. And if securities connected with being embedded in extended family and village structures are lost, governmental and corporate policies can provide security networks for the industrialised human. The industrial context, animated by a meritocratic ethos, may even in some respects be conceived as fairer to the individual than pre-modern contexts. Let us posit, then, that the transition from pre-modern to modern conditions is not necessarily a bad thing. For Illich, at least, the dispute is not between pre-modernity and modernity, but between two alternative modernities.

Since Illich's convivial modernity would safeguard some basic patterns that have been foundational in pre-modern societies, it is useful to compare the role of the individual in pre-modern family and village structures with the role of the individual in modern corporate institutions. A structure such as the family is, in some respects, deeply vulnerable. If, for instance, the father of the family begins to drink too much, it can ruin the entire family. If, on the other hand, a clerk in a large corporation begins to drink too much, the clerk is easy to replace, and the damage to the corporation will often be insignificant. At the same time, the *need* that a family structure generates for each member of the family is absent in the more anonymous and abstract context of the corporation. The awareness that each member is, to a large extent, irreplaceable is liable to foster a sense of duty and loyalty in family structures. In corporations the replication of such organic or cultural ties must be achieved instead by means of legal procedures and the inculcation of work ethics. If these characterisations are fairly on target, could one say that any structure – the pre-modern family or the modern corporation – favours 'human' autonomy more than the other?

Humans, in various agentic capacities, constitute both pre-modern villages and modern corporations. It would therefore be problematic to argue that one is more 'human' than the other. What we can say is that humans in corporations are interchangeable *qua* narrow experts who are needed in specific contexts. Human agents in corporations pursue ends – collective ends – that are often far removed from the individual goods of the humans who compose them. In exchange, the individual humans who compose corporations receive a salary. Such industrial structures, then, offer no enduring roots, but mercenary and meritocratic relations, interchangeability. If it is difficult to make an unambiguous case that human autonomy as such – understood as individual autonomy – is undermined by industrial processes, then what we can say with

a high degree of certainty is that collective autonomy shifts from local pre-modern structures to centralised bureaucratic structures. These bureaucratic structures provide life conditions that are different from those provided by pre-modern structures. On the one side we have the family and the village, and on the other the industrialised army and the corporation. Which one we prefer will no doubt depend on what we value and how we conceive of ideal conditions for human beings. Illich, it could be maintained, is correct in regarding both conviviality and industrialisation as ethical values. And it is only with reference to some values or ideals that we can really *evaluate* these shifts.

On the whole it must be admitted that industrialised society, at least in some important respects, is immensely more powerful than any pre-industrial society. Like no previous society, it can re-organise its environment so that both social and material components of the environment favour industrial production. This capacity, in some respects at least, buffers industrialised citizens against natural environments to an extent never before experienced, gifting them with higher degrees of experienced convenience and security. Should we not therefor conclude that the net effect of skills lost vs skills gained represents a civilisational gain rather than a loss? To the latter we may respond, again using the already established short-, medium-, and long-term grid, that, at least in the long term, industrialisation and the novel ways of life that emerge in its wake must be understood more as a gamble than as either an unambiguous gain or a loss. We simply do not know the extent to which the inheritors of earlier industrialisation processes will be able to cope with all the effects set in motion by industrialisation. In the very short term, some will see only gains. In the medium term, Illich sees important losses that pertain to human autonomy. But perhaps Illich is simply shedding light on a type of disturbance that is inevitable in times of transition. One could argue that it remains the case that, during earlier industrialisation processes, the new skills that were learnt were *needed* in industrial society, and that therefore they represented an investment that humans could make in order to become valued members of a new type of society. Once the transition, which necessarily entailed many inconveniences and disturbances, was completed, the humans inhabiting the industrialised order would, in time, fare much better than their predecessors.

Depending on what we choose to look at and on what we ultimately value, we may see both gains and losses in the same process. If we adopt this more positive view of industrialisation, we nevertheless appear, at present, to be facing something of an impasse.

In the algorithmic era, some key circumstances have changed. If the newly industrialised denizens discovered new ways to be of value to the larger social context *qua* skilled and specialised workers, then Zuboff shows us new and surprising ways in which the newly algorithmised denizens might be of value to new emerging hybrid multi-agent organisations – not *qua* skilled or specialised workers, but as sources of raw material, data-mining, and value-extraction. Admittedly, algorithm-intense hybrid multi-agent organisations still

need human agents *qua* skilled and specialised workers, but in far fewer numbers than earlier forms of organisations. Humans who are not needed as workers are, to varying extents, needed in other capacities, notably as sources of information in aggregate populations. In other words, many humans will not be needed *qua* individuals at all in any meaningful sense, for the individual source of information will no doubt be insignificant and easily interchangeable in such aggregates.

To this it must be added that, in most contemporary technically advanced societies, there are important monetary incentives to encourage humans to spend their time in ways that might be deemed frivolous. There are monetary incentives to promote the activity of gaming; but in what respects would the acquisition of gaming skills represent any gain if the acquisition were to imply the atrophy of previously common critical skills, such as literacy? How about many other highly remunerative online activities, such as the ways of influencers? Most activities no doubt require some degree of some kind of skill set. Cultures, in the phases in which they persevere and flourish, typically promote the acquisition of the critical skills that are needed for individual and collective survival and flourishing. Such skills, one could posit, animate cultures. Meanwhile, cultures typically discourage the acquisition of skills that are understood to be potentially harmful.[72] In contrast, the de-traditionalised and algorithmically defined world of today suggests that we can proceed to choose our ways in accordance with our own personal whims, and that our choice will have little meaningful impact on the social whole.

As we leave more and more important stuff to algorithms, we become free in some sense to pursue whatever frivolous ends our whims may suggest, and free to acquire skills that may be completely divorced from fundamental human needs. But as we 'spontaneously' choose to pursue our preferred ends, we are, at least in some instances, being influenced and/or behaviourally modified by the algorithmic systems with which we interact. In the contexts that Zuboff describes we are unwittingly being put to work, without renumeration, for ends that are unknown and potentially utterly alien to us. If there were such

[72] The human acquisition of skill sets that might be deemed to be of ambiguous value or even harmful may, however, be short-lived. As this is being written, it is reported that AI-generated influencers are increasingly competing with human influencers in the relatively new market of influencers. Some may shed few tears over the potential disappearance of this novel profession. But it is not vanishing. Governmental and corporate organisations can now generate artificial influencers that are customised to their needs and that have the capacity to interact with their followers to degrees that no human influencers have. It has been reported that a famous influencer in Sweden now runs her own company that specialises at generating AI influencers (*Ekot*, April 29 2024), and that '[t]he world's first artificial beauty pageant has been launched by the Fanvue World AI Creator Awards (WAICAs), with a host of AI-generated images and influencers competing for a share of $20,000' (Mouriquand, 2024). Iterations of such phenomena in the algorithmic era cannot be seen as equal to iterations of similar phenomena in pre-algorithmic contexts: the constant accessibility of an increasing range of services is already a much-discussed phenomenon; and now we are fast moving towards AI-generated content that is customised to hook certain personality types and, potentially, specific individuals.

a thing as human autonomy, then this would surely undermine at least the idea we have of it, if not necessarily the actual experience of exercising it. It is beyond dispute that old-world skills to produce food and to build and maintain housing represent social and evolutionary advantages. It remains to be seen whether commonly acquired skills in the algorithmic era will represent anything equally meaningful.

Although present conditions in the emerging algorithm-intense economy, with respect to the expected role of human agents, is markedly different from the contexts of industrialisation considered by Illich, in other respects there is also continuity. Current algorithmic technologies and the aim of general artificial intelligence could be considered a possible culmination of the process against which Illich warns. In a convivial society technologies will be, in Illich's words, in the service of 'politically interrelated individuals rather than managers'[73] (Illich, 2021, p. xii). Illich no doubt had not even begun to imagine the prospects of technologies in service of algorithmic managers, the very extreme antithesis to his convivially structured society.

In *Tools For Conviviality*, Illich is not so much concerned with diagnosing the ills of industrial economies as proposing a way to shape societies that are more convivial. The diagnosis is supplied as background to a possible healing process. In industrial economies, humans want to make technology work for humans; in order to achieve this, Illich argues, they need to school humans for life in service of technology. A convivial society, on the other hand, 'should be designed to allow all its members the most autonomous action by means of tools least controlled by others' (Illich, 2021, p. 20). Illich does not propose a fixed utopia instead of industrial economies. Rather, he tries to show us how the shaping of means in turn could shape human society for better or for worse, where, with reference to Illich's own values, 'better' must be understood to be towards more conviviality and 'worse' must be understood to be towards less conviviality. If we make convivial tools instead of industrial tools, Illich argues, we will lay the groundwork for convivial and communal ways of life. The result will not be one ideal society that is identical all over the planet, but the flourishing of a variety of local convivial societies. Illich's self-avowed purpose is 'to lay down criteria by which the manipulation of humans for the sake of their tools can be immediately recognized, and thus to exclude those

[73] For a fuller description of what would render a society convivial, consider the paragraph in its entirety: 'To formulate a theory about a future society both very modern and not dominated by industry, it will be necessary to recognize natural scales and limits. We must come to admit that only within limits can machines take the place of slaves; beyond these limits they lead to a new kind of serfdom. Only within limits can education fit people into a man-made environment: beyond these limits lies the universal schoolhouse, hospital ward, or prison. Only within limits ought politics to be concerned with the distribution of maximum industrial outputs, rather than with equal inputs of either energy or information. Once these limits are recognised, it becomes possible to articulate the triadic relationship between persons, tools, and a new collectivity. *Such a society, in which modern technologies serve politically interrelated individuals rather than managers, I will call "convivial"*' (Illich, 2021, p. xii).

artefacts and institutions which inevitably extinguish a convivial life style' (Illich, 2021, p. 14). Note here the emphasis on means. Whereas many thinkers put the emphasis on a new type of human – liberated from the constraining yoke of bourgeoise ideology in the case of Marx, or the enslaving mentality of the Christian religion in the case of Nietzsche – Illich does not propose a new type of human, nor does he argue for a new political regime; instead, he proposes that the shaping of different tools would produce a more convivial society. Moreover, Illich is convinced that '[s]ome tools are destructive no matter who owns them', and that '[d]estructive tools must inevitably increase regimentation, dependence, exploitation, or impotence, and rob not only the rich but also the poor of conviviality, which is the primary treasure in many so-called "underdeveloped" areas' (Illich, 2021, p. 26).

To agree with Illich's emphasis does not, of course, imply the negation of the importance of ideology or, as we shall explore later, of worldview and mythological narratives. The way explored here ultimately integrates them all into a holistic or ecological analysis. Furthermore, to agree with Illich's analysis does not imply that we must also agree with his values. One could argue that his understanding of how tools affect ways of life is on target and, contrary to Illich, favour an industrial way of life: the more we use industrial tools, the better. If on the other hand we find value in conviviality, then, to the extent that we find Illich's analysis convincing, we should recognise that we are in a difficult predicament.

If we consider Illich's distinction between industrial and convivial tools, we see that technically advanced societies have not heeded Illich's advice. The difference between convivial and industrial tools is not binary. It plays out on a continuum between two extremes. Let us consider the example of the car. Cars produced fifty years ago were considerably more convivial than cars produced today. When old cars broke, it was often the case that their owners could, if they were moderately interested in how cars functioned, get the requisite spare parts and repair the cars themselves. If the owners were not so inclined, there was the possibility that a neighbour or a friend could help them. The conviviality of a tool can be measured in part by how it structures human relationships. Today, with modern cars, anything that needs fixing structures asymmetrical relationships between owners or renters and the industrialised institutions that produce cars. Mechanical interests and skills are of no use when service requires specialised equipment that is under the exclusive control of industrial institutions.[74] By this criterion, horse carriages were, of course, even more convivial technologies in their time.

The trend of much current technological development is away from conviviality and towards increased dependence on technologies produced and

[74] For Illich's more elaborate discussion of the conviviality of cars and roads, see Illich (2021, p. 36).

controlled by specialised institutions, and, ultimately, radical monopolies.[75] A few decades ago, amateurs were able build their own computers in their homes. At present it is doubtful that there is one person who could understand all the intricacies of the multiple and specialised components that constitute a modern computer. And to the extent that other technologies are computerised – as in the instance of modern cars – it reinforces the race away from conviviality. This movement ties consequential relationships to centralised and specialised institutions. Such relationships are increasingly asymmetrical in knowledge and power. The asymmetry is deepened by our becoming increasingly dependent on specialised institutions for bare existence and for all manner of activities, while we, as individuals, have less and less to offer specialised institutions in the way of skills and resources. Like the money we give them in order to purchase services, we are becoming fungible. With the introduction of the algorithmic surveillance economy, according to Zuboff, relationships become unprecedentedly asymmetrical. Illich helps us to see that such asymmetries can be understood as intensifications of prior asymmetries. The algorithmic technologies that are engaged in the contexts discussed by Zuboff are simply off the conviviality scale: here purported human 'end-users' are being surreptitiously manipulated and instrumentalised for the benefit of techno-institutional ends. Human 'end-users' at the lower ends of technical hierarchies have little meaningful control over the technologies they use; they have no control whatsoever over all the ongoing reconstructions of the technical affordance landscapes within which they live their lives.[76] In a different mythological era, a similar situation would perhaps have produced a cultural judgement that the gods must be playing games with the lives of mere mortals. In our day, the positions of old gods are occupied by tech-overlords – that is, human billionaires and/or collective hybrid agencies.

Many will no doubt object to Illich's conviviality ideal. Indeed, for many inheritors of Enlightenment thinking, families and villages, traditional conviviality-based structures, will often be understood as structures that typically limit the exercise of individual autonomy. Illich, it must be emphasised, envisages a convivial modernity, a condition that would combine conviviality with modern knowledge and know-how while putting limits on industrial growth. It must also be admitted that the exercise of any agency, including the

[75] Some of the more recent radical monopolies are the personal computer and the smartphone. Two emerging phenomena that could evolve into radical monopolies include smart vehicles and generative AI.

[76] An interesting argument could be made on the basis that the algorithmic economy has a fundamental need of humans. Learning machines need humans in order to learn, and the structures criticised by Zuboff need humans in order to make money. Jaron Lanier has argued, on this basis, for something that could be conceived as a less asymmetrical algorithmic economy, in which algorithmic systems do not spy on and steal knowledge from humans, but instead humans have the option to share information voluntarily and to get paid for what they share. Lanier's original ideas can be found in *You Are Not a Gadget* and *Who Owns the Future?* (Lanier, 2011 & 2013).

kind of agency conventionally understood as individual autonomy, will always be limited to some extent by the contexts in which it is exercised. If this is the case, then which type of context, set of circumstances, or affordance landscape, in the short, medium, and long terms, will be more and which will be less likely to favour any given set of ideals, be they ideals pertaining to conviviality or to any other overarching aim? This is the crucial question that, if we accept Illich's understanding, we must learn to ask regardless of whether or not we agree with Illich's overarching convivial aim.

Modernity's preference for modern structures (corporations, administrations, free enterprise) has largely been grounded in the assumption that a meaningful role will always be available for rational and well-educated human agents. By filling such roles, we can receive a form of validation of our social worth and a monetary fungible reward, which correspond to a rough measure of our individual power. In the terminology of multi-agent organisations, this modern assumption implies that, in complex social hybrid multi-agent organisations, there will always be functions that are impossible to automate. This assumption is becoming less and less tenable.

Here lies the tension between the industrial-modern way of thinking and Illich's vision of a more convivial modernity. For the former, individual autonomy is likely to be limited or even frustrated by archaic and outdated cultural structures. Industrial progress is, in a sense, coextensive with individual freedom. By means of the industrial application of instrumental rationality, society could liberate itself from stunting outdated structures and thereby be in a position where its members can exercise their individual autonomy more fully. From Illich's perspective, such liberations from relational, local, and purportedly outdated bonds could be understood as a sleight of hand: yes, on the one hand liberated individuals obtain a new scope for the exercise of their agencies. They become, in a sense, freer, especially in cases where traditional cultural structures put severe limits on individual autonomy. But it is a kind of freedom tied to fungibility and ultimately to serfdom and the loss of autonomy. This type of freedom, as we see later when we consider Ellul, comes with a price, and that price includes the transfer of autonomy from relational structures of personal interdependence, such as family and village structures, to industrial and bureaucratic systems and organisations.

***

Modernity is considered in more depth in part III. Suffice it to say now that high modernity envisaged important tasks for humans as reformers, engineers, and thinkers. There was important work for modern agents to do in modern institutions. This modernity suggests new heroic ideals, meaning, and a purpose for aspiring modern humans. Illich's convivial ideal is not incompatible with this kind of generic modernity. It is incompatible with an overly bureaucratised and overly-industrialised modernity. In part III, high modernity is

distinguished from late modernity. One characteristic of the latter, it is argued, is that the status of values tends to become less absolute and that the human condition is increasingly defined by an experienced need of efficient adaptation to ever more rapidly evolving circumstances. One possible explanation for the transition to late modern conditions, it is suggested shortly, is the increasing prestige of the technological means developed to realise the values held by modern societies. As technological means become ever more impressive, generally held values tend to become softer relative to technological means. In a word, values become *adaptable* to the evolving conditions that are defined by technological innovation and application.

Illich's hopes for a convivial modernity notwithstanding, convivial patterns seem to be most likely to emerge in pre-modern structures of limited scales, as, for instance, in old-style villages and guilds. Since 1973, the year that *Tools for Conviviality* was published, most if not all trends have surely pointed away from conviviality. Industrialised societies have pursued increased industrialisation, and are presently busy pursuing digitalisation and ubiquitous computing. Such impressive feats of systemic engineering do not occur within 'limited scales', which Illich posits as a prerequisite for conviviality, but are undertaken by giant conglomerates of state and corporate interests, which subsequently tend to keep tight control over the maintenance of products and services.

In associations of personal interdependence, when convivially structured, there will be a commonality of purpose, such that purposes pursued by individuals align to important extents with purposes pursued by the collective as a whole. Prospects for the long-term survival and prosperity of individuals will be inextricably intertwined with the long-term survival and prosperity of the collective. As humans leave old relational modes of organisation for the purpose of pursuing careers in bureaucratic and corporate organisations, an important dissociation occurs between individual and collective interests. We should take note of this shift. In organisations of limited scales, such as family, guild, and village structures, assuming individual participation is voluntary, the practical ends of the collective will often be aligned with many of the practical ends pursued by the individual agents who compose them. The individual has a stake in the survival and prosperity of the collective; and for the collective, many of the individuals who compose it will be unique and therefore irreplaceable. Ideally, we could say that individual long-term interests then align with the long-term interests of the organisation of which the individuals are members; in order to satisfy such long-term interests, as always, individual and organisational short-term interests may have to be sacrificed.

Large-scale organisations, on the other hand, will tend in time to dissociate themselves from anything that is tied to the individual interests of the members who compose them. They will pursue grand and abstract goals. Often the visions of large corporations become so grand that they are impossible to realise. Illich asserts that '[t]he setting of abstract impossible goals turns the

means by which these are to be achieved into ends' (Illich, 2021, p. 40). And it is precisely when means turn into ends that we can begin to talk of the autonomy of *technique*. When the efficiency and power of technical means become pursuit-worthy ends in themselves, environments begin to be more aggressively reorganised in order to satisfy the technical requirements of this pursuit. When the functionality of technical systems and technical organisations become more pursuit-worthy ends relative to other modes of human organisations, the degree of technical systemic autonomy increases relative to that of other modes of human organisation. More and more, the agency-venues available to the human agent – liberated from traditional obligations towards family, kin, and region – are to be found in the service of technical and bureaucratic organisations – modes of organisation that are structured so that they facilitate the expansion of all manner of technical systems.

The measurable reward for individuals in service of this type of organisation is money, a fungible commodity that, in theory, can be used to advance individual short-, medium-, and long-term interests. However, the ultimate goals pursued by corporations and other bureaucratic institutions, in contrast to the goals pursued by a family or a village, will rarely have anything to do with individual interests. Organisational goals will frequently be at odds with the individual interests of the human agents who enable them to be accomplished, and they might even become parasitic on the well-being of the very humans who are recruited in order to enable their realisation. In any case, regardless of the ethical status of the collective goals pursued by such organisations, the nature of the individual–institutional relationship has changed: it has become entirely mercenary.

Today, the potential of having AI at our human disposal may give the impression that human agency will be radically expanded and that humans *qua* agents, therefore, will be vastly empowered. If we pay attention to the analysis of Illich, we will see that this pattern is not new – that it, in fact, is a repetition of industrialisation and its promises to liberate humans from difficult and oppressing life conditions. Liberated we may have been, but only to fall into a new dependence: a dependence on the very means used to liberate. Whether this new dependence should be understood as a gain or a loss relative to old dependencies can be discussed. Regardless, we should heed this lesson when we envision human agency and autonomy in algorithm-intense environments.

Jacques Ellul and the autonomy of technique

As dependence on technical systems increases, common sense suggests that the functionality of technical systems ought to be continually improved and rendered more efficient. Illich argues that there are limits to the extent to which this could be accomplished without undermining conviviality. But technologically advanced societies have generally not heeded Illich's recommendation. Instead, they are increasingly busy exploring new frontiers of technical

possibilities, many of which require ever higher degrees of specialisation and bureaucratisation. The prospects for individuals not controlled by institutional tools but in control of *their* tools seem even weaker than in the days of Illich.

To the extent that the aim for industrialisation is not already embedded in a people's worldview as an ideal or ethical value, the analysis of Illich has illustrated how, as humans lose touch with prior life-sustaining cultural patterns and instead become dependent on industrial technologies, it can become an aim that is experienced as a practical necessity. If, on the other hand, the aim is already there as an ideal, through the same dynamics it becomes reinforced by practical necessity. The pursuit of higher efficiency, Ellul argues, *is* the driving force behind the development of modern technology. Ellul distinguishes between '*l'opération technique*', which concerns any process in which a method is used in order to achieve an end, and '*le phénomène technique*'. The latter can be understood as a sort of mindset that first emerges in parallel with technological development, and then plays a leading role in driving the development further: it frames the search for the best means in all domains. The accumulation of means, under the supervision of this peculiar mindset, produces a technological civilisation (Ellul, 1990a, pp. 17–18).

On Illich's understanding, the introduction of something akin to industry in the social fabric produces a rupture of long-established social patterns. Ellul locates the origin of the rupture in the industrially produced machine and, by extension, the factory.[77] The perfection of machines, Ellul argues, requires the further perfection of the industrial production process of machinery. This produced the scientific management of Taylorism, in which the factory is considered a world of its own, closed off from the social world in which it is located.[78] The separation of the mechanism to be studied and perfected from the wider context in which it is located is, for Ellul, at the origin of the autonomy of technique (Ellul, 1990a, p. 121). In the new technical form of organisation that gains traction, questions pertaining to good and evil, to just and unjust,

[77] According to Lewis Mumford, mechanisms such as animal-, wind-, and water-powered mills, if we understand them as machines, are not as anomalous as their industrial offspring vis-à-vis traditional societies. Ancient wind- and water-powered mills require cooperation with natural forces. Animal-powered mills are perhaps a step towards something different, in that their human users aim in a different sense to control the power source for the mill – that is, the animal agents. In all of these instances, all that is required in order for such primitive machines to function – from the skills that enable their construction to the operators, power source, and raw material to be processed – is available in the vicinity of the machines. Compare this with the conditions for modern machinery. Industrially produced machines are introduced into local environments, but they are produced in specialised environments by technicians with specialised knowledge, and they are powered by energy – oil, electricity, etc. – the production of which requires additional specialised knowledge. If the ancient type of mills required cooperation with naturally occurring wind- or water-patterns or with animal agents, these new types of machine could be understood as possessing various degrees of autonomy vis-à-vis the local environment in which they operate. For a problematisation of mechanisation and sources of power, see (Mumford and Winner, 2010, pp. 60–118).

[78] Taylorism derives its name from its pioneer, Frederick Winslow Taylor.

become irrelevant. In the closed-off world of the factory, the guiding priority for all activity is the search for laws that determine technical efficiency.[79]

In considering the ideas of Ellul, we must juxtapose them to what has already been treated and, later, subjects yet to be treated. The theoretical frame developed within the context of this treatise provides a comprehensive lens through which we can interpret the relations between human beings and their environments. By drawing on the perspective of Ellul, and applying some of his insights to the multi-agent system- and affordance-based edifice developed previously, and by including the perspectives of Zuboff, Illich, and others yet to be treated, a well-nuanced and holistic understanding is developed, one that considers the conditions necessary for technique and technology to function properly, the social implications of technical macro-structures, the role of concrete technologies in everyday life, and new possibilities and temptations that arise as a consequence of new technologies.

The rise of the autonomy of technique, as Ellul understands the matter, is *relative* to a loss of autonomy *of other spheres of human activity*. It must be recognised that 'technique' also belongs to the category of 'human activities'. The rise of the autonomy of technique, then, should not be understood as a rise relative to a loss of *human* autonomy in some abstract sense. Instead, as societies recognise its importance, technique gains in degrees of autonomy relative to other spheres of human activities, such as economics, politics, and social conditions. If those spheres have previously tended to determine technical development, Ellul argues that technological societies reach a point where the relationship begins to change:

> Technique elicits and conditions social, political, and economic change. It is the prime mover of all the rest, in spite of any appearance to the contrary and in spite of human pride, which pretends that man's philosophical theories are still determining influences and man's political regimes decisive factors in technical evolution. External necessities no longer determine technique. Technique's own internal necessities are determinative. Technique has become a reality in itself, self-sufficient, with its special laws and its own determinations. (Ellul, 2021, p. 133)

At the time of Ellul's writing, then, the autonomy of technique is still a societal phenomenon. It involves humans. If we accept this, then the extent to which the goal of some AI researchers is realised, that of removing the human from the loop, may also determine the extent to which it will remain a societal phenomenon.

But should we accept this description? In our own day, is it really the case that technique *determines* economics, politics, and social conditions? This is

[79] This type of process may have contributed in the transformation of modernity, from a classical Enlightenment understanding that holds a range of values and rights to have universal validity to that which is retained in late modernity, where only scientific and technical laws *really* have universal validity.

surely far too strong an assertion. A cursory consideration of the rising and declining technical prowess of nations would seem to be enough to contradict it. Is it not irrefutably the case, rather, that deteriorating social, political, and economic conditions frequently undermine the viability of technical research? And is it not the case that, when such conditions improve, the viability of technical research tends to improve with them?

We must consider these categories as dynamic and interdependent sets of conditions. It cannot be a question of a binary option between either technique absolutely determining policy or policy absolutely determining technique. A dynamic understanding implies that, in technological society, technique increasingly sets the conditions to which economics, politics, and social conditions must adapt, rather than the other way around. This in no way implies that the latter are unimportant. Social conditions, politics, and economics are preconditions for technical research and production. That makes them, in a relative or instrumental sense, supremely important in pursuing technological innovation. As their importance relative to technical excellence is recognised, it becomes imperative, in technological society, that they be adapted so as increasingly to serve the requirements of technique.

A telling example of these types of dynamics could be provided by considering general social, political, and economic attitudes vis-à-vis AI research. Few so-called advanced nations – regardless of the political colour of their governments – seem to be willing to abstain from participating in AI-driven change. It is widely understood that AI will have profound implications for all spheres of human activity, including social conditions, politics, and economics. The exact consequences of present and future iterations of AI are unknown, yet most societies appear to be perfectly willing to make sacrifices in order to enable AI research to flourish. The sacrifices include present and future adaptations of social conditions, politics, and economics. In the short term, AI promises to empower some potential stakeholders and consumers. These are the obvious short-term benefits. From this perspective, AI technology could be understood as instrumentally serving the interests of present stakeholders and consumers. In the long term, however, it is widely understood that just about everything will have to be adapted to radically new conditions set by newer and newer iterations of AI technologies. People project both hope and fear into the unknown AI-intense future. There is no real debate about whether or not we ought to move towards it. It is as if it is simply dawning upon us, like fate. We have no option – whether we like it or not – but to give up, or, rather, to adapt current social, political, and economic conditions, or current ways of life, for the sake of participating in the inevitable development or evolution of an AI-intense future.[80]

[80] If we think that this is as it should be, then we seem to be implying that we do not have anything at the present that is really worth treasuring or preserving. Not, that is, anything but the goal of technical innovation. This, it could be argued, is a form of nihilism. It represents a

Or this would be at least a viable interpretation of current dynamics through the lens of Ellul. One could object that, in many instances, AI is in fact being adapted to current norms. We are not, one could argue, giving up anything; rather, we are improving present conditions and ways of life not only by adding a new powerful technology, but also by adapting that technology in order that it aligns in important ways with current norms and values. Nevertheless, in view of its long-term effects, it is difficult to imagine how the construction of AI-intense environments will affect individuals and human populations. To the extent that it is practically possible, it is hardly surprising, even if we accept Ellul's interpretation, that there would be attempts to adapt new technologies in order that they become more compatible with current norms. But do we care so much about current norms that we are willing to sacrifice technological innovation in order that current norms might survive in the long term? Marginal Amish communities, even today, are prepared to make this sacrifice. The mainstreams of technically advanced societies are not. Technically advanced societies have undergone a sort of inversion of traditional societies – Amish communities included – in that, more and more, rather than technique being evaluated relative to how it favours cultural ideals, cultural conditions are being evaluated relative to how they favour technique. Again, this must not be understood as a binary either/or, but as dynamic and complex conditions that exercise mutual influence upon one another.

So far, Ellul's analysis seems to concern very broad societal categories – that is, general fields of human activity, of which technique is one. It is not yet clear how this relates to the previously discussed dynamics concerning the autonomy of human agents and the autonomy of social structures such as the family, the village, and the corporation.

Humans have been practising techniques, making tools, and using tools since time immemorial. Nevertheless, complex machines could be understood as something of a historical anomaly in the lifeworld of human beings. A machine is not a simple tool. It is a complex of parts that, like living organisms, can move with unity of purpose.[81] The functionality of machines requires that

surrender to the pursuit of power – but power as an abstraction, since the extent to which the power being created will ultimately benefit actual humans is unknown. Nevertheless, according to the mindset that seeks the best possible means in all domains, to abstain would mean to become stagnant, uncompetitive, and weak relative to other agents. To the extent that this remains the case, and no other countervailing values are brought to the fore, technique will no doubt continue to drive change in politics, economics, and social conditions.

[81] Can a machine be understood to be 'alive'? Drawing on the insights of Jonathan Pageau, we can understand a machine to be symbolically 'alive' in the sense that, like other living organisms, it consists of environmental stuff gathered into a form of organisation that expresses a unified purpose. The biblical language for 'environmental stuff' is 'dust', and we are informed in Genesis that we came from dust and will return to dust. Death, then, could be understood to manifest when there is disunity of purpose. Actual death occurs when the dust that formerly composed the living organism no longer expresses any purpose whatsoever (Pageau, 2024). To the extent that machines are made for purposes, this symbolic life cycle could also be applied to machines.

adequate pre-conditions be instantiated. If human beings live, thrive, and die in life-worlds, machines function in machine conditions. In the modern industrial era, the perfection of machines requires countless interventions not only in the interior design of machines, but also in areas that concern the conditions under which machines are produced and the conditions under which they are intended to operate. In order to achieve optimal machine functionality, environmental and, ultimately, human conditions must be adapted to the needs of machines. 'Technique', Ellul clarifies, designates the totality of means and methods that adapt environments, including human beings, to the functionality of machinery (Ellul, 1990a, p. 2).

The Taylorist factory is an early example of a major feat of technique. In Frederick Winslow Taylor's organisation, closed off from the organic, cultural, and social world in which it is located, environmental conditions and human agents are scientifically managed in order to create the best conditions possible for technical functionality and efficiency. The industrial factory is a multi-agent organisation in which the optimal functionality of industrial machinery can be understood as a tactical goal. Specialised humans assume the role of the agents who make it work. Whereas the people in my father's village learnt many skills that were generally useful in the convivial contexts of family and village cohesion, factory workers need to learn narrow skills that are useful only in specific industrial contexts at various stages of industrial development. Technique, for Ellul, encompasses all kinds of means – educational, propagandistic, diverting, pharmaceutical, policiary, etc. – that can be used in order to re-adapt agents who are accustomed to agricultural village conditions to industrial factory conditions. *Technique, in service of efficiency, re-adapts the way that human agents express agency.* If they happen to be learners of skills that are generally useful in the context of local or regional convivial structures, techniques can be used to make them become learners of skills that are narrowly useful in the service of technical efficiency. In the historical process as it played out, industrial agents also became, as part of the bargain, professional wage-earners. In the earlier stages of industrialisation and automation the demand for human agents was high: there were no other types of agent that were adequate for the purpose of developing and maintaining the functionality of industrial machinery.[82]

If techniques, in the service of overarching efficiency, were previously used in order to re-adapt the way human agents express their agency, adapting human populations to emerging technical conditions, then at the present both sides of this equation appear to be subject to transformation: algorithmic machines can be used in order to influence human agents, and, as artificial agents

[82] If human agents remained in high demand, a grim fate awaited other types of agent that had previously been in high demand, especially work horses. The tragic fate of horses could be interpreted as an omen of one possible fate of human agents in algorithmic-intense environments.

are increasingly able to do the work that human agents used to do, the controllers of algorithm-intense multi-agent organisations may find it expedient to influence humans to accept other roles.

That technique is becoming autonomous means that all of the processes and activities that revolve around technological efficiency begin to form a *world* of its own. Technique obeys no cultural tradition. Technicians are busy discovering *its* laws. Moreover, the growing accumulation of techniques and their ongoing evolution is so rapid and revolutionary that the cultures in which the evolution occurs do not have time to adapt and integrate technique into their local traditions. If we are accustomed to think of the effect of a *particular* technology or machine as predictable, this is in no way the case with the effect of *technique as a whole* (Ellul, 1990a, p. 12). Increasingly, Ellul argues, it becomes the case that technique effects unpredicted changes in the ways of life of those who use it or who are exposed to it. A technological society is one governed not by allegiance to conviviality or to any other cultural or ethical ideal, but by the means of technique deployed to accomplish the end of adapting environments for the benefit of optimal machine conditions.

It should not be too difficult to find examples from our current times that could serve as counter-evidence to such sweeping generalisations. First, it must be said that since technique, as Ellul defines it, represents an accumulation of means, nothing in principle would prevent us from using techniques in order to destroy machinery or otherwise frustrate the progress of technological society. Even if such use of techniques may indeed occasionally occur, as in the case of the historical Luddites, nonetheless it represents a marginal use by fringe elements in technological societies. The question that has concerned us here is not how technique *could* be used, but how it generally *is* or *is likely to be* used in the context of technological societies. Is it generally used in order to adapt environments to machines? If Luddites are rare, the use of techniques for multiple other purposes must be very common. PR techniques, to mention only one example, are frequently deployed on behalf of all kinds of clients with all kinds of interests. Does this not undermine Ellul's analysis?

If we are to use Ellul's understanding as an interpretative lens, then again we must be careful not to interpret the lens itself as a sort of binary either/or categorisation. That technique is becoming autonomous means that it is becoming *more* autonomous relative to other fields of human activity. It does not mean that everything in society will be immediately and efficiently adapted in order to instantiate optimal machine conditions. Technique does not become incarnated in the form of an all-powerful dictator. Human society, with all its diverse forms and concerns, still exists. Rather, as Ellul expresses it elsewhere, one could understand it as technique and increasing layers of machinery grafting themselves on to human society (Ellul and Porquet, 2004,

pp. 29–30). Technological imperatives, then, do not determine all events outright; they increasingly shape events.[83]

Another possible objection could be made on the basis of the observation that machines are frequently adapted to human needs and to human psychology. This by no means undermines the analysis of Ellul. Both machines and humans are constituent parts of environments. If the overarching goal is to adapt environments for optimal machine conditions, then it makes perfect sense to adapt not only human behaviour but also machine behaviour, so that the two types – humans and machines – function as smoothly as possible together. If we revert to Zuboff's analysis, there is a potentially more sinister angle to the phenomenon of the adaptation of machines to humans. That which we may perceive as easy-to-use or human-friendly may also serve the function of a Trojan horse. The algorithmic economy analysed by Zuboff could be understood as instances in which machines with high degrees of autonomy use *techniques* in order to adapt humans to the functionality of machinery. In the case of surveillance capitalism, the adaptation consists of rendering human beings more predictable. This seems significant, for a key feature of functional machinery is predictability. *If, at the time of his writing, Ellul differentiates between technique and machines, we are now, it seems, moving towards an era in which the use of technique increasingly becomes incorporated into machinery that is at the stage of attaining ever higher degrees of autonomy. In the age of AI, machine behaviour becomes increasingly unpredictable to humans. Simultaneously, machines are, presumably, rendering humans more predictable*. This, at least, seems to be the implication of the analyses conducted so far. If this is so, then these changes represent a significant break with modernity as we have historically known it.

We are aware of many unintended consequences of technical systems that figure prominently in public discourse, such as pollution and climate change. Ellul's analysis intersects with the problem that Wiener identified earlier: the more we automate, the more acutely we will find that the agents in charge of technology will lack the cognitive capacity to use technique predictably and wisely. If Ellul is correct, then, in view of the medium- and long-term effects of technique as a whole, no humans will be available to use it predictably. Technique will to some extent, like nature, follow an evolutionary or co-evolutionary pattern. At least this will continue to be the case for as long as the mindset of seeking the best possible means in all domains remains the

[83] Some concrete technologies may, at first glance, seem to contradict Ellul's analysis. Consider, for example, the washing machine. The obvious effect of washing machines in everyday life is to reduce labour requirements in homes; it liberates humans so that they can pursue other ends, including professional careers. This taken into account, Ellul's point is that the washing machine also generates new needs. For the prospective consumer of washing machines and other industrially produced goods, the need is to acquire the means necessary to purchase them; and in order for washing machines to be properly produced and maintained, a whole range of needs arises pertaining to human servicing of the industrial structures that produce and service machines.

dominant mindset in technological societies. But if we accept the analysis of Illich, then the degradation of a more general usefulness of humans in convivial contexts is making humans dependent on ever more efficient industrial technology, and, therefore, it seems likely that a practical need to embrace this doctrine will grow stronger.

Ellul argues that our civilisation has become a civilisation of means, in which means are simply more important than ends (Ellul, 1990a, p. 16). This is another way of saying that the means have become ends. This, according to Ellul, differentiates contemporary civilisation from all previous forms of civilisation. From an evolutionary perspective, one could argue that all civilisations have been civilisations of means. In pre-modern cultures, things such as family continuity and cultural autonomy could probably be understood in many instances as important ends in and of themselves. Evolutionarily, on the other hand, a good case could be made that the pursuits of such ends have proven to be instrumentally useful as means for survival. Collective forms of cultural organisation have manifested the most viable contexts in which human individuals have been able to survive and procreate. This implies that ends that, from a cultural perspective, are valued as pursuit-worthy in and of themselves could also be understood from an evolutionary perspective as instrumentally useful – as means – to the ultimate end of survival. If we adopt a purely evolutionary understanding of means and ends, then technological society could be regarded as a new form of collective organisation in which human beings may or may not prove to be able to survive and thrive. As humans develop new means – or as new means evolve – it is only to be expected that humans will find new ends for which the new means are fit. On this understanding, there are no permanent ends. On the other hand, Ellul presumes some 'natural' or 'ideal' condition – a condition in which humans are able to exercise their agency freely. As technological society undermines that genuinely human condition, humans experience alienation. Within the framework of the present treatise, no firm position on any natural or ideal condition is taken, and it is understood that experienced alienation can also be the result of a worldview–environment mismatch.

An implicit attachment to inherited values, combined with an implicit or explicit embrace of the evolutionary perspective, could explain much of the ambivalent enthusiasm that animates many discussions about AI. Thinkers such as Nick Bostrom (2014) and Max Tegmark (2018) express fears about the possible consequences of AI. However, although many people do fear that an AI-intense society would be more dangerous and/or less humane than our current society, hardly anyone argues that we should therefore abolish the venture. If we opt out, it is customary to reason, we will be in a worse position relative to the actors who opt in. This implies that technology, or at least technology-mediated power, is more important than any inherited cultural values or ends that might be undermined by technological innovation. And in order

to justify the acceleration of AI-development, stakeholders are busy inventing new ends that fit the new means.

Potential transhumanist applications, such as cognitive enhancement and superintelligence, represent some of the more fantastical new ends that become theoretically possible owing to the emergence of new means. For many transhumanists, the realisation that such ends, owing to new means, become theoretically feasible then serve to justify the acceleration of AI research. The previous reasoning about ideology in relation to affordances suggests that such means-driven inventions of new ends or values occur all the time. As the algorithmic affordance landscape changed, the end of the smart home was invented. The end of the smart home took precedence over the previous and more convention-grounded end of the aware home. Previous ideology and/or worldview patterns were adapted to justify the novel end of the smart home. This is not meant to negate Zuboff's analysis, but to suggest a shift of emphasis. Following the reasoning of Ellul, it would in fact be possible to agree with Zuboff that something like ideology or worldview ultimately has priority. But it would not be the ideology of surveillance capitalism, nor even the ideology of capitalism – cf. Mozorov (2019) – but, rather, the worldview of modernity or of technoscientific progress in its entirety that carries on to frame matters so that new and unexpected outcomes follow at an increasingly rapid pace.

Earlier Kissinger *et al*. observed that, as we become aware of major gains that are achievable through AI, the option not to use AI may increasingly be seen as perverse. Ellul presents a number of features that characterise modern technique, some of which are of notable relevance to ongoing AI developments. One such feature implies that the choice of using technology becomes automated (Ellul, 1990a, p. 74). In other words, wherever AI can measurably produce desired outputs more efficiently than was the case prior to AI, we have no choice but to use it. *Technical ethics*, it seems, simply demands that the most efficient means be used. The aforementioned Amish illustrate the counter-example to the dynamics spelt out in the preceding paragraphs. Here local communities decide to subordinate technique – the total assemblage of means at their disposal – to the end of a peculiar ideal of local cohesion; such communities thereby opt out of technological society and become marginal exceptions to it. Their choice about using technology, therefore, is not automated – or one might say that it is not automated with reference to using the most efficient means. They have other imperatives to which the efficiency of means is subordinated. If, on the other hand, we wish to stay and be a part of technological society, our choices will be, to a large extent, if not automated, then at least biased in favour of using the most efficient means. We must, we feel, opt in.

Here we must add that, in the case of AI, we may not be witnessing only the automation or bias of the *human* choice to use AI. Once sufficiently autonomous AI systems are in place, decisions to use technique might become

'automated' in a more rigorous sense. Humans may then simply be removed from the decision loops of artificial multi-agent systems.

Another feature that Ellul describes is the so-called 'unity' of technique. One technique produces the demand for another. *Le phénomène technique* embraces all techniques; it constitutes a whole. Whereas a particular technology can be made for a purpose, the evolution of technique can be understood as a causal process, in which new techniques are responses to the requirements created by previous techniques. Autonomous technique, then, constitutes a force in its own right (Ellul, 1990a, p. 90). To be sure, humans could intervene and influence the turn of events. But for Ellul it is unlikely that the development of technique and science would follow any ennobling trajectory that coincides with the betterment of human conditions.[84]

What could better illustrate the unity of technique than the emerging interconnectedness illustrated by Zuboff's examples? Gadgets that we purchase and place in our homes, in our pockets, and around our wrists are now increasingly constituent parts of systemic wholes. Watches, telephones, and home appliances have gone from the status of separate tools for individual users to agentic limbs in hybrid multi-agent organisations.[85]

Ellul argues that, as opposed to non-technique-related values, technique retains universal validity in the contexts of technological society. Modern technique, moreover, has universal *appeal*, even at the outer margins of technological societies. Wherever modernity encounters more ancient cultures, technique ends up gaining the upper hand, undermining local culture (Ellul, 1990a, pp. 106, 113). Some of the processes by which this can occur have already been described by Illich. To this we could perhaps add the plausible and non-negligible attraction that modernity will tend to exercise on pre-modern subjects. To the encultured pre-modern person who stands at the edge of two worlds, modernity promises liberation from constraints and hardships and the potential of unimaginable riches and liberties. Modernity could, in a sense, be perceived as an individual power-booster compared with the collective power of pre-modern cultures. At the same time, one could argue that all the features that characterise modern technique contribute to boost the autonomy not so much of individuals as of technique. The collective power that previously

[84] As an example, one could cite a category of techniques, such as 'police techniques'. Ellul argues that police techniques, if they are allowed to run their course, will transform entire nations into gigantic concentration camps (Ellul, 1990a, p. 93). The existence of other categories of technique within the whole of technique makes the evolution of technique as a whole less predictable.

[85] Here, again, we can draw on the insights of Pageau (cf. footnote 81). Disparate and seemingly discrete gadgets in public spaces, in our homes, and on our bodies also appear to be in a state of communion with one another. Unified as limbs in a body, these disparate and narrowly purposed gadgets can also express the unity of a higher purpose. We can understand this in analogy with how purposeless stuff, or dust, becomes *alive* when it is organised in such a way that it expresses a unity of purpose. To the extent that the analogy is on target, our inanimate life-milieu, it would seem, is coming symbolically alive.

inhered in pre-modern institutions increasingly inheres in technique. The autonomy of technique, according to Ellul, is so strong that technique accepts no moral judgement or limitation. Moral judgements, on the other hand, are frequently passed on human agents who 'misuse' technologies. This has relevance for the expected evolution of multi-agent systems. If we follow the line of reasoning pursued here, then there is a major potential glitch that would need to be eliminated from current hybrid multi-agent systems, namely, the non-predictability of the non-technical human agent.

> This progressive elimination of man from the circuit must inexorably continue. Is the elimination of man so unavoidably necessary? Certainly! Freeing man from toil is in itself an ideal. Beyond this, every intervention of man, however educated or used to machinery he may be, is a source of error and unpredictability. The combination of man and technique is a happy one only if man has no responsibility. (Ellul, 2021, p. 136)

If the human agent is to have a fruitful role in the context of such systems, it must, according to Ellul, be a human agent relieved, like technology itself, of all responsibility – that is, a predictable and interchangeable technical performer of well-defined subtasks in the context of a larger technical system or bureaucratic organisation.

If the industrial machine was originally an anomaly vis-à-vis the cultural and organic environment in which it was placed, then, according to Ellul, the encultured human person now increasingly becomes the anomaly in mechanised environments. Human beings are not evolutionarily adapted to cope with the fast rhythms of mechanised and algorithmic environments. Here we recognise something akin to Wiener's timing problem. The imperative to improve technical systems produces the incentive either, to the extent that it is possible, to eliminate the need for human agents in multi-agent systems, or, to the extent that it is possible, to mechanise human beings so that they become more compatible with the rhythms of machines. A third way consists in making technologies more adapted to how humans function. A priori, this may represent an alternative to the mechanisation of the human agent envisioned by Ellul. A process of mutual adaptation between humans and technologies might eventually enable the flourishing of the kind of partnership envisaged by Kissinger *et al*. The problem is that the adaptation of algorithmic technologies to human beings would no doubt require that, as we have learnt, technical systems integrate much knowledge about human beings. It is this third way, arguably, that is pursued to a great extent by the structures problematised by Zuboff. The smoother the technology-mediated experience is, the easier it should be to 'hook' human users. Paradoxically and unfortunately, this third way could both eliminate the need for human agents at important levels of multi-agent systems and organisations and effect the instrumentalisation – arguably a form of mechanisation – of human agents.

Ellul recognises that human societies never become completely mechanised. Technical systems are not the same as technological society. The latter is the kind of society that makes room for and promotes technical systems, that is always eager to adapt and change in order to accommodate new technical systems. Technical systems nevertheless need humans to function. Ellul argues that technical systems are grafted on to technological society in analogy with how the machine is grafted on to nature. The machine is made of nature, and cannot exist without nature; but the machine does not transform nature into machine. Likewise, society is a natural phenomenon, and technical systems do not abolish it as such; they do, however, begin to shape society to the effect that, primarily, the requirements of technical systems become satisfied. Industrial mining – another technique-intense system – is a prerequisite for machine production. The systems involved in industrial mining and machine production are grafted on to both nature and society. The machine interferes with a previous ecological order. And human society begins to adapt, to the effect that both ecological and social order are shaped for the benefit of technological pursuits. Thus technique has become one, but not the only, determinant factor in technological society (Ellul and Porquet, 2004, pp. 29–30).

The reasoning of Illich and Ellul complements that of Wiener and Susskind. Of these thinkers, only Susskind is open to a potentially happy or harmonious ending for the paths of technical development currently being pursued: a future of leisure. We must, however, consider the potential cost of such leisure. If the price for leisure would imply giving up much or all that we hold dear, then this could represent a sacrifice that we might not be, nor should be, willing to make. In a sense, Zuboff's description of surveillance capitalism represents one possible model for leisure in algorithm-intense environments.

Interestingly, surveillance capitalism also represents a potential corroboration of Wiener's hypothetical predicament. As we continue to automate functions and to increase the technical complexity of that on which we depend, Wiener argues that cognitive demands on humans will also increase. Surveillance capitalism could be understood to demonstrate the conditions under which this hypothesis is valid or not valid. As humans inhabit contemporary algorithm-intense environments, then the cognitive demands on them will increase, provided that they value some cultural or political life-forms or ends to the extent that they want to preserve them or bring them into being. If, for instance, we were to share Zuboff's values, then we would probably need to become aware of all the complex and surreptitious structures that interfere with and define our lives, lest we lose our individual autonomy. We can no longer be simple users of consumer products; we must learn about the complex ecology of technical systems, how technical systems are used vis-à-vis humans, and the various roles that humans can enact vis-à-vis and/or as part of hybrid multi-agent systems. We must, in a sense, become like the Amish, but in a life-milieu that implies much heavier cognitive demands on the humans who inhabit it. We do not need to embrace *all* Amish values, but rather their

attitude vis-à-vis technology. Like the Amish, we must learn to evaluate critically the functions that different technologies will fill in the context of the way of life that *we* happen to value. If we value a way of life more than the blind evolution of technical systems, then we must also learn to subordinate technologies to our higher values. Susan Maushart, in *The Winter of Our Disconnect*, implicitly embraces this attitude in a secular family context (Maushart, 2011).

On the other hand, for humans who inhabit contemporary algorithm-intense environments and who do not value anything in particular, the cognitive demands will not necessarily increase. If the inhabited environment happens to be designed so that it is pleasurable, the value-less human inhabitant may even be in for a smooth ride. In such cases, it may make some sense simply to hang on to the evolutionary process and to discover where it leads. Most humans, however, will at least value survival. Therefore, even persons who embrace a relatively nihilistic evolutionary view will tend to take some risks seriously. Thus the development of ever more powerful algorithmic systems compels us, at the very least, to think seriously about potential risks. But here we do not need to become like the Amish. Edward Lee chooses instead to embrace the concept of co-evolution, a progressive and forward-looking view of human beings and machines as mutually interdependent in co-evolutionary processes (Lee, 2019).

If earlier waves of industrialisation enabled new forms of human society, in which one human-intense and human-centred social order (e.g., family and village life) was supplanted by another human-intense and human-centred order (industries and bureaucracies), then this, according to Russell, Susskind, Kissinger, Schmidt and Huttenlocher, appears to be about to end. From now on, they all agree, the human-intense or human-centred mode of organisation is over. It appears that there are no new skills that would enable most of us, of our own *human* accord, to become needed, productive, and valued members of the types of organisations that are emerging. Even if we disagree with Illich's ideals and instead favour the way of industrialisation, we now need to consider the new challenges that we are facing in algorithm-intense environments. If human skills will no longer be in demand, then, in a sense, we will become useless – useless, that is, with reference to the maintenance and development of the structures that sustain our survival and ways of life. Perhaps a productive and useful role will nevertheless open up to human agents, *qua* partners with AI. If, on the other hand, the predicament of Wiener were to deepen in algorithm-intense environments, then human skills and knowledge would become more in demand than ever. In fact, the cognitive requirements placed on human beings might then exceed human capabilities, so that, again, humans become useless – useless in that, in this case, we would be unable to cope with the complex structures that we ourselves have erected between us and the natural environment.

Illich and Ellul paint an even bleaker picture. In the words of Illich: 'The hypothesis was that machines can replace slaves. The evidence shows that, used for this purpose, machines enslave men. Neither a dictatorial proletariat, nor a leisure mass can escape the domination of constantly expanding industrial tools' (Illich, 2021, p. 10). The observations of Illich and Ellul describe a transfer not only of power but also of consequential agency and autonomy from human agents and convivial human contexts to industrial and technical organisations.

In earlier historical stages of this transfer, humans were perhaps *qua* individuals initially able to exercise fairly high degrees of autonomy vis-à-vis, and especially through, industrial and technical organisations. They could become wage-earners and receive high modern status. If Illich and Ellul are right, we now risk being subordinated and limited to degrees that extend, in important respects, beyond the limitations imposed by pre-modern family and village structures. Technical and bureaucratic organisations are increasingly turning into algorithmically defined multi-agent structures.

If we were to this add the apprehension of Wiener, then, in the wake of further mechanisation and automation, we could also expect unforeseen dysfunction owing to the limited cognition of human overseers. If we follow the logic of technological society spelt out by Ellul, this would only incentivise society to replace humans with algorithmic technologies. But is this a solution to Wiener's predicament, or does it represent a further aggravation of the predicament? To this, alas, no decisive answer can be given.

Algorithmic exploitation of features inherent in human nature

Wiener, Illich, and Ellul did not live to see the so-called age of AI. Yet they foresaw, from their respective angles, many of its challenges. If many of the effects of current algorithmic technologies appear to corroborate their understandings, then this suggests that, in some respects, there is continuity between current algorithmic technologies and the automation technologies that preceded them. Nevertheless, many are convinced that algorithmic technologies also represent something new and revolutionary.

Let us consider a much-discussed phenomenon that appears to have emerged as a consequence of some current algorithmic architectures: the encouraging of narcissistic impulses. Mark Coeckelbergh writes, on the topic of self-improvement, that 'if the current technologies boost this wrong kind of self-love, then we have a problem. If our screens become mirrors for our narcissistic self, we miss the chance for empathic engagement with others and an openness for the world around us: a non-destructive and much better form of self-improvement' (Coeckelbergh, 2022, p. 27). If some current algorithmic architectures encourage us to go down potentially destructive paths, how are we to understand the relationship between human agents and constructed or reconstructed environments? Are we to understand that phenomena such as

the encouraging of narcissistic impulses are deliberately intended by human developers of algorithmic systems? Or do they simply 'emerge' unexpectedly?

First, as previously argued, it is implausible to understand the agent–affordance relationship as proceeding exclusively from the free and abstractly rational choice of autonomous individuals. On this rationalistic view, new affordance landscapes offer rational and autonomous agents new ranges of possibilities to realise ends, and individual agents can then freely and rationally choose between them. The more plausible view is that new affordance landscapes elicit certain responses from given agents by virtue of the congenital and encultured nature of those agents and by virtue of their previous commitments. Rational thinking, depending on the overarching goal or goals for which it is exercised, can then be used to justify giving in to elicited responses or to dissuade action. Vis-à-vis human agents, but probably not vis-à-vis donkey agents, some contemporary algorithmic architectures appear to elicit narcissism-boosting responses. Some, but not all, human agents give in to such elicitations.

If we were *naïvely* to consider some fairly sophisticated narcissism-eliciting architecture, structured like the systems that Zuboff describes, it would perhaps be too easy to intuit that the narcissism-inducing feature must have been deliberately designed by the system's human operators in order to hook users. Knowing how complex are the intricacies that define some agent–affordance relationships, we could offer the alternative explanation that the feature emerged unexpectedly as humans ventured into novel affordance landscapes. Knowing what we have learnt about the workings of algorithmic technologies, we could also consider the explanation that the feature could be a deliberate intra-systemic means developed for the purpose of attaining a higher goal. At this juncture, things get weird in novel algorithm-intense environments, for humans are no longer the only agents involved. Only humans can interpret and understand the meaning of phenomena as 'narcissism-eliciting'; yet narcissism-eliciting features, it seems, can now be devised autonomously by algorithmic systems. Algorithmic systems treat phenomena that humans interpret as narcissism-eliciting as value-neutral subgoals deemed to be instrumentally useful to the accomplishment of a main goal. This implies that humans could now be used as means by artificial agents. Narcissism-eliciting affordances, then, could be incidental to and/or deliberate in view of money-making or some other purpose. 'Incidental to and/or deliberate' is meant to convey the inherent ambiguity in the tension between an agent and its algorithm-intense environment: the narcissistic-eliciting feature may have been deliberately construed by human operators or incidentally effected by algorithms as an instrumentally useful means to *any* overarching goal.

In the light of this, how should we assess the plausibility of the algorithmically enabled world of leisure envisaged by Susskind? In such a world, most activities that we conventionally think of as work would be done by powerful

algorithm-intense multi-agent systems. This implies that we humans would have to surrender most of our native powers to such systems, or to those who are supposedly in control of such systems. If things then turn out relatively well, humans may find that they benefit in important respects from them; but they will also be at their mercy. In the sphere of human activities, the category of leisure would supposedly still be reserved for humans. The analysis of Zuboff, and phenomena like that of narcissism-eliciting algorithms, suggest that a future world of algorithmically enabled leisure would be morally problematic. This, however, is perhaps the lesser problem, for the analyses discussed in part II suggest that the future realm of leisure is likely to be reserved for autonomous human beings to no greater extent than the domain of salaried work. Many of our so-called leisure activities are already increasingly defined by algorithmic architectures. If current developments provide reliable indications of future paths, what should we expect of an algorithmically enabled leisured future? Who or what, for instance, would be best suited to propose leisure policies: human beings or AI? If leisure activities require the acquisition of any demanding skills, why would AI not be used instead?

One could respond by arguing that human beings are intrinsically active beings and that AI would not change human nature. Humans, then, will always *want* to learn, regardless of the quality of the AI systems that happen to surround them. This, it must be admitted, may well be the case. But to *desire to learn* and to *be motivated to learn* are two separate things. When I was a child, I daydreamed of having the skills of a superhero. At the very same time I often found it difficult to find the motivation to learn some of the more basic skills that were widely understood to enable members of society to become functional and valued social agents. Much of the motivation that I did find no doubt derived from the expectation of becoming, through the acquisition of skills, a needed and valued member of society. In the world that Susskind envisions, all such incentives, it seems, must vanish.[86] Critical knowledge and skills are therefore likely to recede from human beings. Human beings, then, will most likely be – in a debilitated and/or domesticated state – at the mercy of powerful algorithm-intense multi-agent systems and organisations.[87]

Kissinger *et al*. at least seek contexts for more substantial functions for humans. If their hopes were realised, humans would perhaps be better off than

[86] Would people not then pursue knowledge for the sake of intellectual curiosity, as the leisured Greek philosophers did? Some no doubt would. Regardless, since little or no knowledge or skill would actually be needed, individuals would be free to pursue their own idiosyncratic inclinations. To the extent that the algorithmic-systemic complex were to understand such sprawl as potentially dangerous, it would make systemic sense to modify the inclinations and behaviours of human populations in order that they become more uniform.

[87] Why 'and/or'? A debilitated human state, open to unpredictable, slothful, and potentially harmful spontaneity, may actually impede the functionality of algorithm-intense multi-agent systems. From a functionalist point of view it would then make sense to domesticate the debilitated population. But from the point of view of many presently dominant norms, such domestication would probably produce yet another debilitated state.

in Susskind's future. Partnership not only sounds nice; it also implies some socially useful purpose for human activity. If there are any perennial features that have characterised all varieties of human cultures, then this must surely be counted among them: that human knowledge and skills have been used for socially and collectively useful purposes.[88] The notion of human–AI partnership, then, could represent a continuation of this perennial pattern. What kinds of partnership between humans and artificial agents, then, are likely to evolve as a result of human interactions with evolving algorithmic affordance landscapes?

We need to take into account that Kissinger, Schmidt and Huttenlocher represent elitist interests. If we can imagine instantiations of different kinds of AI that display varying degrees of intelligence and systemic power, then we must not forget that human beings also come in different configurations, and that human beings occupy different systemic positions. Could it perhaps be the case that the structures that Zuboff labels 'instrumentarian power' and 'surveillance capitalism' are already the product of a kind *or rather several kinds* of 'partnership'? Today, the kind of relationship that people are able to have with chatbots could suggest possible future hierarchies. Some versions of ChatGPT are available for free, but all data that is generated can be accessed by its corporate owners. Meanwhile, people at the higher ends of technical hierarchies have access not only to more advanced versions of chatbots, but they also wield control over vast algorithmic networks. At the lower ends of such algorithmic networks, or technical hierarchies, people may get the impression that their lives will improve if they use the services that are made available to them, or if they form 'partnerships' with phenomena that appear under the guise of artificial agents. If they get that impression, then it might be because the architectures are designed incidentally and/or deliberately to 'hook' users. However, use or partnership at the lower ends of technical hierarches are unlikely to mirror qualitatively the kind of use or partnership that will be occurring at the higher ends of such hierarchies. Zuboff tells us again and again that surveillance capitalist structures produce unprecedented asymmetries. At the lower ends of such hierarchies we may find lonely people who are delighted to be able to form partnerships with phenomenological agents that display intelligence, humour, beauty, and so on. In a wider perspective, such phenomenological agents may be functional limbs in a hierarchical multi-agent organisation, at the top of which people like Eric Schmidt form more consequential forms of partnership with more powerful AI systems. The following question raised by Zuboff's analysis remains, and it gains in gravitas: If a thermostat can serve hidden agendas, then what hidden purposes could more advanced artificial agents be serving? Whose interests will they be likely

[88] And if we keep applying the insight of Pageau (cf. footnote 81 and 85), then stuff or dust gathered for no purpose – including in the form of human beings, other animals, and machines – symbolises death.

to serve? Will we again be encouraged to commit to structures that satisfy short-term interests and/or cravings of users at the lower ends, as these very structures are surreptitiously and methodically satisfying the long-term interests of those in control of technical hierarchies?

Regardless of the ideological bents of people in charge of powerful algorithmic technologies, the mere construction of such technologies represents events that in and of themselves ought to cause us to worry. Contemporary denizens in technological societies are – or should be – familiar with the varieties of techniques that can be used for the purpose of behavioural control and behavioural modification.[89] Algorithmic technologies offer not only unprecedented opportunities to automate such techniques. By means of ever more potent systems, algorithmic technologies could increase the knowledge available to propagandists. Unsupervised learning, for instance, could give the totalitarian-inclined more knowledge about how to manipulate people to cooperate with ends that may be alien to them. And the ever deeper embedding of human subjects in algorithm-intense environments would offer new and unprecedented contexts in which humans could be propagandised and manipulated around the clock in all sorts of ways. Unfortunately for those who oppose totalitarianism, maximal technical efficiency, in the algorithmic contexts discussed by Zuboff, appears to require a totalitarian approach. If we seek to construct algorithmic systems that function well in social environments, these systems will need to be able to make good environmental predictions. In order to be able to make good predictions, they will need to *know* much about humans. Knowledge alone is unlikely to be enough: systems will also be required to *act* in order to make humans more predictable.

The embeddedness of humans in algorithm-intense environment deserves close scrutiny. Earlier, a pattern that appears to be repeated again and again was spelt out: first something is marketed and accepted as a convenience, then it becomes indispensable. Perhaps the final destination of this train of thought is *disappearance*? Mark Weiser is credited as the inventor of the concept 'ubiquitous computing'. In the early 1990s, Weiser pioneered a new way of thinking about human–machine relationships. Zuboff quotes Weiser: '"Machines that fit the human environment instead of forcing humans to enter theirs will make using a computer as refreshing as taking a walk in the woods […] The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable"'[90] (Zuboff, 2019, p. 198). Another way to think of disappearance is mergence. The clothes we are wearing – industrial-modern offspring of the old arts of clothmaking and tailoring – do not, fortunately, disappear. But they become part of us to such an extent that they become integral to our identity and purposeful

[89] See, for instance, Jacques Ellul (1990b), Edward Bernays (2011), and Richard Thaler and Cass Sunstein (2009).
[90] Zuboff quotes from Mark Weiser (1991).

unity. This is the case in a very basic sense even if we disregard additional fashion-associated identities: we simply cannot move around in society without *any* clothes. In cold regions, we risk our lives if we are not well-clothed. In a similar sense, it is becoming increasingly difficult to manage one's personal economy without the intermediary of various iterations of algorithmic technologies. The author of this treatise still has never used a smartphone. But how long before such non-compliance with emergent norms simply relegates people to marginal positions like that of the Amish? Like the added-on layer of clothes, new and continually evolving iterations of algorithmic technologies are merging with our personal, social, and legal identities.

Not all human subjects are embedded in equal ways. Some humans exert various degrees of control over the systems in which the majority of humans are embedded. Any kind of partnership between people who exert control over such systems and future versions of AI should cause us to worry. Zuboff has already given us an idea of what such systems can accomplish today. Next, we briefly consider two specific ways in which algorithm-powered systems could increase the control over human agency. Both involve the exploitation of knowledge of features that could be understood to be inherent in human nature. We could think of them as the way of an ever more efficient framing of human instrumental rationality, and the way of an ever more efficient manipulation and reinforcement of human desires.

Harnessing rationality and desire

Some thinkers, such as Ray Kurzweil, envisage future AIs that are able to operate completely autonomously from humans, AIs that will be able to sustain themselves and to evolve independently of humans. For now, at least, the functionality of algorithmic systems is inconceivable without human enablers. Larger technostructures, such as the organisation of Google, are not pure artificial multi-agent systems but human-artificial hybrid multi-agent organisations. The intentionality of the human agents who are in charge of such systems and organisations will naturally, to some extent, be social: it will, we can posit, therefore seek to engage in various ways with the larger social world that encompasses the organisational structure. This has been the case with all human types of organisation. Let us, again, consider algorithm-intense types of organisation in relation to their nearest predecessors, the modern administrative pre-algorithmic bureaucracy.

Governments, corporations, and other collective agents, in addition to more open and frank means of communication and persuasion, have always been able to resort to propaganda and behavioural modification techniques in order to influence humans both in and beyond their respective organisations to support and/or work for the collective ends of their organisations. The terms 'propaganda' and 'behavioural modification' are used in the broadest sense possible: the construction of a railway, for instance, could in some instances

be viewed as a form of behavioural modification technique, and any campaign that methodically seeks to influence humans to use it could count as integrative propaganda – that is, methods that seek to integrate humans into new technical life-milieus.[91] Translated into the terminology developed previously, this means that humans would be incentivised to use their agentic powers so that they served organisational ends, or that organisations harnessed the agentic powers of humans in the service of organisational ends. One might object: would not individual human ends also often be served? That they would, indeed. We shall see, however, that individuals could be influenced and manipulated to pursue ends that happen to coincide with the interests of the larger and more powerful collective agent. One of the criticisms that could be made against surveillance capitalist structures is the following: they produce contexts in which individual short-term ends coincide with corporate short-, medium-, and long-term ends, but not with that which *ought* to be individual medium- and long-term ends. If we make this argument, then the 'ought' hinges on some moral notion concerning ends that ought to be pursued, or on some notion of human flourishing that is impeded by surveillance capitalist structures. One such notion could be the desirability of the convivial life conditions favoured by Illich – life conditions that, in turn, would imply that local convivial structures would need to retain high degrees of autonomy. It would then be better for local human social bodies to retain high degrees of autonomy rather than to surrender autonomy to distant and centralised bureaucracies.

The techniques at the disposal of pre-algorithmic organisations, although powerful, have their limitations. To an important extent, the limitations of such techniques depend on the social structures in which their targeted population is embedded. To the extent that humans are embedded and well-integrated in pre-modern and/or convivial social structures, again, like Amish communities, they are likely to be less receptive to such techniques. Most of us, however, are embedded and well-integrated in the latest configurations of technical life-milieus. We depend less on extended family structures and to varying degrees on corporate, governmental, and other forms of modern organisations.

---

[91] For 'integrative propaganda', see Ellul (1990b). The purpose here is not to argue that it is fruitful to view every form of environmental intervention in terms of behavioural modification techniques. It certainly will often be more fruitful to understand the construction of a road or a railway from other angles. However, to the extent that a railway is built in order to make it easier for an administered population to travel a certain route, and to the extent that other techniques are used to make people choose that route, there is an intent to change human behaviour. In this case, road construction could thus be viewed as a behavioural modification technique that is deployed together with propaganda techniques. Techniques that are more intuitively understood as 'behavioural modification techniques' could, of course, also be deployed in the example given. Nudges may be given in order to prod people to choose the railway instead of taking the car on the old road. (For 'nudges', consult Richard Thaler and Cass Sunstein (2009).) Road taxes may be raised to discourage the use of cars. The old road may be left to fall into a state of disrepair so as to make it increasingly dangerous to use. In a civilisation reduced to means, anything and everything could be instrumentalised for the sake of ulterior ends.

The construction of algorithmic-intense life-milieus radically changes the contexts in which previously used techniques can be deployed. As algorithmic technologies become ever more potent, human subjects become increasingly embedded in algorithm-intense life-milieus. If recent trends provide reliable hints of future trends, then we can expect that powerful algorithm-intense organisations will become even more efficient in harnessing human agentic powers. Here we briefly consider two possible approaches to the harnessing of human agentic powers: the harnessing 1) of human instrumental reasoning and 2) of human desire.[92]

***

Harnessing of 1) instrumental reasoning. In the words of Bert Rockman, from the entry on bureaucracy in *Encyclopaedia Britannica*, 'bureaucracy' denotes a 'specific form of organization defined by complexity, division of labour, permanence, professional management, hierarchical coordination and control, strict chain of command, and legal authority' (Rockman, 2024). Whereas the earliest forms of bureaucracy had already arisen in the ancient world, there is a common understanding that modern societies, at some stage of their development, tended to become increasingly and exceptionally bureaucratised. Bureaucratisation is generally done for the purpose of rendering the administration of affairs more efficient, and it presumably becomes worthwhile when the affairs that require administration exceed some critical limits of scale. Today, when many corporations exceed the scale of ancient empires, the modern form of bureaucracy has become a necessity for the administration of governmental and corporate affairs.

Bureaucracy can be understood as a product of instrumental rationality applied to organisational behaviour in analogy with how algorithmic systems can be understood as a product of instrumental rationality applied to artificial systems. If we apply the paradigm of multi-agent system, we can see not only continuity between the two, but also potential evolution – away from one form towards another. The old-fashioned bureaucracy would be a pure multi-agent organisation run exclusively by human agents. Current technically advanced bureaucracies will be hybrid organisations run by bureaucratic human agents who increasingly apply artificial agents to more and more domains.

If bureaucracies have their undisputed advantages, many critics have pointed to their disadvantages. Rockman provides a brief resumé that includes the following rather unflattering perceptions of bureaucracy: 'excessive rules and regulations, unimaginativeness, a lack of individual discretion, central

[92] In addition to these and other possible subtle manipulative means, we must not forget formal means of legislation and police enforcement. To the extent that the interests of legislative powers align with those of corporate algorithm-intense organisations, the gathered momentum for radical structural change in the social fabrics of humankind may be great indeed.

control, and an absence of accountability (Rockman, 2024). Bureaucracies, furthermore, are perceived to be lacking in adaptability and to be prone to organisational dysfunction. Rockman concludes that 'the characteristics that make bureaucracies proficient paradoxically also may produce organizational pathologies'[93] (Rockman, 2024).

If, as suggested by the multi-agent paradigm, algorithmic systems can be understood as a continuation or extension of a more conventional form of bureaucracy, it may be worthwhile to consider, briefly, how human agency and autonomy can be affected in bureaucratic systems. But surely, one might object, we must not consider human agents to be passive objects at the mercy of bureaucratic systems. The very meaning of 'agent' implies an entity that is able to act on its environment; and human agents, so far, represent the most advanced type of agency. It stands to reason that human agents will control bureaucracies, not the other way around. Would not rationality, for example, provide protection to the human agents who exercise it from undesirable bureaucratic influences? Not if we listen to Zygmunt Bauman, who reaches the following conclusion: 'It appeared that when God wanted to destroy someone, He did not make him mad. He made him rational' (Bauman, 2008, p. 142).

In *Modernity and the Holocaust*, Bauman argues that the genocidal concentration-camp complex administered by the Nazis, rather than reflecting a regression to a more primitive pre-modern mindset, reflects instead another side of modernity. Modern genocide, Bauman argues, is different from primitive genocide, in that it does not represent an end in itself; rather, it is an end that is instrumentally useful for the accomplishment of a more important end. In the context of modernity, the ultimate end is a 'better' world, such as a racially purer and/or socially more just and/or technologically more excellent world. In order to achieve any desired utopia, impediments to its achievement must be overcome or eliminated. The means used to realise the modern utopia, in their most general sense, are provided by instrumental rationality. Instrumental rationality, furthermore, is understood as objectively valid, as opposed to values, which are understood as subjective.

Here we must point out, of course, that instrumental rationality was not invented in modernity. One could ask: could not acts that 'primitive' cultures performed, including genocide, also be understood as having been instrumentally useful for any ulterior end that such cultures might have sought? Indeed, early bureaucracies deployed by ancient civilisations, such as in Egypt and the Roman Republic, should no doubt be understood in this sense – as efficient

[93] Arguably a case could be made that this would tend to hold for all collective endeavours. In part III we see that the pattern also applies to the structure of worldviews. The very features that in any given time period and in any given set of contexts may render a worldview heuristically useful may, at later stages, when skill sets and other circumstances have changed, decrease the heuristic value of the worldview. In the particular case of modern worldviews, it is argued that they motivate the humans who inhabit them to produce the very changes that will decrease the heuristic value of those worldviews.

means to ends. Likewise, the wars conducted by such civilisations must have been conducted instrumentally in order to accomplish some more ulterior end. By implication, should not the same hold for any genocide carried out by such civilisations?

However, Bauman's main point concerns the relationship between means used and ends pursued, or rather, means and values adhered to. Whereas ancient civilisations were well aware of the practical value of applied rationality, they applied it in a different cultural context within which cultic values could be understood not only as objectively valid and absolute, but as overriding and limiting many things that were practically doable. In such contexts, human sacrifice could be performed for cultic reasons.[94] Cultic doctrine could (and still can) limit the scope within which technical means could be employed to carry out cultic ends. Cultic reasons, to be sure, are also reasons, and any practice performed for cultic reasons could then rightly be understood to be of instrumental value to any cultic aim, such as tribute to the gods. But whereas entities such as gods were considered to be of objective and supreme importance in pre-modern contexts, Bauman's point is that the only thing that retains objective validity in modernity is instrumental rationality. If at some given time period there happen to be values that to some extent limit the exercise of instrumental rationality, it is nevertheless widely understood that such values cannot *really* be absolute; they are not *sacred* in the sense that we understand pre-modern societies to be limited by a sense of the sacred. Humans, then, still use instrumental rationality to accomplish ends, but with an *awareness* that values have become subjective and adaptable.[95]

Bureaucracy, on Bauman's view, is both an organisational product of applied instrumental rationality and a vehicle through which instrumental rationality is exercised. But bureaucracies are not mere instrumental means that humans can use for their ends. Provide a bureaucracy with a goal, and it acquires a dynamic of its own. Its various instrumental limbs begin to be trained to accomplish their various partial goals, which are deemed to be instrumentally useful to the overarching goal. This echoes some of the key insights of Illich and Ellul. Provided with a goal, a bureaucracy, in the terminology previously developed, can acquire agency and autonomy that may long outlast the usefulness for humans of the accomplishment of the goal that was initially provided. In time, the bureaucracy may begin to invent new goals in order to justify its existence.

[94] For the practices of human sacrifice, see, for instance René Girard (2010).

[95] The *awareness* that values are adaptable is what matters here. The analyses in part III imply that values have always been and will always be adaptable to some extent; but this refers to a perennial human condition, one that does not necessitate any awareness of this matter on the part of humans. Awareness represents one of the traits that describe the outlook of humans inhabiting late-modern conditions. It should not be assumed, however, that such awareness must be unique to modernity. It could even be a recurring phenomenon, one that typically manifests in transitional times.

In pre-algorithmic bureaucracies, of course, new goals are invented by humans who have come, in various respects, to depend on the continued existence of a bureaucracy. Given all that we have learnt about algorithmic systems, especially how they can invent subgoals in order to accomplish an overarching goal, would this still be the case in hybrid bureaucracies – that is, in algorithm-intense hybrid multi-agent organisations?

Bauman's book is an in-depth problematisation of a modernity that, as it turns out, is two-, or, rather, many-faced. It covers a wide range of social and moral problems that can emerge in bureaucratic processes. Of special interest for the purposes of this treatise is the way in which, as Bauman convincingly argues, powerful bureaucracies can be *in a position to construct circumstances within which the instrumental rationality of human beings* – and not only human beings who are employed in bureaucracies – *become conditioned to serve ends that favour the ultimate goal of bureaucracies*. Options that in a present set of circumstances could be justly dismissed as false dilemmas could, by means of environmental and legal reconstructions, be turned into circumstantially valid dilemmas and ultimately, perhaps, into inescapable dilemmas.

This process of bureaucratic construction or reconstruction should be obvious if we consider some algorithm-sustained virtual worlds. If we wish to socialise with others through some given social media platform, we have no option but to use any of the technology-mediated means that have been made available to its users. Suppose that, upon being exposed to some content, one had the option of either hitting a 'like' button or doing nothing at all. We could say that this dilemma is circumstantially valid under the conditions that apply on a given algorithm-sustained platform. One still has the liberty to withdraw to the so-called real, conventional, or default world, where the range of possible responses to stimuli is considerably larger and, crucially, not defined in advance by a bureaucratic-algorithmic structure. Yet the conditions defined by algorithmic systems are more and more often not restricted to virtual worlds. Today, if we wish to be part of society at all, we increasingly have to use algorithm-sustained platforms in order to do our banking. We now do the work that bank-employed clerks previously did for us. This was initially marketed as an improvement, a convenience. It sometimes *is* convenient. However, as conventional banking services are being cancelled, such new conditions are turning from being optional into inescapable dilemmas: either we do the bank's work for them, or we withdraw to a marginal existence.

Today, these patterns are mirrored in how people relate to new AI-related phenomena, such as ChatGPT. As individuals, we can initially choose when and how to use ChatGPT as a tool. To say that we ought to either use it or not use it at all could be construed as a false dilemma. Surely we are able to use it rationally for our own personal ends, where and when that serves our aims? But lo and behold, as more and more people – individual and collective corporate and administrative agents – begin to use such technologies for their respective ends, the conditions change slightly. Various professional agents

will increasingly find themselves in positions where, because of fierce competition, they face either going out of business or the imperative of relying increasingly on AI. Among freelancers, those who wish to stay in business may feel compelled to re-educate themselves as so-called prompt engineers. The AI hype is in full swing: it is the future. Thus the logic or rationality of *self-preservation* can be solicited from all angles: either join forces with *it* or fade into the margins. If we join, however, we should be aware that, in time, autonomy will migrate towards the structures that control the means on which we consent to become dependent. Our participation inescapably serves the medium- and long-term goals of algorithm-powered structures, over which we who interact with them at the lower ends of technical hierarchies can exert little or no influence. If and when all autonomy has faded from pre-algorithmic social structures, circumstances will have changed: conditions that used to apply circumstantially will have become inescapable.

Over twenty years ago, long before the AI hype, Bauman saw this pattern disturbingly mirrored in the Nazi concentration camps. The Nazi overseers were not satisfied with merely administering the camps themselves, in accordance with instrumental rationality tied to their goals. The number of internees grew to enormous proportions. German staff were increasingly needed elsewhere. The bureaucratic agents in control of the camps managed to construct circumstances under which internees cooperated in their own subjugation and destruction. They created semi- or pseudo-autonomous bureaucracies run by chosen internees. Under the conditions set up by the overlords, internee-subjects were able to choose rationally between options, but all of the options that could be chosen served only the medium- and long-term interests of the oppressors. Typically, internees were induced to choose 'freely' the *lesser evil* of several options. The choice of a lesser evil in fact represented an act of voluntary cooperation with evil on behalf of the chooser. This was done by engaging the 'rationality of survival' and the 'rationality of lesser evil' of the internee-administrators:

> Those who embraced the 'save what you can' strategy had been first marked as the victims. Those who had marked them as victims created a situation in which things needed to be saved in order to exist, and thus the calculation of 'loss avoidance', 'cost of survival', 'lesser evil', was set in operation. In such a situation the rationality of the victims had become the weapon of their murderers. But then the rationality of the ruled is always the weapon of the rulers. (Bauman, 2008, p. 142)

The internee-administrators were given the option to choose between saving some internees at the expense of others: choose a hundred persons to be sacrificed, or we will choose a thousand for you. The choice to sacrifice a hundred persons could then be conceived of as a choice that saved nine hundred. This, perhaps, represents the starkest way in which rationality was recruited to serve the goals of the overseers. According to this way of reasoning, the choice gave

the internees a short-term reprieve that enabled survival; this reprieve was accomplished by an act that facilitated the short-, medium- and long-term goals of the overarching system, goals that included the extermination of the internees. To the extent that internees could be recruited to cooperate, the Nazis managed to set up largely self-administrating, or semi-/pseudo-autonomous multi-agent systems that, on their understanding, were administered by subhuman agents.

Bauman's example is perhaps unfortunate, for the purpose here is certainly not to reduce the algorithmic subjects being discussed to Nazism. However, Bauman argues that these camps represent merely one example of a very modern use of instrumental rationality. So, what would happen if we exchanged the example of the concentration camp for a bureaucratically organised hospital, one in which bureaucratic decision-making could regularly produce the effect of, say, saving any number of lives by means of deprioritising the treatment of certain pathological conditions? In bureaucratic economies of limited means, this, presumably, must occur on a regular basis. Although it is always ethically problematic, it must nevertheless be accepted as a modus operandi. In bureaucratic organisations, prioritisations and de-prioritisations will be guided by the overarching values that frame the bureaucratic system. In a more and more confusing modernity, one in which values are understood as relative and in which values can even be invented as we press towards progress, the historical camps and the hospital represent two possible products of the only thing that modernity as a whole recognises as objectively valid: instrumental rationality. The methods used in the camps, Bauman argues, permeate in many subtle ways all of the post-Holocaust structures that remain of modernity, which would include the bureaucracy of our hypothetical hospital. If the present purpose of the hospital happened to be construed on the basis of some values, such values could be expected to undergo change in the days to come. Current debates about euthanasia illustrate this process well.[96] If we cannot rely on the universal validity of the values that define the purposes of hospitals, then the organisation of hospitals could be used for all sorts of ends in the future. The same, in principle, should hold for the concept of concentration camps. It should also hold for the structures problematised by Zuboff. *More importantly, this logical structure, within which everything is instrumentalised for the benefit of an overarching goal, appears to be inherent in the technical workings of algorithmic systems.*

[96] The infamous parts played by members of the medical profession under the governance of 20th century totalitarian regimes are well-known. An ardent critic of euthanasia, Wesley Smith describes how changes in the medical professions, guided by the principles of eugenics, actually *preceded* the emergence of Nazi ideology in Germany (Smith, 2016, pp. 37–8). Historically it could perhaps be argued, then, that hospitals began to embrace purposes served by concentration camps long before there were any concentration camps. We would need, of course, to consider much more historical evidence in order to begin to make such an argument.

When ‘hard’ circumstances in the real world are not arranged so as to determine absolutely the possible venues for instrumental rationality, circumstances can increasingly be determined absolutely in virtual worlds; and to the extent that we are nudged, impelled, and even compelled to act more and more in algorithmically defined worlds, this is likely to become a more and more pressing issue. To return to the quote that opened this section, Bauman goes on to lament: ‘Under sharply asymmetrical power conditions, rationality of the ruled is, to say the least, a mixed blessing. It may work to their gain. But it may as well destroy them’ (Bauman, 2008, p. 149). Taking into account the sobering lessons of the Milgram experiments,[97] he draws the conclusion that we may now need to ‘fear the human who obeys the law more than the one who breaks it’ (Bauman, 2008, p. 151).

One could reasonably object that, if everything were instrumentalised for the benefit of goals, then this would imply that we still have goals and values in relation to which that which is instrumentalised is subordinated. We live in an age of a fairly large variety of values, when ideologies vie with one another. Yet, is there any ideology that elevates instrumental rationality to the highest position? Even surveillance capitalism, as defined by Zuboff, values money-making above all things. But this is simply how things must be by definition: means will always be used to accomplish ends. We could respond that it is precisely when values that pertain to cultural or universal ends begin to be experienced as relative and adaptable that multiple factions will begin to compete over which values ought to prevail. Perhaps the urgency with which different factions defend their values betrays the fact that, in the social arena, values have become relative. Meanwhile, all ideological positions will agree on the objective instrumental value of technical systems and bureaucratic organisations. This implies that it becomes imperative for everyone, regardless of ideology, to exert control over and develop the power of means. In regard to all other values, humans are understood as subjects that can be persuaded; in regard to rationality-enabled power structures alone, everybody is in agreement. This implies that the possession and improvement of rationality-enabled power structures are becoming an overarching end in and of itself. The ends for which we will use them at present are likely to change over time; but it is of supreme importance that we do not lose control over the means, regardless of which ends we will pursue tomorrow. If we assent to this view of things, then we should not be too surprised when modern bureaucracies, including hospitals, begin to pursue ends that we happen to think are revolting and/or primitive. Expect, rather, that the ends that are pursued are likely to be unstable and impermanent and to change over time.

[97] ‘The Milgram experiments’ refers to Stanley Milgram’s account, in *Obedience to Authority*, of a series of experiments in which subjects were put in situations where they were instructed to obey a malevolent authority (Milgram, 1975).

Many of the moral problems associated with bureaucracies bear on questions of responsibility. The compartmentalisation of tasks within larger bureaucratic frames, which is a logical organisational product of applied rationality, makes it easy for individuals to dodge moral responsibility. When bureaucracies misbehave, who is to blame? Unfortunately, as bureaucracies increasingly turn into algorithm-powered multi-agent organisations, it does not become easier to assign moral and legal responsibility. Let us consider the example of the algorithm-powered worlds of social media.

Social media platforms are often accused of being responsible for generating polarisation and other undesirable effects. Questions of where responsibility ought to lie cannot be altogether separated from questions concerning in whose or *what's* interests the effects are produced. To be sure, it may be in the interests of some people to generate polarisation. Is the solution to the present problem of polarisation as straightforward as identifying individuals who have an interest in polarising people? No. Instrumental rationality is now wielded in bureaucratic structures not only by human bureaucratic agents but also by algorithmic systems that attain various degrees of agency and autonomy. Increasingly, algorithmic systems drive organisational agentic behaviour. And we learnt earlier that algorithmic systems set to predict human preferences and behaviour not only analyse data, but also act in order to make data more predictable. In many respects, a polarised population will be easier to predict than a non-polarised population, just as it will certainly be easier to predict the actions of an agent that can only choose between a 'like' button or no action at all than it will be to predict the actions of a person not limited by any bureaucratic or algorithmic structure. Thus, barring any algorithmic instructions to the contrary, actions resulting in the polarisation of human populations are likely to be effected by algorithmic systems whenever polarisation is reckoned to be instrumentally useful to any main goal. If this is so, our assent to being part of the either/or logic of polarisation will serve the short-, medium-, and long-term interests of organisational predictive analysis; the extent to which polarisation serves those who are polarised is debatable.

If we care at all about these dynamics, then, to the extent that we consent to become more and more deeply embedded in algorithmically defined environments, *we need to pay keen attention to the choices that are set out before us*: if an option appears to serve our short-term interests, fine; but whose long-term interests does the *entire range of available options* serve? Is the exercise of our reason, within the limits set for the exercise of reason by any bureaucratic-algorithmic superstructure, to the benefit of our own persons and/or the kind of sociality that we happen to value? Will the exercise of reason in such contexts benefit inter-personal autonomy or conviviality? Or will the mere exercise of reason within the limits defined by algorithmic architectures instead strengthen the autonomy of distant bureaucratic structures that are alien to us?

The bureaucratic framing of rationality problematised by Bauman becomes even more relevant under the current and ongoing transformations of

organisations – from human-intense towards algorithm-intense organisations. Let us consider, again, the example of the Nazi camps. Here the bureaucratic agents in control of the camps constructed circumstances in which internees were incentivised to cooperate in their own subjugation and destruction. From a strictly organisational point of view, the internees of the camps become incorporated *as bureaucratic agents*, at the lowest ends of the hierarchy, into the bureaucratic organisation. A gulf, one might argue, separates this example from the economy discussed by Zuboff. But if we remove the moral dimension of the ends pursued in each case from our consideration, is there still a gulf between the two? Many of the gadgets discussed by Zuboff *only* spy on us; they do not enrol us as active agents into a bureaucratic organisation. On the contrary: in many instances they empower us as individual agents in our lives. We must concede that, at first glance, the cases seem starkly dissimilar, even from a strictly organisational point of view. The dissimilarities, however, may not be as meaningful as they seem. Various scenarios could be imagined in which, disturbingly, humans are enrolled – with or without their own knowledge –actively to carry out important functions in algorithm-intense bureaucratic systems. One of the simplest functions imaginable is precisely to provide a bureaucracy with reliable information. Only, in the case of algorithm-intense organisations, the difference – and it is a difference that makes the algorithm-intense bureaucracy potentially even more dangerous than the human-administered camps – is that humans may be enrolled as bureaucratic agents *unawares, at the lowest ends of technical hierarchies, and by algorithmic systems without human intervention*.

The principal function of such steps as 'Verify you're a human' can be to incentivise humans to perform tasks; the information derived from such task-performance can then be used to train algorithms. To the extent that such procedures are implemented, the implication is that humans are already unwittingly being recruited as agents for organisational purposes – in this case, for the purpose of training artificial agents. If in the concentration camp example humans were recruited as agents of their own destruction, we could envision an analogous situation here, in that humans, albeit unwittingly, are being recruited as agents for the purpose of removing humans from the loop in all kinds of organisational tasks. Algorithm-intense organisations are not typically interested in exterminating the human agents that they recruit. They could be; their value-informed purposes, like those of the concentration camp and the hospital, could be anything. But the internal organisational goals of algorithm-intense organisations imply making themselves, to the extent that it is possible, independent of the human agents whom they recruit.

To opt not to sign the end-user agreement of the thermostat that Zuboff exemplified means that there is a good chance that the thermostat will not function. If this is so, then it is no doubt better not to acquire the thermostat in the first place. To opt out in such situations always means to give up some perceived convenience. But the more one opts in, the more one will become

habituated to the conveniences associated with opting in, and the more difficult it will be to retreat. To opt in could in some instances be tantamount to consent to become an integral part of a bureaucratic apparatus for the sake of pleasure, security, or any other form of convenience. Ellul predicted that the logical end-point of technological society would be something like a concentration camp. Here the concept of 'concentration camp' does not imply any of the horrors associated with the Nazi death camps. Bauman, too, argues that the general principles that animated the historical concentration camps to an important extent animate the structures that remain of modernity. These principles structure administrative spaces in which internees, professional agents, or citizens are able to administer certain affairs, in which they are able to use their reason and make rational, and, of course, irrational choices, but within the limits set by the administrative apparatus, to the effect that the aggregated behaviour of the agents involved serves some overarching administrative goal.

Piecemeal acts of opting in, even as such acts may seem insignificant at the time when they are done, may have profound long-term repercussions for the viability of the worldviews of the persons opting in. The acquisition of a 'smart' thermostat is perhaps unlikely ever to undermine the heuristic viability of any worldview. But a gradual opting in on more and more offers will change the practical life-milieu of a person in profound ways. Prior affordance landscapes will be exchanged for new affordance landscapes. Such changes have previously occurred many times over. The key novelty in the process that is now under way is that we are increasingly opting in to become embedded in environments that are coming 'alive', environments in which new types of consequential agencies and agents are in the process of emerging. It should not be taken for granted that worldviews that have been heuristically viable in modernity will remain viable in such radically changed environments.

In the case of the Nazis, any rational choice that internees could make was actually harmful to the internees. But this need not be the case in the more general contexts that are under consideration. Some will no doubt argue that the technical structures problematised by Zuboff genuinely serve the interests of those who use them. We could easily concede that they are generally made to serve at least the short-term interests of users at the lower end of technical hierarchies; a good case could nevertheless be made that they are usually intended rather to serve the long-term interests of other agents, and that such long-term interests typically run counter to those that *ought* to be the long-term interests of 'users'. Certainly, with a view to structural autonomy, services offered to so-called end-users are generally designed to meet several objectives: typically, even as they are designed to offer convenience to consumers, their overarching purpose is to strengthen the autonomy of the provider. Remember that, according to Zuboff, the real customers of such services are not the consumers to whom they are marketed, but the corporations that purchase the information that is harvested from consumers.

With adequate technical means, circumstances that in principle are similar to the concentration camp but that are experienced as convenient and pleasant by 'internees' are certainly possible to arrange. Possibly, the smart home and, by extension, the smart city would fit this conceptual pattern.

Harnessing of 2) human desire. The exemplification of the second approach to harnessing human agentic powers, by means of harnessing desire, is briefer. In many respects it is structurally similar to the harnessing of instrumental rationality. In order to suggest the potential reach of this approach, a theory of desire is briefly spelt out.

René Girard's mimetic theory of desire postulates that human beings desire intensely, but that they also do not know which objects among all the things that can be desired are in fact worthy of being desired. Desire is something of a metaphysical category, and as such it is distinct from appetites such as hunger and sexual attraction; but desire is nevertheless frequently intertwined with appetites. In order to find that which is worthy of being desired, humans look to their peers. That which appears to be desired by prestigious persons in an observer's peer group must, the observer intuits, be desirable. And so we learn to desire things such as BMWs and punk rocker charisma.[98]

We have repeatedly been confronted with the executive side of algorithmic predictive analysis – that is, the acts in which algorithmic systems engage in order to make their environments more predictable. If the environment of algorithmic systems is social, if in part it is composed by human beings, then such acts could be understood in many instances as behavioural modification. Behavioural modification techniques that are implicitly based on something like the mimetic theory of desire have no doubt been used for a long time already by publicity agencies in order to make their targeted audiences desire marketed products. We are induced to desire objects because we are shown that the objects are desired by persons who are prestigious and successful with reference to specific peer groups, be they in the categories of business, sports, or music, etc.

In the algorithmically defined environments of social media, techniques based on something like the mimetic theory can be employed continually in order to reinforce previous inclinations. When such techniques are combined with the already well-established praxis of customising advertisements on the basis of previous clicks, the navigators of algorithm-powered platforms will face highly sophisticated behavioural modification complexes indeed. In the contexts of social media, peer groups – animated by the socially and deliberately conjectured prestige of influencers – can be invented and shaped in accordance with the interests of any client. Presently, as the phenomenon of AI-generated influencers is gaining traction, we must consider the likelihood of

[98] This brief summary is sufficient for the purposes pursued here. For a more detailed account the reader is referred to Girard (2011).

customised influencers being generated specifically to meet the personality traits of targeted individuals. And if AI-generated influencers proliferate, what are we to think of the ontological status of their followers? If one chooses to follow an influencer with 70 000 followers, perhaps the influencer itself and 90% of its followers will turn out to be AI-generated – a customised existential bubble designed to absorb a relatively small group of individuals who happen to share some key traits.

Manipulation of desire can play a considerable complementary role to that of the framing of rationality discussed previously. If both human rationality and desire are framed so that they contribute to accomplishing the overarching ends of a manipulating agency – such as the goals of a hybrid multi-agent organisation – then the potential predicament becomes even deeper than previously suggested. Technically, if algorithmic systems manage to frame the expression of both human reason and desire, they will have succeeded in making their social environment more predictable. This, given that predictability is of such value, implies that there is a purely technical or practical incitement to make this happen.

Some will argue that the mimetic theory of desire misses the mark. If this were indeed to be the case, then this would by no means necessarily diminish the risks being discussed. The mimetic theory was included as one among several possible theories in order to illustrate how algorithmic architectures, through knowledge of human psychological dynamics, could manipulate and frame desire. Let us now suppose that the mimetic theory is inadequate. The fact remains that humans do desire. And as we already know, many algorithmic systems of today and tomorrow will not only act, but they will also *learn*. By using methods such as unsupervised learning, who knows what the ever more efficient AIs of tomorrow might learn? Perhaps they will extensively map human desire patterns and learn how human desire is structured from scratch. Such acquired knowledge could then be applied to human populations in order to make them more predictable.[99]

Methods used by algorithmic systems to manipulate desire will no doubt include telemetry and gamification. Telemetry, Zuboff informs us, is a technique originally applied to animal tracking (Zuboff, 2019, p. 202). Telemetry employs various types of wearable or ingestible technologies that are used for data collection. Combined with ubiquitous computing telemetry can be applied extensively on human populations. Zuboff further informs us that telemetry can be combined with gamification techniques that engage participants in 'performance based contests' and 'incentive based challenges' (Zuboff, 2019,

[99] It is not argued here that it does not matter how human desire is structured. Anyone intent on manipulating human desire would indeed need to engage with that problem. The point here is that, while academics may argue about how human desire is structured, AIs may learn statistical patterns that describe the structure of human desire from scratch. If various academic theories of human desire miss the mark in some respect, algorithmic systems could end up 'knowing' the structure of human desire better than any human.

p. 215). These techniques combined enable the design of contexts within which twinned incentives work to maximise telematic scope: first, the less information users consent to share, the worse a service provided will function; the next step is to make it attractive to share personal information. This can be accomplished by recontextualising the sharing of, for instance, the details of user's physical exercises to the effect that the sharing of information becomes an incentive to exercise even more. When sharing of information becomes gamified, users are incentivised to share information not only with the provider of a service, but also with other users. Applied to the domain of physical exercises, motivations can include physical strength and physical beauty; through gamification desire for strength and beauty can be reinforced by means of exploitation of competitive dynamics. How competitive dynamics can be efficiently exploited can be derived from the mimetic theory of desire; but as has been suggested, if that theory misses the mark in some respects, a manipulator – human, artificial, or hybrid – can draw from other sources.

If all the means at the disposal of powerful algorithm-intense hybrid multi-agent organisations are employed to more and more efficient effect, there are reasons to expect that the desire of human beings, in addition to the exercises of their instrumental rationality, will be increasingly *harnessed* in the service of organisational long-term ends. For Zuboff, the duality of algorithmic machine architecture is central to the dynamics that have been described: in order to work, the machinery must be able *to know* as well as *to do*. This duality, through the many examples we have explored, redefines the agent–environment relationship of human beings vis-à-vis their environments. The high-modern cartesian human agent used to be conceived of as the supreme agent: a knower and controller of matter in time and space. But the construction of an environment that is able *to know* as well as *to do* threatens to reverse this relationship. In the words of Zuboff: 'This convergence signals the metamorphosis of the digital infrastructure from *a thing that we have to a thing that has us*' (Zuboff, 2019, p. 203).

Concluding discussion

One of Zuboff's overarching concerns relates to the likelihood that we will be able to find something that resembles a home in the algorithmically defined life-milieus of tomorrow (Zuboff, 2019, pp. 3–5). This is also one of the stated concerns of this treatise. Perhaps it is a mistake to assume that humans need contexts of conviviality, as understood by Illich, or that humans will necessarily experience alienation in technological society, as argued by Ellul. Perhaps humans will find new ways to flourish and make themselves feel *at home* in algorithmically defined worlds. What is 'home', anyway, if we posit contexts in which values and goals are becoming more and more adaptable to the evolutionary processes of the means at a society's disposal? If this is indeed the kind of context in which we live, then values pertaining to the notion of

home will most likely also be pliable and adaptable. And if ideal notions of home are adaptable, then cultural notions in general will no doubt also be adaptable.

In the introduction we were confronted with Walsh's understanding of the relationship between a people's sense of home and their worldview. If we accept his understanding, then if a people happen to be relocated to settings that are alien to the sense of home communicated by the worldview they inhabit, we can predict that they will experience something along the lines of alienation or a meaning crisis. This situation, Ellul would argue, is analogous to the situation of contemporary humans in technological society. Ellul, if he takes worldview into account at all, seems to argue from more essentialist premises. 'Natural' life conditions of human beings are transformed into 'unnatural' life conditions; hence the alienation. What if there are no 'natural' life conditions? What if the conditions that humans tend to experience as natural are simply conditions to which their worldviews happen to be attuned? If this were indeed to be the case, then resettlement in conditions experienced as alien would probably incentivise the adaptation of worldview structures.

In part III we consider the potential changes that worldviews and mythological notions can undergo in adaptation to environmental changes. If we still adhere to the understanding of Walsh, then, once worldviews have sufficiently changed, the conceptions of the homes of tomorrow could be quite different from current conventional notions. If there is such a thing as a human nature, however, there will likely be limits to the changes that all these structures – mythological narratives, worldviews, and homes – can undergo.

The dynamics described by Illich, whereby the autonomy of human agents in local social contexts is undermined, make the technocentric society described by Ellul appear almost a necessity. The perceived necessity of such structures, no doubt, lays the foundation for the ideology- or worldview-like structure that Ellul calls *le phénomène technique*, which incentivises denizens to organise a technological society that bends over backwards in order to create conditions that enable technique to flourish and to evolve. It might be possible to counter these dynamics by, for instance, producing convivial tools and cultivating various kinds of practice that are conducive to convivial ways of life. In our times, industrial and technical structures continue to bear out their logic. The contemporary individual is one who forms the most structurally consequential, albeit not necessarily the most emotionally charged, relationships not with family or village members or personal friends but with techno-industrial institutions. In our own life-narratives, important family members and personal friends may occupy places of honour. But who or what really sustains our way of life? Who or what enables our income? Who or what enables the modes of our transportation? Who or what enables entertainment, travel, and medical treatments? The arena of the locally uprooted and cosmopolitical individual is *the world*, not the locally situated extended family and village. To the extent that parts continue to sustain and cohere with the whole,

we enter the era of a 'liquid modernity', a modernity that is not without its pleasures and venues for human flourishing.[100]

But the era of liquid modernity, as Bauman characterises it, may already be drawing to its end. From the angle developed in this treatise, the function of 'liquidity' will depend on the existence of rewarding agency-venues for *human* agents in the organisational structures of modernity. The evolving bio-technical hybridity of current multi-agent structures threatens to eliminate the need for human beings in human societies. Meanwhile, signs of social breakdown all over a world that is technologically more advanced than any world before it suggest some deeper dysfunction. They suggest, perhaps, that the reasoning of Wiener and Ellul is on target: the addition of more and more complex layers of technology and automation will put enormous strains on human cognition; far from creating a world of leisure, it will set up contexts that require super-human cognitive abilities, and, barring such abilities, contexts that will yield dysfunction and unpredictability. All of this bodes ill for the expectations of a harmonious and leisured future.

None of this eliminates the possibility of future human–AI partnerships as envisaged by Kissinger *et al*. The dynamics discussed in this section should make us question just what such partnerships might mean for the majority of people. Bauman's argument, that the principles that animated the historical concentration camps are similar to the principles that animate the structures that remain of modernity, could serve as a warning, but also as a lead-up to an important question: Would we consent to be confined to a concentration camp if the life-experience in it were more pleasant than the life-experience in the real world? Today, as tensions rise all over the world, and as we are drowned in apocalyptic narratives about climate change, pandemic menaces, foreign enemies, popular uprisings, and, indeed, AI apocalypse, while algorithmic technologies are being continually improved, such conditions could no doubt be constructed, at least to some extent. It is the Matrix-problem from the point of view of a pre-Matrix society: Should we construct and submit to the Matrix?

The insights of Bauman and Ellul suggest that a well-informed and decisive answer to this question would be difficult to find for humans who are embedded in the lore of modernity. Whereas many strands of Enlightenment thinking originally promoted the universality of values, late modernity definitely retains the objective validity of instrumental reason. To be sure, in a world where constant change is understood to be the normal condition, numerous proponents of modernity still argue for the universality of their respective values. Do such arguments reflect some genuine belief in the universality of values? Or do they reflect, rather, the need for political strategy in a context where change is understood to be the normal condition? The latter, it could be

[100] Bauman applied the concept 'liquid' to a peculiar light and software-based modernity (Bauman, 2000), to love (Bauman, 2003), and to the age of uncertainty (Bauman, 2007).

argued, is more likely. Meanwhile, all proponents of modernity who disagree with one another on values agree implicitly on the universal validity of instrumental reason. The earliest shoots of this late modernity plant emerged, it appears, even at the apogee of the Enlightenment; for what is the categorical imperative of Immanuel Kant if not a reduction of universal values to the exercise of something like instrumental reason?

The role that Zuboff attributes to surveillance capitalist ideology, it has been argued, is exaggerated. It is not that ideology has no function. It is, rather, that, at least in these instances of continual environmental reconstructions, it is not possible to make any neat separation between ideology and environments. The chief enabler of instrumentarian power, however, is not surveillance capitalism, nor even capitalism, but the modern and late-modern elevation of instrumental rationality as the supreme value.

The ability to control environments represents a timeless temptation for humans. In the contexts that have been discussed, if or when new affordances are revealed in the process of technological development so that interested parties realise that behavioural data from humans can both improve product development and be used to exert control over humans, something like instrumentarian power is already a given. To use Ellul's terminology: the choice is automated. That which, on Zuboff's analysis, may be more directly tied to a surveillance capitalist ideology concerns the use of behavioural surplus data and behavioural futures markets. It is possible that, in a technological society with a slightly different bent, behavioural futures markets might not be created. But then, if that society were really a technological society as envisaged here – one that primarily organises in order to create the best conditions under which technology can flourish and evolve – there is no reason why it should not invent behavioural futures markets as the technostructures evolve, or, for that matter, why it should not invent any other new goal that happens to fit well with the current state of technology. The technological society of late modernity is one in which only instrumental rationality is recognised as objectively valid. By extension, its fruit – technique – is recognised as ontologically real. Values, on the other hand, can be invented to fit evolving circumstances.

Conclusions Part II

Reconstructed affordance landscapes in algorithm-intense environments will likely have radical implications for human agency and autonomy:

1) Algorithmic systems or artificial multi-agent systems will present various sets of affordances both to the human consumers to whom they are made available and to the stakeholders who exercise control over them.

2) For the first time, humans and human populations will also present affordances to technological systems that embody various degrees of agency and autonomy.
3) Algorithmic systems in the context of hybrid multi-agent organisations will present sets of affordances to the human agents at the upper levels of technical hierarchies that differ from the affordances presented to consumers at the lower ends of technical hierarchies.
4) Human beings and human populations, by the mediation of algorithmic systems in hybrid multi-agent organisations, will present new affordances to the human agents at the upper levels of technical hierarchies.
5) Given that we accept some or all of the human vulnerabilities discussed in part II (propensity to seek the satisfaction of short-term wants at the cost of long-term interests, propensity to enter frameworks within which our own reasoning may be deployed against our own interests, lack of control over our own desires), algorithmic systems deployed in the context of hybrid multi-agent organisations represent new and powerful means for exploiting human vulnerabilities.
6) By incentivising humans to act within algorithmically defined environments, the scope of human options could be framed so that human actions benefit the long-term interests of the operators of hybrid multi-agent organisations.
7) Algorithmic systems that learn the patterns of desire of human beings could be used to make human desire serve the overarching goals of algorithmic systems and/or the long-term interests of the owners of algorithmic systems.
8) Algorithm-intense hybrid multi-agent organisations might increasingly be able to absorb humans *qua* bureaucratic agents or assets into their organisational structures, to the effect that humans, unaware of the organisational function that they fill, will 'work' for the benefit of such organisations, potentially to their own disadvantage.
9) For all of these reasons, increasingly algorithm-intense environments imply the intensification of Wiener's predicament. If we wish to preserve some scope for spontaneity and human autonomy, then we will need to exert our brains more, not less, and we will need to pay careful attention even to the most mundane aspects of our life-environments, even to the potential various interests that could be served by a thermostat. If we fail to rise to the challenge, then we will face a potential overwhelming of human capabilities – an overwhelming executed by environments that humans have engineered.
10) It is difficult to see that we are moving towards anything that could be characterised as conviviality in the way that Illich understands it. But we are not simply being absorbed into old and tested industrial structures. Rather, we are being introduced into new contexts in which

humans can form ostensibly meaningful relationships with machines. But this is also a context in which the lines between categories become blurred and in which meaningfulness itself becomes a question of appearance. If we 1) use a so-called smart technology or 2) form a relationship with a machine, it will be difficult for us to grasp the underlying meanings of the processes in which we are engaged. Assuming that our own cognitive capacities do not become radically enhanced, we will be entering a more confused condition.

11) Together, these scenarios portend widening power asymmetries between humans at the lower and higher ends of technical hierarchies and, potentially, a loss of meaningful agency and autonomy at the lower ends of technical hierarchies.

Part III: New tensions – modern worldviews, non-modern life conditions

Thus far we have considered how algorithm-intense environments are likely to affect the expression of human agency and autonomy. Here the task is to consider how algorithm-intense environments will affect modern worldviews. In order to do so, we first consider, in a more general sense, the relationships between worldviews, world-making, and environments. A generic modern type of worldview is proposed. This type of worldview, it is argued, is animated by a generic modern mythological narrative. The modern type of world, it is also argued, is compatible with a specific philosophical anthropology. Finally, we examine how this modern philosophical anthropology might begin to adapt as humans become embedded in algorithm-intense environments, or, if it is maintained in its present state, what other adaptive measures are likely to develop. The larger part of the introduction that follows is devoted to a consideration and elucidation of the phenomenon of modernity.

Introduction to worldviews, mythological narratives, and modernity

A worldview is understood here to consist of sets of assumptions that frame how persons who inhabit it view and understand themselves and the world.

In considering how worldviews might be affected in algorithm-intense environments, we run into the problem of deciding which worldviews we should consider. Modern societies are characterised by the co-existence of a plurality of worldviews. We could understand the 20th century as having been defined by several dominating types of worldview: capitalist, liberal, communist, fascist, secular humanist, and a number of religious worldviews.[101] The 21st century then may see the eclipse of some previously dominating worldview-types, and the emergence of some novel types, such as AI-centric and transhumanist worldviews. We cannot consider here how AI-intense environments will affect every single worldview that prevails in contemporary societies. Instead, 'modernity' is treated as an inclusive and epoch-defining category. The working hypothesis is that worldviews that historically have generated conditions that can be understood as 'modern' will have some key patterns in common.

[101] If these paradigmatic examples are cited, it must be understood that an actually instantiated worldview, as opposed to possible theoretical reconstructions, will not be a purely political or purely religious or purely metaphysical worldview.

In order to explain further how these common patterns animate modern worldviews, the influence of an epoch-defining mythological narrative is inferred. The core tenets of modern worldviews, it is argued, gain their purpose and meaning from being in dialogue with this mythological structure. One of these core tenets is the understanding of the nature of the human subject and its relationship with its environment or world, namely the modern philosophical anthropology. This understanding in particular, it is argued, will be directly challenged as humans become more deeply embedded in AI-intense environments.

Some of the worldviews – notably religious ones – that could be understood to have been shaped by the mythology of modernity may trick us into thinking that they must be pre-modern. If we speak of a modern Christian worldview, we must differentiate between worldview and religion, with the understanding that the Christian religion was originally shaped in contexts defined by pre-modern mythological narratives. Such narratives, although no longer epoch-defining, have not altogether vanished. In contemporary contexts, it is therefore possible to inhabit both modern and pre-modern Christian worldviews. Some core traits that differentiate the two are discussed.

The content of a worldview will, by default, consist largely of beliefs and understandings that have been learnt in a way that gives them the status of pre-reflective and taken-for-granted assumptions. An inhabitant of a worldview can become aware that many of its structures are arbitrary or out of touch with the world that is actually inhabited. An inhabitant of a worldview can, then, set out actively to reform the worldview. In other instances, a worldview can simply be abandoned, whereupon a worldview conversion may ensue. If a person also becomes aware of being under the influence of a nebulous mythology, then the inhabitant can opt either to agree with or to oppose currents suggested by the mythology. No *single* person can reform a mythology. If indeed there are such things as epoch-defining mythologies that influence the evolution of worldviews, then they must surely be very thick structures that will tend to endure for centuries. When they do undergo change, the change is likely to result from complex structural changes in the world and/or environments in which the mythology prevails. For instance, structural worldly changes can render a narrative maladapted for heuristic use. An epoch-defining mythology, then, can wind up being maladapted to the world that, indirectly, it has produced.

Several of the thinkers cited earlier could be understood as opposing in some respects some of the currents that are suggested by the mythology of modernity. None of them could be said to have reformed the mythology. Even to this day, it is assumed, the mythology stubbornly exercises its epoch-defining influence. However, changes in human life-milieus may ultimately render core tenets of this mythology increasingly out of touch with reality. This, in turn, would render the mythology increasingly obsolete and therefore also invite mythological adaptation.

The mythology that is proposed here is not accessible anywhere as an authorised version. The theoretical reconstruction of the mythology will no doubt seem very familiar, as it is a version of one of the most common and best-known stories ever told. For although the narrative is not straightforwardly accessible as an authorised version, its many tokens are ever present in our midst. The narrative structure is indeed so prevalent that we continually fail to perceive it as *mythological narrative* and instead interpret it as the *progressive or teleological history of humankind*.

Since modernity arose in the western world, the reconstruction will take the form of a western-informed narrative. Facets of the modern world, notably western ways of approaching technology and science, have since spread all over the globe. The question of the extent to which other scientifically and technologically advanced societies, such as Japan and China, can be supposed to have been shaped by such mythological structures must, for now, be left open. Some fruits of modernity, notably the western approach to science and technology, have now become integrated all over the globe. Other notions such as 'human rights', 'liberal democracy', and 'communism' are also fruits of modernity: these are considered ideas that, to varying degrees, ought to be integrated or opposed all over the world. If we limit our interest to advanced science and technology, then we may see little difference between the western world and the Far East. The aspects that are of special interest here, namely, a rudimentary philosophical anthropology that supposedly inheres in modern worldviews, might be more parochial in scope. If this turned out to be the case, then that which follows would perhaps be of higher relevance to western readers, but not without interest to readers from other cultures that, in their respective ways, participate in the modern world. Time may inform us further on this matter.[102]

First in part III, some theoretical concepts that will be useful in understanding worldview structures are discussed. Second, characteristics that render a worldview modern are inferred. Third, a narrative structure for the mythology of modernity is proposed. The narrative structure is based on general and commonly shared understandings of the history of humankind. We consider how some prominent thinkers – a pioneering anthropologist and two philosophers – have engaged with this narrative structure. Finally, we consider how the modern understanding of the human being is challenged in AI-intense environments, and some of the adaptive measures that might develop as a result of that challenge.

[102] What are we to make, for instance, of the influence of Shintoism over contemporary Japanese society? Since Shintoism is generally understood to be an animistic belief system, any answers to this question could be of great relevance to the analyses conducted in this part of the treatise.

The remainder of this introduction is devoted to the clarification of a concept that is at the centre of all these discussions: modernity.

'Modernity' is a concept used to describe a sort of society, social order, or mentality. There are many different ways to define modernity. One commonsensical and inclusive way to set up limits for modernity could be to say that the concept denotes the type of order that succeeded more traditional and parochial pre-modern ways of life. This would not tell us much. We would, on this account, no doubt be living in modernity, and the populations of Europe would, to varying degrees, have been living in modernity for the last 100 to 500 years. Within this time-span, societies have undergone such rapid and profound changes that we who now live in technologically advanced societies barely recognise that we live in an order that is similar to that of our grandparents, let alone to any order that may have prevailed 100 to 500 years ago.

We have already become partly familiar with Bauman's conception of modernity. The discussions concerning Bauman in part II suggest a considerably narrower understanding of modernity. Elsewhere Bauman has theorised a modernity that he labels 'liquid modernity' in contradistinction to a more 'solid modernity' (Bauman, 2000, 2007). Liquid modernity is characterised by increasingly fast and disruptive changes and decreasing institutional stability, which in turn makes it difficult for humans to have long-term projects, and instead prompts them to be increasingly flexible and adaptable. Through this concept, Bauman seeks to describe modernity as it has developed in late 20th century liberal societies. The use of a qualifier to 'modernity' suggests that there are several different modernities, and/or that modernity might follow a trajectory of change over time. On the latter view, we who live at present must be living in late modernity or, perhaps, in *a* late modernity.

Thinkers who, like Bauman, write about 'late modernity' mean to stress that there is a continuity between something like a 'high modernity' and a 'late modernity', but that the conditions that define the latter are markedly different from those that define the former. In contrast, thinkers who are associated with the term 'postmodernity', such as Jean Baudrillard (1981) and Jean-Francois Lyotard (2005), stress rupture: according to them, we no longer live in modernity. Typically, such thinkers describe a condition in which the belief in progress and in grand narratives has collapsed. Institutions tied to such a belief then also lose their status. Left is the individual who, in a sense, becomes the arbiter of his or her own reality. There is no macro-agreement. Each individual has his or her own story, or micro-narrative. Other thinkers still, such as Bruno Latour, argue that we have in fact never even been modern (Latour, 2010).

We have now touched the edges of the concept of modernity, evoking the possibility of a late modernity, a postmodernity, and the notion that we have never been modern to begin with. How, then, should we understand modernity as a core concept? Innumerable thinkers have written copiously about the subject. Here, in this introduction, we primarily consider Romano Guardini's

understanding of modernity. Later we turn to Charles Taylor for key insights that pertain to the philosophical anthropology of modernity.

Guardini's representation of modernity, and especially his understanding of the role of culture in modernity, will enable us to grasp the tensions inherent in the phenomena labelled as modernity, notably between original modern ideals and the life conditions that have issued from societies that have historically adhered to such ideals.[103] This understanding will further strengthen one of the core themes argued in this treatise, namely, the theme of continuity in transformation. However, the theme of continuity may not fit well into Guardini's own understanding of the matter.

Guardini, in the 1950s, argued that the western world was undergoing changes that transferred it out of modernity into a new condition. According to Guardini, modernity is characterised by special conceptions of nature, the human being (he often uses the gendered word 'man' for 'human'), and culture. Modern Europe was structured around the following ideals:

> a nature subsisting in itself; an autonomous personality of the human subject; a culture self-created out of norms intrinsic to its own essence. The European mind believed further that the constant creation and perfection of this "culture" constituted the final goal of history. (Guardini, 2019, p. 68)

By the mid-20th century, Guardini argues, all of these conceptions are unravelling. Modernity is characterised by a reverence for 'Nature' as a source of harmony. Nature, however, is increasingly understood as ambiguous. On the one hand, we become trained to see nature as

> insensate order, as a cold body of facts, as a mere "given," as an object of utility, as raw material to be hammered into useful shape […] Technological man will remold the world; he sees his task as Promethean and its stakes as being and non-being. (Guardini, 2019, p. 74)

On the other hand, humans feel increasingly responsible for their universe. This implies that they must 'take care of being' (Guardini, 2019, p. 72).

If modernity professed individualism and a faith in the autonomous self-made genius, the new human type, according to Guardini, is absorbed by technology and rational abstractions (Guardini, 2019, p. 76). The cognitive requirements of the complex new order of the world are becoming exceedingly high.

> The work of dominating the world calls for a union of skills and a unity of achievement that can only grow from quite a different attitude. This new attitude is revealed by the evident fact that the coming man renounces an

[103] Guardini's representation of modernity may also elucidate a possible reason for the sense of *loss* that is implicit in the analyses of Illich and Ellul: a sense of loss of a natural, harmonious, and autonomous view of self in relation to the world.

> idiosyncratic life for a communal form, that he surrenders individual initiative for a given order of things. (Guardini, 2019, p. 84)

This 'given order of things' is not 'Nature' as experienced by modern humans; for the new human no longer experiences nature as Nature, but as something mediated by abstractions, mathematics and technology. The new nature 'is beyond our common experience. If experienced by a few here and there, it is done in an enigmatic way through an order of things to which man cannot speak' (Guardini, 2019, p. 90). Guardini is uncertain about what this could imply. It may, he admits, imply that the boundaries for human experience are being expanded.

Changes in views of nature, human beings, and culture are all interdependent. It was nonetheless stated earlier that the conception of culture is of special interest for the present inquiry. On Guardini's understanding of modernity, culture, from the point of view of modernity, is thought of as an objective good. The presupposition is that humans were originally at the mercy of a merciless environment. Culture, in modernity, represents power over nature.[104] One of the key goals pursued by modernity is constantly to improve and refine culture, and thus by implication to increase power over nature. It is to be understood that the culture of the civilised modern human includes all fields of autonomous[105] activity, from arts and letters to technology and science. This refinement, it was believed, would result in a corresponding and continual improvement of human life conditions. The tragic developments of the first half of the 20th century revealed the naïvety of this belief. Culture is not an unambiguous good. Culture denotes ambiguous mediating structures that can be used for both good and evil. The goal of the constant improvement of culture, therefore, ushers modernity into crisis.

The crisis hearkens back to modernity's own conception of the instrumental value of culture. Without *any* culture, human beings would indeed be very vulnerable in nature. The moderns were thus right in their valuation of culture in relation to nature. They were wrong in their assumption that cultural refinement beyond all limits was unproblematic. For if, in an original condition,

[104] In pre-modern contexts, culture may also have been understood as means that, *to some degree*, represent power over 'natural elements'; but here culture's *mediating qualities* are likely to be more strongly emphasised, for nature represents something that can be neither fully understood nor controlled. For pre-modern subjects, a healthy culture would have been one that framed its inhabitants in a way that they did not overstep their limits and so upset an established balance; modern culture constantly urges us to push the limits and move beyond the limitations of previous generations. If pre-modern culture aims at stability and perseverance, modern culture aims at progress and the exploration of new frontiers.

[105] Yet another trait of modernity, as Guardini understands it, is that all fields of pursuit become 'autonomous' spheres of activity. In pre-modern Europe, everything was bound together and structured into an ordered hierarchy by revelation. In modernity, science, arts, and letters become autonomous spheres of activity, not bounded or limited by any mysterious pre-modern hierarchy. This understanding puts Ellul's conception of *technique as autonomous* in a broader context.

danger primarily originated in the natural environment, we have been entering a condition in which danger primarily originates within the very means used to ward that danger off – that is, within culture. What is it specifically in culture that poses danger? The answer is: power.

> To exercise power means, to a degree at least, that one has mastered "the givens." Power over "the given" means that man has succeeded in checking those existential forces which oppose his life, that he has bent them to his will. Today the sceptre of power is wielded by the hand of man. He has extensively mastered the immediate forces of nature, but he has not mastered the mediate forces because he has not yet brought under control his own native powers. Man today holds power over things, but we can assert confidently that he does not yet have power over his own power. (Guardini, 2019, p. 109)

The world of tomorrow, according to Guardini, will be shaped by the problematic relationship between a powerful but unhinged culture and a nature out of joint.[106]

The structures analysed by Zuboff, which we could also understand now in terms of multi-agent structures, must, if we adopt Guardini's lenses, be understood as integral parts of this new culture. In such multi-agent systems and organisations we may find human agents who fulfil varieties of roles, many of which will be outright alien to the human self-understanding that dominated during the times of high modernity. Nature, meanwhile, represents a constant apocalyptic danger, as it is widely believed that the technostructures used by humankind – in fact, the integral machinery of their culture – continually wound and derail the processes of nature, thus provoking a nature out of joint, a nature that is vengeful. This picture of our contemporary times seems distinctly non-modern if we compare it with modernity as Guardini describes it.

But has the western world really or thoroughly left modernity behind? Although contemporary conditions could be justly characterised as markedly different from the condition that prevailed during modernity as described by Guardini, some defining modern characteristics continue to manifest in our own times. Forms of organisation that arose during modernity – modern

[106] Viable venues for power leverage can, in a sense, be expected to mirror modernity's *autonomous spheres of activity* (cf. previous footnote). To the extent that human society is knit together in some form of unity of purpose, as, for instance, to render service to God, the higher purpose, in important respects, will inform the human use of power, just as such a higher purpose once informed the presently autonomous spheres of activities of science, the arts, and letters. When human society is 'liberated' from such common unity of purpose, it should be expected that power will be leveraged in the service of varieties of purpose that were once understood to be lower, subordinated to, and regulated by a higher purpose, but that, the hierarchy broken, can henceforth be understood as justified in and of themselves. Liberated from celestial and other non-worldly authorities, humans can increasingly be expected to feel free to use power for ends that in previous contexts would have been understood as objectively good or evil. Indeed, increasingly, humans appear to be in positions where they consider themselves free to define good and evil without submitting to any authority higher than their own.

bureaucracies, factories, automation machinery – continue to operate in full swing. Many still express a belief in progress. Observing the messaging of corporations and governmental organisations, one could easily get the impression that *progress by means of science and technology* still qualifies as a core cultural dogma. The popularity of thinkers such as Steven Pinker could also be taken as evidence that Guardini goes too far in his assessment.

Pinker's *Enlightenment Now* is written as a defence of the ideals of reason, science, humanism, and progress. To be sure, the Enlightenment tradition defended by Pinker cannot be equated with modernity as a whole. Still, it could be understood to represent one facet of modernity. It shares with the larger category of modernity the fundamental aim of escape from error and obscurity by means of applied reason. Pinker points out that 'the Enlightenment endorsement of reason' should not be confused 'with the implausible claim that humans are perfectly rational agents' (Pinker, 2018, p. 8). It is precisely because we are habitually animated by 'irrational passions and foibles' that we need to apply reason methodically in order to overcome them.

The Enlightenment tradition is the context in which the methodical application of reason is understood to have been cultivated. The practices and habits of reasoning that have come in important respects to define the modern age were, supposedly, initially *cultivated* by Enlightenment thinkers. We could also think of this as the *culture* of the Enlightenment. Here it is possible to interpret a thinker such as Pinker as someone who, with respect to 'culture', applies the view specified by Guardini vis-à-vis culture in modernity onto a narrowed-down culture of the Enlightenment. With respect to the more general culture of modernity, Pinker would no doubt agree with the successor view – that is, that it is ambiguous, a possible vehicle for both good and evil. Enlightenment culture, on the other hand, provides us with frameworks by means of which we can hold our beliefs accountable to objective standards. By trial and error, through Enlightenment culture, we can learn the errors of our ways and improve our knowledge. Mistakes will still be made, but in time and piece by piece, the methodical exercise of Enlightenment virtues will enable us to replace lore with more and more reliable method, superstition and error with more and more knowledge. Even our self-knowledge will become increasingly scientifically informed (Pinker, 2018, p. 10). Science alone, Pinker recognises, cannot guarantee progress. It needs to be complemented by another heritage of the Enlightenment. Humanism, Pinker argues, is needed more than ever in order to instantiate an adequate ethical framework and to encourage humans to accept cosmopolitanism and 'our citizenship in the world' (Pinker, 2018, p. 10).

The bulk of Pinker's book consists of presentations of evidence that he argues should persuade us that the Enlightenment tradition is indeed the vehicle that has permitted us, and still permits us, to improve our chances for human flourishing. That which should concern us here is the extent to which facets of modernity still thrive today. The popularity of thinkers such as Pinker

indicates that some facets of modernity still animate the intellectual and popular imagination.

Although Pinker cannot be said to express modernity in a way that perfectly corresponds to Guardini's characterisation of modernity, he continues to be animated by faith in progress. In our day, few if any can have the view portrayed by Guardini of culture as a whole. But Pinker singles out a narrow chunk of culture, the culture of the Enlightenment, which continues to carry the function that culture as a whole allegedly carried during high modernity. In some other respects, Pinker no doubt fits the patterns that Guardini argues characterise that which succeeds modernity – that is, an ambiguous view of nature and of culture as a whole. As for the narrow chunk of culture that is represented by Enlightenment thinking, it could still be thought of as an unambiguous good, *the* vehicle that permits humans to escape from superstition and increases opportunities for human flourishing. Thus, it becomes imperative that we continue to safeguard and perfect this culture.

Even if a thinker such as Pinker is hardly representative of the larger population, we can conclude that some of the fruits of modernity continue to mature in our times, even as we are on the verge of something new. We are standing, perhaps, with one foot rooted in modernity, while the other is feeling for solid ground in the obscure beyond. We cannot yet decisively leave modernity, for there is not yet anything solid to hold on to beyond its increasingly outdated premises.

It is possible, as we have seen, to define modernity in different ways. Rather than giving a precise definition of modernity or of a modern way of life, it is posited that life conditions for the last few hundred years, first in the West and then, to various extents, all over the world, have been indirectly generated by a type of worldview that is different *in type* from all worldviews that can be conceived of as 'pre-modern'. During this long stretch of time, worldviews, like life conditions, have undergone much change. A plurality of worldviews manifests, many of which are in significant conflict with one another. It nevertheless makes sense to label such worldviews as 'modern' to the extent that they are extensions of a much simpler and vaguer narrative – one that is characterised here as the mythology of modernity. Something like this modern narrative, it is posited, is still motivating populations to support or go along with the reconstructive endeavours of contemporary societies.

The conceptualisations currently under development imply that it is possible to argue that – according to some specific criteria – people who inhabit modern worldviews are not modern; one could even argue that no one has ever been modern. If it is posited that we are influenced by a mythology of modernity, all that is meant is that we are induced to think and behave in certain basic ways with regard to human self-understanding and ways to achieve a good society; the end-result of this thinking and behaving might be very different from that which could be expected on the basis of specific modern

ideals, and it would therefore be possible to maintain that, historically or sociologically, and according to some specific criteria, we no longer are modern, or perhaps never have been.

In general, wrong or unintended results can be explained by an inadequacy of means, given factual circumstances, with respect to desired or intended objectives. In the ecology described here, inadequacy can be located on many different levels, including those of worldviews, knowledge, and know-how. Crucially, in the algorithm-intense environments under construction, inadequacy can be distributed all over hybrid multi-agent organisations and locked within pure algorithmic systems. Given that we can expect that some inadequacy will always be distributed over various levels in agent–means relations relative to the goals being pursued, we should always expect the actual effects to be somewhat different from the intended effects in processes that involve many unknown factors and dynamics.

It will be maintained, nevertheless, that *modern worldviews* have tended to generate conditions that we could plausibly describe as *modern conditions*. However, modern conditions can be quite diverse in character. A 'modern condition' is understood here as any condition within which an inhabitant of a modern worldview can continue to navigate the world in a way that, at least theoretically, can be purposeful and meaningful. If modern worldviews were to cease to generate modern conditions, then modernity – on the level of assumptions, ideas, and general attitudes – would find itself radically challenged. Many of us are still living in conditions that *can* be, but that do not necessarily *need* to be, experienced as modern. Increasingly, they are borderline conditions.

Modern worldviews

What type of worldview fits with and produces modern conditions and/or technological societies?

In order to answer this question, a fairly conventional general understanding of the concept of worldview is presented. This understanding centres on 'big questions'. The approach is also sufficiently interactionally dynamic in that it has the potential to integrate the entire systemic ecology – with its complex affordances – discussed in parts I and II.

In a concise and comprehensive paper, Ann Taves, Egil Asprem and Elliot Ihm build on an understanding of worldview as 'a complex set of representations' related to the following big questions:

> (1) ontology (what exists, what is real), (2) epistemology (how do we know what is true), (3) axiology (what is the good that we should strive for), (4) praxeology (what actions should we take), and (5) cosmology (where do we come from and where are we going) […] (Taves, Asprem and Ihm, 2018, p. 208)

The writers posit that worldviews integrate in complex ways with 'ways of life' and 'world-making', but that worldviews belong uniquely to a 'species with the capacity to create what anthropologist Maurice Bloch (2008) characterizes as "transcendental" social worlds, that is, worlds that go beyond face-to-face transactions' (Taves *et al*., 2018, p. 208). Non-human animals inhabit mere transactional social worlds, in which, again according to Bloch, 'animals assume roles and produce groups based on "a process of continual manipulation, assertions and defeats"' (Taves *et al*., 2018, p. 210). Moreover, humans are not necessarily the only creatures who generate answers to the big questions. Even simple organisms, the writers argue, can enact implicit answers by means of 'world-and-self modelling capacities': humans, however, are unique in their capacity to articulate, reflect on, and approach the big questions *as* questions.

Here we also retain the understanding of Brian Walsh discussed in the general introduction. Drawing on previous sections, we can recognise that ways of life can be understood as embodied praxeological answers to questions. Architecture, in turn, can be understood as structural forms *for* ways of life that *imply* answers to questions. This applies to conventional architecture and, also, to algorithmic architecture. Finally, embodied praxeological answers and formative (architectural) implicit answers can be more or less compatible with worldview-patterns that may or may not be articulated by the human agents who inhabit architectural spaces or who enact implicit answers through praxis. If things are understood in this way, then it should be possible to see how worldviews, imbricated in ways of life and environmental structures, can be more or less in agreement with or more or less at odds with changes that occur in ways of life and in environments.

Taves *et al*. bring to the discussion another understanding that the reader will recognise from the explanation of how learning machines function and from the introduction of the concept of affordances. The human agent is understood to act in a 'lifeworld'. Like learning machines, higher biological organisms are able, more or less accurately, to learn how to predict their environments. A 'world-model' is understood as something more encompassing (or, if one prefers, more basic) than a worldview: it is an embodied and enacted set of predictive inferences of the agent's environment. If worldviews are concerned with the articulation of explicit questions and answers that generate a normative dimension, then world-models are concerned with the way things can be expected to work. The generation of a world-model 'co-constitutes a self-model, a set of expectations about […] bodily extension, needs and so on' (Taves *et al*., 2018, p. 208). World-models generate plausibility landscapes within which ways of life can be enacted and worlds can be constructed, and within which worldviews can make sense of the whole. A world-model can be understood as the cognitive interface by means of, or through which, an agent is able to comprehend and exploit environmental affordances. If simple organisms, through their world- and self-models, and given their pursuits, can

be understood to enact implicit answers to big questions, then this implies that algorithmic systems, through the world- and self-models that they may construct, and given the goals that they are set to pursue, may also enact implicit answers to big questions.

The authors offer three useful definitions that pertain to the agent–affordance relationship:

> An affordance is a possibility for action provided to an organism by things and creatures in its environmental niche, given the organism's particular sensorimotor, perceptual, and cognitive abilities (cf. Ramstead, Veissière, & Kirmayer, 2016, p. 3). An environmental *niche* can be defined as the totality of affordances available to a particular population in a particular environment, a "landscape of affordances" (Ramstead *et al.*, 2016). An individual organism's world emerges from the local and situated attempt to leverage immediately available ensembles of affordances in the niche. (Taves *et al*., 2018, p. 208)

The authors also provide a useful distinction between natural and conventional affordances: 'In contrast to natural affordances, which generate action based solely on environment–organism relationships, conventional affordances depend on cultural schemas that provide shared expectations relative to specific things, persons, places, or events' (Taves *et al*., 2018, p. 209). Ways of life can then be understood as 'the organism's habitual patterns of interaction with affordances in its world. […] Since an organism's predictive models of itself and its environment are generated from interacting with its niche, world-and-self models are *co-constructed* in ways of life' (Taves *et al*., 2018, p. 209). The authors suggest that all organisms embody answers to at least some of the big questions 'in the shape of affordance-based world-and-self models'; such embodied answers notably relate to that which exists, to that which an agent ought to strive for, and to how an agent should proceed in order to realise its goals:

> Viewed from an evolutionary perspective, everything revolves around the *praxeology question*: How should the organism act in any given situation? The organism's action is determined from its best prediction of *what is* (ontology) in accord with *the affordance-based goals and values embodied in its self-model* (axiology). (Taves *et al*., 2018, p. 209)

This description, it seems, could also be judiciously applied to many artificial systems. When the concept of multi-agent systems was introduced earlier, a simple model for how intentional systems, such as human and artificial agents, can interact with environments was also suggested. Intentionality presupposes the ability to have inner representations of exterior states of affairs. Inner representations, it was claimed, can more or less accurately reflect exterior states, but they can also suggest, crucially, how exterior states *ought* to be arranged; in the case of non-moral agents, such as animal and artificial agents, the

'ought' represents states that are deemed to be instrumentally useful to the accomplishment of *any* goal.

The specific concept of worldview, on the other hand, is limited 'to those social animals – primarily humans – who have the cognitive ability to develop and use conventional as well as natural affordances (Ramstead *et al.*, 2016)' (Taves *et al*., 2018, p. 209). To the extent that humans have certain inferential capacities, they are capable of turning enacted and embodied implicit answers into explicit semantic content, to ask the big questions and to consider alternative answers. We can speculate about untried ways to do things. We can travel mentally back in time and revisit past 'transcendental social worlds' in order to consider how things have been understood and done in previous times. In contrast to transactional social worlds, transcendental social worlds include forms of 'essentialized roles and groups, that is, roles and communities that *exist in the imagination*, independently of the individuals that comprise them. […] Worldviews [in the understanding of the authors] emerge together with this "transcendental social" and are an integral part of its organization' (Taves *et al*., 2018, p. 210).

We can now elaborate on a feature that differentiates human agents from animal and artificial agents. It can safely be said that, in any system that interacts with an environment and that can be characterised as a learning system, the manifestation of a worldview already implies a manifestation of a world-model; but the manifestation of a world-model does not imply the manifestation of anything as sophisticated as a worldview. Both a world-model and a worldview could be more or less adequate for an inhabited world/environment, but their adequacies or inadequacies cannot be evaluated using identical criteria.

The concept of the world-and-self model, together with natural and conventional affordances, should enable us to integrate worldview, way of life, worldly constructions, and environments into an organic whole. A worldview already presupposes a world-and-self model; on the basis of the more prosaic and factual assumptions and hypotheses of the world-and-self model, it forms a normative structure that potentially enables its inhabitants to make meaningful sense of a world. At least, this will tend to be the case, provided that the worldview is sufficiently attuned to the world *actually* inhabited.

Having made the distinction between worldview and world-model, it is now possible to understand an important dynamic of human–world interaction. It has previously been stated that modernity is characterised by a multitude of worldviews, many of which are in conflict with one another. In some instances, it may be possible to claim that the worldview of a particular individual or group of individuals is so out of touch with reality that it makes those who inhabit it dysfunctional in society. But a worldview that is in disagreement with the world as it is currently configured does not in itself imply dysfunction, at least not in a practical sense. Provided that the world-model, the *predictive* model, is adequately attuned to the way in which things function, it

should be possible to inhabit all manner of radical and diverging worldviews and still function in society. Let us posit instead that a worldview becomes dysfunctional when, for whatever reason, it can no longer adequately suggest any long-term purpose and meaning that can be experienced as useful and/or plausible by the agent that inhabits it. But even in such instances, as the example given by Walsh demonstrates, it may not be as straightforward as simply blaming the worldview. Just as diseases can be triggered by environmental factors, in the cases of which medication might not provide the best remedy, so too a crisis of meaning can be triggered by an environment or a world that is alien to the inhabited worldview. All that could be plausibly argued in such instances is that the worldview is no longer adequate to the task of making sense of the human environment/world relationship. But one may still be justified in thinking that the type of environment that does not match the worldview is not an environment that ought to be inhabited by human beings. One could be – and many are – convinced, instead, that we ought to make the *world* more attuned to the norms suggested by the worldview that one inhabits.

Now that a general understanding of worldview has been outlined, a few worldview traits that could plausibly be expected – at least, this will most likely have been the case up to now – to produce modern conditions are suggested. Let 'modern condition' stand for any kind of condition that is likely to be perceived as eminently controllable and manageable by means of reason, science, and technique, and/or likely to be perceived as eminently suitable for the flourishing of a buffered self.[107] As has been repeated again and again, a characteristic trait of modernity, especially late modernity, is a certain pluralism. Contemporary modern societies tend not to be defined or framed by a single religion or a single '-ism'. Liberal societies boast that they have room for many lifestyles and political persuasions. Common contemporary answers to the five big questions can engender very different worldviews. While some contemporary worldviews can be at odds with modernity and technological society (remember the Amish), many worldviews that are at odds with one another can still be modern, compatible with, and, to a large extent, producers of modern or borderline modern conditions. To the extent that such worldviews cease to be producers of modern conditions, they can be expected, at least for some time, to prolong the production of highly technological conditions.

Modern worldviews, it is posited, can differ in the types of answer that they suggest to the questions of epistemology, axiology, praxeology, ontology, and cosmology, as long as certain key aspects are emphasised. For instance, a worldview that suggests that only the material world exists (ontology) and that we are the result of arbitrary molecular combinations, and that when we die we simply cease to exist (cosmology) can be modern. A worldview that

[107] 'Buffered self' is discussed under the heading 'Changes in philosophical anthropology'.

suggests that there is also a spiritual reality alongside material reality (ontology), that we are God's creation, and that when we die we return to God (cosmology) can also be modern. But none of them is necessarily modern.

A worldview is rendered modern by certain emphases in the type of answer suggested to the questions of epistemology, axiology, praxeology, and ontology, and, crucially, in the self-understanding or philosophical anthropology suggested by an enacted self-model, the worldview as a whole, and the narrative structure that is categorised here as mythological. As the mythological narrative is treated separately, the category of cosmology is omitted; but it is to be understood that the mythology also suggests answers to cosmological questions. The following type of answer renders a worldview modern: In our pursuits of knowledge (epistemology), as it applies in this world, we must primarily rely on reason, time-tested scientific methods, and evidence-based research. We should primarily strive (axiology) to improve human life conditions here and now. The best ways (praxeology) to improve human life conditions are by using reason methodically and reliably, to pursue scientific research, and to develop technological solutions to most if not all identified problems. Whatever other dimensions of reality may or may not exist (ontology), material reality somehow stands out as constituent of the human habitat-arena: it is primarily this dimension of reality in which humans can and ought to apply their ingenuity and industry. To these answers other answers could be added, as long as the other answers do not annul the primary answers.[108] A modern theistic worldview, for instance, could add to the answers already given that, to some extent, revelation is another reliable source for God's ultimate designs (epistemology), that the improvement of life conditions presupposes that we need also to strive for knowledge about God's designs (axiology), and that the practice of science and technology needs to be complemented by the practice of certain virtues and/or ethical enquiries (praxeology) so that 'improvements' become compatible with God's designs. On such a modern view, changes incompatible with God's designs would not count as progress.

The self-understanding, or philosophical anthropology, suggested by this type of worldview is one in which the human being is an agent that is largely in charge of the forging of its earthly destiny or the enabling of its earthly flourishing. Divine or spiritual dimensions may be admitted; but if they are, it is still with the understanding that the human agent inheres a mature state in which it is to use the worldly resources at its disposal in order to improve conditions in the here and now.

A worldview as so far defined, if not complemented, is liable to be an abstract and 'wobbly' construct for any agent that happens to inhabit it. It will

[108] The reader will note that the humanism that Pinker advocates is missing. Pinker's version of humanism is a possible extension to the 'bare necessities' proposed here; but so too, we could agree with Bauman, is the anti-humanist ethos that animated the historical concentration camps.

contain beliefs and convictions – but be tied to no narrative. In order to provide conditions for the emergence of durable purpose and meaning, the worldview structure needs to be complemented by a narrative structure. The narrative structure is developed here under the category of mythology. Another and perhaps simpler option would have been to include it in the category of cosmology within the concept of worldview being used. The epoch-defining trait of the narrative, however, merits a separate and grander category.

The mythological narrative, the theoretical reconstruction of which is undertaken shortly, vaguely suggests where we come from, how we liberate ourselves from initial limitations, and what our *telos* is. The narrative is an epoch-defining narrative that shapes and reinforces key content that make worldviews modern. It is this ideal, normative, and teleological self-understanding that is likely to be challenged in algorithm-intense environments. How such challenges are met will likely decide the fate – obsolescence or evolution – first of modern worldviews and, ultimately, of the epoch-defining narrative itself.

Mythology in modernity and beyond

First, a very simple narrative about the history of humankind is presented. It should be familiar to all readers. It is embraced by popular consciousness, and generally does not contradict the understanding of most scholars. It starts in pre-history, reaches the present, and suggests a continuation.

Once we were hunter-gatherers. Primarily because of the cognitive capacities of the human species, human tribes could colonise most of the planet. As centuries and millennia passed, we learnt to make more and more advanced tools. Nomadic ways of life were abandoned for settled agricultural ways of life. The Stone Age gave way to the Bronze Age. Alongside rural ways of life, new urban ways of life were enacted. Societies began to industrialise, and laid the foundations for a continual technological advancement of human life conditions. From all indications, it may be expected that tomorrow will be an even more technologically complex version of today.

Although, so far, no value is embedded in the narrative, there is a teleology that almost asks to be read into it: as a species, we follow a developmental trajectory, from immature to mature, from lower to higher, from primitive to advanced know-how and knowledge. 'Developmental' implies, in contrast to 'evolutionary', the kind of fixed pattern according to which a human being develops from infant to child and from child to adult. We are not obliged to read something like the pattern of the developmental teleology of a human being into the history of humankind; but it is easily done.

The narrative of the development and rise of humankind is echoed in innumerable instances of popular culture all over the world, as well as in academic disciplines.[109] In proposing a theoretical reconstruction of the inferred

[109] Some examples of academic work are discussed later in the text.

mythology of modernity, nothing original is presented. The reader will recognise that, like the story presented above, it will be something very familiar. It is also argued that it is a story with which many prominent thinkers have engaged. Indeed, it could justly be argued that it is also the story with which *we* have been engaging, piecemeal, until the present moment.

That the reconstruction is categorised as 'mythology' should by no means be taken to imply that people who argue along similar lines are necessarily wrong, nor that they are necessarily right. It is to be expected that a dominant epoch-defining narrative will reflect real patterns in human history, albeit vaguely and inaccurately in relation to factual events and processes. What makes it mythological is not the factual truth or falsity of the historical events and processes that it suggests, but its normative and formative influence over the epoch that it defines. A mythology, in the sense that it is given here, need not be in either agreement or disagreement with particular facts in history. If it is to become epoch-defining, it must nevertheless resonate with actual human experience of the world and its history. If a narrative fails to resonate, it will certainly also fail to become an epoch-defining mythology. If an epoch-defining mythology ceases to resonate, its status will be undermined.[110]

Some of the thinkers we have covered propose that we are in the process of exiting modernity, or that we have already left it, or that we have never even been modern. In contrast, here it is posited that most contemporary technological societies are still being motivated through the influence of an epoch-defining modern mythology, especially in how they approach the construction and maintenance of society. The world-making output that results from a people's being informed by such a mythology need not resemble any particular ideal 'modern condition'. Even if one agreed with postmodernists that contemporary societies have undergone a loss of belief in grand narratives, it could nevertheless be maintained that, as long as none of the succeeding micro-narratives gain the status of a grand narrative, and as long as not all have lost faith in some tenets of the old mythological narrative, the mythology is likely to retain much of its persuasive force. Here it is argued that the status of this mythology itself, owing in part to the output (such as algorithm-intense environments) that it motivates its inhabitants to produce, is being increasingly undermined.

A few words need to be added on the subject of 'modern condition'. In its most general sense, the concept refers to any kind of life condition in which it becomes not only practically possible but also expedient to practise the kind of way of life that modern worldviews suggest ought to be practised. This could imply, for instance, that it becomes expedient, in *a* modern condition (but, it must immediately be added, not in *all* possible modern conditions), for one to take charge of one's own affairs, to assume individual responsibility

[110] It should follow that there might also be narratives that resonate strongly, but for too brief a time and/or among too few people for the narratives to ever become epoch-defining.

for one's actions, with the underlying assumption that humans in general, provided that they are adequately raised and educated, can understand their role in society in a way that is adequate for participation in democratic processes. But what if a people who inhabit a modern worldview that suggests such a way of life were to enter conditions in which such a way of life remains theoretically possible, but in which it is no longer experienced as expedient? What if such a way of life begins to be experienced as increasingly burdensome, or if it increasingly fails to make sense? Such changes in life conditions would probably have profound repercussions on worldviews and the experienced relevance of narratives. Such dynamics, at least, could be used to map out or characterise borderline conditions. Further removed still, of course, are conditions that are outright incompatible with the practice of the ways of life that are suggested by modern worldviews.

Here we recognise the by now well-known complexity that must be considered in the context of design and implementation of all types of automated systems. We cannot rely simply on *intended* effects; we must look further and seek to foresee possible *actual* effects. Let us again consider the problems described by Wiener in part I related to intended versus actual effects. Could we understand some of the difficulties that emerge in our environments as the result of an analogous process in the worldview–world relationship – that is to say that the world that is *actually* constructed always differs from the world that a collective of persons, based on the worldview they inhabit, intend to construct? This must certainly be the case.

The reconstruction of the mythology of modernity is presented in the form of a narrative, which in turn suggests a philosophical anthropology. Both the narrative structure and the philosophical anthropology that it suggests are vital to the whole of modern worldviews. If we remove the philosophical anthropology, then modern worldviews cease to be modern. If we remove the narrative structure, then all worldview content becomes abstract, ceases to make holistic sense, and begins to fragment and then fall apart. This is so because, to the extent that they *do* make sense, human lives follow narrative patterns.[111]

The reconstruction takes the form of a narrative that suggests a basic approach to the big questions. This particular narrative is one that fits the western imagination particularly well. In important respects it overlaps with many popular and scholarly descriptions of human history. The reader may recognise notions from previously discussed thinkers such as Ivan Illich, Jacques Ellul, Zygmunt Bauman, Romano Guardini, and Steven Pinker. However, the narrative is evaluative in a sense that runs counter to some of these thinkers'

[111] This certainly also applies to postmodern conditions defined by a complete loss of belief in grand narratives. Micro-narratives would then compete with one another for human allegiance. In time, one of the micro-narratives might gain the status of a new epoch-defining mythology.

reasoning. To make it simple, it is given a beginning, a mid-stage, and an open-ended modern phase.[112]

Once upon a time, we have been told, people lived in tribes. Inhabiting savannas, shorelines, steppes, or deep forests, they conceived of themselves as merely one type of agency among many other types of awe-inspiring agency. All agencies were embedded in their respective sacred environments. Nature appeared to be a wholesome force in itself. The order perceived in nature functioned as a model in accordance with which such peoples could work out their ways of life and imagine their cosmologies. Nature could not be questioned, only mediated in pursuit of equilibrium. To trespass on the limits imposed by sacred nature was understood as hubris. The act of hubris threatened to undo the coherence of the tribe. Harmonious commerce with the surroundings of the tribe, on the other hand, promised survival or, at best, limited prosperity. For the inhabitant of the mythology of modernity, something like this represents the original condition.

In another age, people looked down on this distant past, when people were subject to the terrors of their superstitious beliefs. The human creature had reached a level of self-awareness that squarely distinguished it from other creatures. Nature was no longer populated by awe-inspiring agencies. It was no longer sacred. The sacred, along with the awe-inspiring agencies, had migrated to the realm of the transcendent. Human society now modelled itself in accordance with a higher order. In relating itself as *human* to that higher order, humanity breached its parochial boundaries and attained a universalist perspective. The beauty of a divine order was understood as an image of the well-ordered human soul and society. Hubris was now understood as intentions and acts that defiled such beauteous order. Defilement led to dissipation, and dissipation to disorder. Pious contemplation of divine order and adjustment of society in agreement with it, on the other hand, promised salvation. For the inhabitant of the mythology of modernity, something like this represents the middle condition.

In our time, people tend to look back on these distant pasts, through interpretative and evaluative narratives such as the one presented here, as backward and superstitious stages on the way towards a rationally and/or scientifically informed modernity. The mid-stage de-sacralisation of nature and the attainment of a universalist perspective are counted as gains. To some (but not all) moderns, the mid-stage belief in another type of sacred and in a transcendental agency are counted as just as superstitious as the beliefs of our more distant predecessors. The moderns who still retain a belief in some form of the sacred and/or in a transcendent agency nonetheless tend to align with other moderns

[112] As in the previous case of modern worldviews, it will be easy to imagine many variations on the same narrative structure: some, for instance, with three, some with four, some with five stages. The reconstruction presented here should be understood as only one of many possible variations on the same pattern.

in their understanding of the human condition here on Earth. The human subject has reached a state of maturity in which the responsible thing to do is to take charge of things and to improve the world. Improvements to life conditions are possible mainly by means of reason and the further development and refinement of science and technology. To the extent that it is possible to control things in this world, we humans, and no others, are the controllers.[113]

This teleology of a progressive liberation from natural limitations and superstitions is basically the one inaugurated by the Enlightenment era. The means available for accomplishing the teleology are understood at once as existing inherently in the autonomous individual agent (capacity for reason) and as the product of the collective use of reason (a peculiar reason- and science-informed culture). Together, individual reason and its cultural product reinforce each other.

Modern worldviews, modern scholarly theories, modern political ideologies, and so on will provide many more details and specifics than the simple and vague narrative presented here. In one modern mythology, we could posit, many contradictory modern views can incubate. Steven Pinker argues that democracy is a major facilitator of human flourishing (Pinker, 2018, pp. 199–213), If progress is measured by increased chances of human flourishing, then democracy – given that we assume with Pinker that democracy facilitates human flourishing – could be understood to represent progress over prior non-democratic modes of governance. Still, some historical Enlightenment thinkers, such as Thomas Hobbes and Voltaire, tended to favour enlightened absolutism. One thing that unites these Enlightenment thinkers with Pinker is the conviction that it is time for humans – those who are enlightened – to rise above superstition and archaic tradition, to take charge of their own affairs, and to improve the lot of humanity.

If Enlightenment thinkers have reached contrary positions on a number of subjects, this is no thorn in the side of the Enlightenment. A thinker of Pinker's inclination – one who values democracy primarily as a means to a more ultimate end – could argue that Hobbes and Voltaire, like many other Enlightenment thinkers, simply reached bad conclusions. In time, if enlightened thinkers do not relent in their pursuit of reason-based progress, time will weed out bad ideas whereas good ideas will be retained. On this view, it could be argued that the triumph of democracies has proven, or that a coming triumph of democracies might prove, Pinker's position to be the better one. For the purposes pursued in this treatise, it would be more interesting to argue that Hobbes and

[113] To be clear, this admittedly crude picture of human history is not presented in order to imply that all or even a majority of human beings, even in so-called modern societies, must subscribe to it as veridical. People in general no doubt have many different and often contradictory intuitions and views about human history. If we posit that this consistent historical narrative represents an epoch-defining mythology, then this does not entail that everybody must subscribe to its veracity; it implies that it should be vaguely familiar, at least in parts, as dominant tropes to all who inhabit modern worldviews.

Voltaire may actually have been right *in their times*, and that other ideals may have been right *in succeeding times*, and that other ideals still may be right *in times to come*. Let us pursue this line of thought.

As Europe slowly woke up from its dark mediaeval slumber, the argument could begin, knowledge was scarce. Its populations were tyrannised by superstitions propagated by the Church. Only a few heroic enlightened people existed. It was only by means of the enlightened rule by such people that the tyranny of the Church could be broken. Then, as the broader swathes of Europe's populations began to be educated, the need for enlightened absolutism began to wane. More and more enlightened individuals became fit to take charge of their own affairs. For such populations, guided primarily by reason, democracy would probably be the superior alternative. We could also characterise this kind of development as a succession of *different modern conditions*.

The argument presented here is admittedly crude, and should not be taken to represent all Enlightenment thinkers. Nevertheless, to an extent it mirrors the understanding of John Stuart Mill. If Mill is famous for his advocacy of individual liberty, he appears to have been of the view that liberty as a 'principle' cannot apply to conditions in which humankind is not yet 'capable of being improved by free and equal discussion'. For Mill, '[d]espotism is a legitimate mode of government in dealing with barbarians, provided that the end be their improvement, and the means justified by actually effecting that end' (Mill and Gray, 2008, pp. 14–15). To be clear, Mill is probably not referring to 'enlightened' despotism in the context from which he is quoted here, but to despotism in general; and he argues that it is only justified in supposedly 'barbarian' conditions.

This line of reasoning is interesting in the context of this treatise not because we must agree or disagree with Mill, but because it combines a high ideal of the human agent with a practical endeavour to re-educate humans so that social conditions begin to approach the ideal.[114] Furthermore, it makes the assumption that the functionality of any fairly complex society will rely on a set of specific agent-types. The agent-types that supported the social structures in Mill's British context could not be reasonably expected to provide any

[114] The argument implies that means to ends must adapt in accordance with concrete circumstances, and that there are no particular means that will simply be right or best at all times and in all places. For Pinker, the ultimate end is not democracies, but increased opportunities for human flourishing. Democracies, perhaps, represent a crucial step towards that goal. Let us accept this view for the sake of our argument, and explore some possible implications. At the time in history when democracies became viable, the mindset was no doubt such that, to an extent, their constituents had already become accustomed to thinking of themselves as individual responsible stakeholders in statecraft. The expansion of suffrage rights no doubt proceeded as the result of persistent popular struggles *and* a vital recognition by the upper classes of the fundamental services rendered and value added by the popular classes to the state. Presently, as Susskind informed us in part I, in democracies we are increasingly told that there will be no need for our skills in the future. If we will be in a position to add value to the whole, then the values that we add might just be extracted from us, as illustrated in part II. Given such trends, what are we to think of democracies and human flourishing?

equivalent support in a feudal Japanese context, and vice versa; these two societies required different modes of education and training for their key agent-types. And here we could revert to the (by this stage) repetitive line of questioning. What will become of human populations as more and more artificial agents appear to excel in the display of enlightened virtues – instrumental rationality, knowledge, and know-how – over and beyond levels which can reasonably be expected from averagely skilled or even expertly skilled human agents? What will happen if humans become de-incentivised to learn critical skills and to pursue critical knowledge? To what extent will humans in such contexts experience that they have reached an adult, mature, and responsible historical state? To what extent will they find that they are well-equipped to participate in important decision-making processes expected to affect the fate of their polities? To what extent will they experience that they can exercise important degrees of control of their own personal lives?

An epic rise of humankind, a goal, and the modern philosophical anthropology

Here we pay closer attention to the pattern of an epic rise of humankind in the narrative presented earlier, and to an emerging *modern notion of the human agent*. Let us revisit the three stages of the narrative of modernity. Let us look on these stages briefly and partially through the eyes of some scholars, whose scholarly interpretations could be understood as a sort of engagement with the narrative of modernity. The work of prominent thinkers such as James Frazer (1996), Karl Jaspers (1953), and Charles Taylor (2007), to mention just a few, offer understandings that imply something similar to the narratives that have been proposed. These scholars, of course, are not directly preoccupied with the theoretical reconstruction presented in this treatise. It is more likely the case that *I*, the author, have been inspired by them and by many others in the acts of inferring and reconstructing such a narrative. These scholars endeavour to untangle and explain developments in the fields of anthropology, philosophy, and history. However, if the inference made here is justified, they also engage, of course, with an epoch-defining mythology; we all do. Be that as it may, in considering some notions proposed by these thinkers, we might hope that we shall arrive at a more concrete notion of how a modern outlook on history, the present, and the future could be conceived. We should then be better able to figure out whether and how a modern outlook could be plausibly retained in algorithm-intense futures.

Frazer, a pioneering and influential anthropologist, discerned a historical progression from early religion in the form of fertility cults towards scientific knowledge (Frazer, 1996). Jaspers popularised the idea of an ‘Axial Age’, and affirmed the notion that history has a goal (Jaspers, 1953). Through the modern eyes of a follower of Frazer and Jaspers, a very brief history of humankind could read as follows.

In prehistoric times, our ancestors were embedded in natural environments. They were subject to complex cycles and random events that occurred in nature. In their own imagination, and by means of fertility cults and animal and human sacrifice, as imagined and chronicled by Frazer and other early anthropologists, their fate was mainly to adapt to rather than to shape the vagaries of their habitats. Moreover, nature was understood to be inhabited by the sacred: human agents had no option but to adapt to the more awe-inspiring non-human agencies that they imagined were embedded in nature. From the point of view of the modern observer, even many of the grandiose archaic civilisations represented but very minor steps on the journey towards modernity. We had to wait for something like the *Axial Turn* for a genuinely new beginning. The notion of the *Achsenseit* (*Axial Age*) was popularised by Karl Jaspers in *Vom Ursprung und Ziel der Geschichte* (*The Origin and Goal of History*), the title of which, incidentally, supplies some possible key categories of a well-rounded epoch-defining mythology. According to Daniel Austin Mullins *et al.*, one of Jaspers' motivations was in fact to 'salvage what he called the "spirit of Europe" in the midst of post-war devastation' (Mullins *et al.*, 2018, p. 598). A function of mythology, in accordance with how the concept is used here, is precisely to impose a common narrative – that is, a common historical purpose and meaning – on the people who are informed by it. This, it seems, overlaps with Jasper's own understanding of the meaning of history.[115]

Since Jaspers popularised the notion, the Axial Age has been much debated by scholars. The Axial Age is usually located between 800 BCE and 200 BCE. During this time, a series of revolutionary developments are understood to have occurred independently from one another across the regions of the eastern Mediterranean, Persia/Iran, India, and China. According to Mullin's and his co-writers' overview of claims made in prominent works on the Axial Age, it is supposed to have seen the 'emergence of universalizing moralizing religious/philosophical traditions […] of a universal egalitarian ethic […] of a doctrinal tension between the transcendental and mundane orders among elites […] of second-order thinking or the increased reflexivity and critical evaluation of existing religio-philosophical expositions among elites' (Mullins *et al.*, 2018, p. 601).

The article of Mullins *et al*. represents an attempt to evaluate empirically claims that researchers have made in reference to the Axial Age. By means of the methodology that they employ and the admittedly limited data selection that they use, the authors find that claims often appear to be based on anecdotal narratives, that researchers have a tendency to cherry-pick data that validates their claims, that it is doubtful that the phenomena lumped together under the Axial Age really manifested in an 'age', and that there is little support for many axial claims (Mullins *et al.*, 2018, pp. 605, 607, 612, 613, 615). They

[115] In *The Origin and Goal of History*, Jaspers considers the meaning of history in the third and final part (Jaspers, 1953, pp. 231–276).

assert, nevertheless, that many of the phenomena characterised as 'axial' appear to have emerged during the course of human history and that 'axiality' merits further study.

This is not the place either to defend or to attack the notion of an Axial Age. The claims advanced by Jaspers and other axial theorists may be historically accurate, as far as we know. Note well, however, that the lumping together of a set of profound changes so that we produce a discrete age of amazing progress, and the suggestion that history and human life have a goal, also amount to captivating story-telling. If humans construct meaning through narratives, then it stands to reason that it is also a distinctly human thing to do. On the other hand, it does not come as naturally or as easily to us – agents who inhabit narrative structures – to perceive historical changes as value-neutral evolutionary processes that are contingent on external circumstances. We need narratives to make sense of things. One way to structure captivating narratives is by means of powerful analogies. Thus, for instance, we may perceive historical events in analogy with how we perceive the development of an organism: from infant to child, from child to adult, and from adult to… Well, we shall see whether there might be a narrative-mythological continuation of this teleology beyond senescence and decay.[116]

In the case of Frazer, primitive religion appears to have been considered not only from a modern point of view but also with a very modern frame of mind. In the introduction to the abridged version of *The Golden Bough*, George W. Stocking, Jr writes that 'Frazer thought of religion in Taylorian[117] terms as essentially a philosophic system built on mistaken premises, and his writings […] could very well cause thoughtful readers to reconsider beliefs fundamental to traditional Christianity' (Frazer, 1996, p. xx). This implies that the relationship of primitive humans to primitive religions must have been similar to the relationship of modern humans to philosophy, and perhaps even to science – that is, that the function of religion must have been understood by the primitive subject similarly to how the function of modern philosophy and science is understood by the modern subject. Frazer sees a progressively ordered development that carries humankind from magic to religion and, ultimately, to science. Stocking's further comments on and quotes from Frazer bring to light a thoroughly modern worldview:

> Science emerged as men, 'after long ages', began 'to realize that entreaty [to gods or to God] is also vain', and once again tried compulsion, only this time 'within narrower limits and in a different way' (in Marett and Penniman 1932:

[116] A common mythological continuation of this pattern is renewal. After death, the life cycle starts again. Later we shall see how some people envisage a linear continuation from something like a 'mature' or 'adult' human stage, via a transhuman transition, towards a superior posthuman stage.

[117] The Taylor referred to is the anthropologist Edward Burnett Taylor, a contemporary to Frazer. The abridged version referred to is Frazer's own abridged edition of his original twelve-volume edition.

> 41-2). Later, Frazer spoke in less linear terms of 'interwoven threads' of black and red and white, and cautioned that science itself might be superseded by 'a more perfect hypothesis'. But there was never any doubt that the 'hope of progress – moral and intellectual as well as material – in the future is bound up with the fortunes of science, and that every obstacle placed in the way of scientific discovery is a wrong to humanity' (below, pp. 854-5).[118] (Frazer, 1996, p. xxi)

Crucially, they – the ancients – are imagined as having been mistaken in their premises, the error of which is now being rectified on an ongoing basis by modern enlightened scholars. Few contemporary scholars of religion would embrace Frazer's understanding of religion. It should also be mentioned that Frazer was controversial in the budding anthropological field of his day. Many contemporary anthropologists have been preoccupied with re-evaluating so-called 'primitive' ways of life.[119] Indeed, it has become something of a trend in our late-modern times to relativise and sometimes to devalue the dominant state of affairs in western societies in comparison with alternative non-western and/or pre-modern ways of life. As Guardini describes the matter, we have become too aware of some of the potentially dangerous consequences of technologically empowered civilisation to ignore its ambivalent value.

If we accept that the modern imagination is structured by the notion of a goal and a progressive view of history, then what precisely makes it modern? Within specific worldviews, goals can assume different forms. To the extent that worldview-goal structures are modern, we should be able to re-state specific main goals in more general terms, such as 'liberation from undesirable limitations' and/or 'empowerment' and/or 'enabling of human flourishing', *with the understanding that the primary means to reach such goals involve the use of reason, the advancement of scientific knowledge, and the exercise of technical know-how*. But what could we say about how the human being itself is being imagined?

Let us consider some possible key changes in human experience that may have paved the way from pre-modernity to modernity. Charles Taylor, a contemporary philosopher, engages with the history leading up to modernity in a way that is less marked by *alignment* with the story of modernity – that is, with its many and varied ideals and end-goals. Taylor, instead, tends to see ambiguity in most developments, so that that which in some respects can be represented as a gain usually also implies a loss. In *A Secular Age*, he is primarily concerned with phenomena that may have enabled the emergence of a secular age to begin with. The extent to which a secular age – or modernity – represents a gain or a loss is left open to further discussion. Moreover, Taylor is preoccupied with 'accidental causes' for change towards modernity. Axial theorists, by contrast, tend to direct our attention towards pivotal thinkers

[118] The latter quote is from the end chapter of *The Golden Bough* (Frazer, 1996, p. 854).
[119] See, for instance, Marshall David Sahlins (1972).

and/or ground-breaking ideas and their historical implications.[120] On Pinker's understanding, the heroic feat of some thinkers in the context of the Enlightenment originally enabled much of what we think of as modern progress. In Taylor's typical analyses, we are given to understand that, once they have gained traction, ideas that are forged for very specific purposes often tend to spur evolutions that reach far beyond the designs of the thinkers who originally forged them.[121] One of several leading catalysts for change towards modernity is represented by developments in mediaeval theology – developments that, during the times in which they occurred, had nothing to do with the pursuit of modern ideals but that nonetheless ended up laying the foundation for a modern secular culture[122] (Taylor, 2018, pp. 25–218). In these analyses Taylor builds and expands on previous works, such as Max Weber's treatise on Protestant work ethics (Weber *et al.*, 2012).

Whereas Frazer and Jaspers, in the sense that they embrace a progressive and universalist view of history, could be considered fairly exemplary representatives of mainstream modern thinkers, Taylor is difficult to categorise. The reason is that his understanding of a transformative process, from an enchanted (pre-modern) age into a secular (modern) age, does not imply a 'progress view' of history. Indeed, in parsing his analyses, one gets the inescapable impression that, although much has undoubtedly been gained, much has also been lost. We cannot know whether ultimately the gains will outweigh the losses or the losses outweigh the gains.

Taylor's discussions of changes in human self-understanding are of special interest to the present topic. As European societies undergo a number of other

[120] As for Jaspers' own understanding of what caused the Axial Age, after having considered a few possible explanations, he concludes that it is something of a mystery and leaves the question open (Jaspers, 1953, pp. 13–18). Jaspers is more concerned with the meaning of the Axial Age. Unlike culture-specific grand events, the Axial Age, according to Jaspers, is common to all people. The reason for this is that it supposedly occurred in several key regions independently of one another. On Jaspers' interpretation, the Axial Age represents a historical stage in which '[t]he whole of humanity took a forward leap' (Jaspers, 1953, p. 4). This reflects a progressive and universalist view of history. But the main point for Jaspers eludes mere historical facts, for the meaning of history in some sense depends on our awareness of, our being shaped by, and our reflections over the Axial events. The post-axial subject is one that, in overcoming narrow self-enclosed historicity, is becoming increasingly aware of a shared humanity: 'Our present-day historical consciousness, as well as our consciousness of our present situation, is determined, down to consequences I have only been able to hint at, by the conception of the Axial Period, irrespective of whether this thesis is accepted or rejected. It is a question of the manner in which the unity of mankind becomes a concrete reality for us' (Jaspers, 1953, p. 21).

[121] If this appears to coincide with how the social impact of technologies is understood in this treatise, then this is probably no coincidence. If we find either the analysis undertaken in part II or Taylor's analysis plausible, then we should be willing to entertain the notion that what holds for technologies also holds for ideas, and vice versa. For what are technologies if not materially instantiated ideas?

[122] Note the similarity of pattern with a historical development commented earlier, namely that of the clock being invented by pious monks in order to structure their prayer life. According to Lewis Mumford the clock, unexpectedly, afforded a thorough reorganisation of society in accordance with the industrial mode of production (Mumford and Winner, 2010, pp. 12–18).

important changes, a porous self-understanding gradually gives way to a buffered self-understanding. The 'porous self' represents the human creature as a subject that is under the continual and sustained influence of opaque spiritual forces that are embedded in its life-environment. At its most extreme, everything in the human environment will have spiritual significance. The influence exercised by such non-human and non-visible agencies is only vaguely grasped. In order to resist, cope with, or invite the influence of such agencies, special knowledge, special know-how, and special rites are required. In such contexts, symbolic intermediaries that mediate between 'this world' and 'that world' will be put into place. This means that the porous self is not fully an inhabitant of 'this world'– that is to say, a world that can be subject to transparent analysis and practical reconstruction. He or she is torn between 'this world' and another dimension or realm that is opaque to practical understanding.[123]

The 'buffered self' represents a self that is shielded from the forces that affect the porous self. The image of the buffered self, in Taylor's account, replaces the porous self as belief in spiritual beings fades and as prominent thinkers adopt a more empirical and utilitarian approach to nature. The Judaeo-Christian belief that human beings are created in God's image also plays a role in the emergence of the buffered self. Of course, that belief can have many implications, depending on how God is understood and the context in which the belief is believed. Be that as it may, in the western part of the Christian world in the late Middle Ages, nominalism was developed as a way to defend God's omnipotence. Nominalism implies that there can be no laws in God's creation that God cannot annul or change. It is always and everywhere God's will that upholds the order of things. God is an absolutely sovereign and all-powerful creator and ruler. In comparison with earlier and rival mediaeval understandings, such as that of Thomas Aquinas, things and creatures of the created world now lose their inherent purposes. Instead of a disposition that allowed for pious reverence towards the inherent beauty of a created world, all reverence is to be directed to God. God alone is worthy of our admiration. Moreover, as humans alone are understood to have been created in God's image, it becomes possible to infer that humans alone retain a purpose.

Taylor argues that, in early modernity, the already entrenched understanding of human beings as created in God's image began to evolve into an understanding of human beings as God's co-creators. It goes without saying that God imagined as all-powerful cannot be under the influence of some petty spiritual forces embedded in nature. For creatures assumed to be created in God's image, it becomes plausible to dismiss the influence altogether. As the enchanted view of nature waned, belief in such forces waned also. Human co-creators, intuiting the purposes of God, were becoming busy reorganising

[123] Is it an imprudent leap to posit that an analogous pattern would apply to humans embedded in algorithm-intense environments?

society in accordance with higher ideals. And as the notion of an interventionist God began to lose its force among the populations of Europe, human beings alone remained as stewards and caretakers of a world reduced to instrumental usefulness to *human* purposes.[124] The purest expression of this philosophical anthropology is perhaps best represented by the popularised understanding of cartesian dualism, in which, by means of the proper use of reason, a mind (*res cogitans*) is given the purpose of becoming a sort of controller of matter in time and space (*res extensa*). According to cartesian dualism, in order for the mind to be able to control its environment, it must first exercise control over its body. Thus, the human body, like the human environment, becomes instrumental to the purposes embraced by or invented by the mind. Cartesian dualism is by no means a position that modern worldviews must necessarily include; it is, rather, an extremely influential fringe position in the larger ecology of modernity.

A means-to-end structure has now been identified. Within this structure, the means of reason and a science-informed culture are used to facilitate the betterment of human life conditions or, alternatively, to increase the opportunities for human flourishing. A human self-understanding that co-evolves with this structure has also been described. It is akin to a buffered intellect that sees its purpose as being in control of its environment – an individual mind that, to a large extent, is understood as autonomous.

As has already been stressed, modernity contains a plurality of worldviews. Nevertheless, if a view is to count as modern, it should be built on these basic foundations. With these as a base, innumerable extensions can be added to the modern structure.

Objections and clarifications

Some of what has been brought to our attention is now problematised further. If we accept what has been stated until this point as adequate descriptions of how significant numbers of subjects in modern societies will tend to view their world and understand their role in it, then how does this view correspond to how societies are actually structured? How inclusive is the conception of 'modern views'? How could modern views be differentiated from non-modern views?

The degree to which humans understand themselves as 'buffered' must be considered. The cartesian dualism between a mind that has the ability to control on the one hand, and the matter in time and space that is to be controlled on the other, has become a caricature. Incessantly attacked by critics, it has become infamous. But if so much effort has been spent criticising Descartes, it may be worth pausing to ponder the extent to which people actually tend to

[124] For the porous/buffered distinction, see (Taylor, 2018, pp. 27, 37–41, 83, 137, 300–301, 539–540). For the impact of nominalism, see (Taylor, 2018, pp. 97–8).

view themselves-in-the-world in this way. Indeed, much of the criticism is made on the assumption that humans are embodied creatures, that there is no mind that is separate from the body, and therefore that the distinction between *res cogitans* and *res extensa* is misleading, if not outright false.

Whether or not cartesian dualism accurately describes ontological properties is a question that is separate from whether or not it captures an age's view of itself. Well, does it capture some specific age's view of itself? The ambivalent reply is 'Yes and No'. Yes, because it seems to represent a teleological conclusion to the development of the buffered self. It may well also have captured the view that high modernity had of itself. However, the 'Yes' is increasingly turning into a 'No', because it is becoming increasingly doubtful that many people are still able generally to view themselves as controllers of matter in time and space to any high or meaningful degree. An even more common experience in late-modern times, no doubt, is a lack of control. But it is precisely in this tension that we must locate cartesian dualism, not as a description of reality as it tends to be experienced at any given time, but as an ideal set against and above common experience. *Modernity is a mindset endeavouring to be more and more in control over matter in time and space*. During high modernity, popular experience may have reflected the sense that society was approaching the ideal; in contrast, during late modernity popular experience may reflect the understanding that we are distancing ourselves from the ideal. At the very opposite of the control ideal we find a pre-modern philosophical anthropology, in which humans embedded in animated environmental contexts are in positions where they need to placate animistic forces rather than exert technical control over environments.

If we relocate the notion of *absolute control* to an ideal rather than to actual experience, then how should we think about the notion of the buffered self? By means of the concept buffered self, Taylor seeks to conceptualise something that adequately describes people's actual experience of self in the new kind of world that was emerging in early modernity. But is the concept sufficiently supple to be integrated into the varieties of worldviews that we, in accordance with previous categorisations, ought to recognise as modern? Is it compatible with a Christian worldview? The answer is that it is one of the factors that would render a Christian worldview modern. In a Christian worldview, the status of the buffered self cannot be separated from the understanding of God's role in the created world. A Christian belief in a divine hierarchy replete with greater and lesser spiritual beings and a fallen counterpart of demonic beings is perhaps not particularly compatible with a buffed self-understanding. On the other hand, a Christian belief that is more aligned with the developments described by Taylor would be quite compatible with a buffered self-understanding. On such a modern Christian view, humans are God's noblest creations, created in the very image of God. And if God is a reasoner and a creator, it could plausibly be inferred that, by analogy, humans ought to enact a relationship with the created world that is similar to the God-world

relationship. Humans, then, become like God, or, rather, like gods. There are no longer any lesser angelic beings to be petitioned or demonic forces to be feared.

It must be recognised that the worldviews of actual persons will often not be purely modern or purely non-modern. Worldviews informed by Eastern Orthodoxy or Roman Catholicism, which affirm the existence and activity of a divine hierarchy of beings and demonic powers, may be more or less modern, depending on how the whole is conceived. It may also be the case that the self-understandings of Christians in general would tend to be less buffered than the self-understandings of secular materialists. After all, Christians believe in at least one spiritual being of awe-inspiring power. Whether or not this is the case, within the line of inquiry here pursued, it may turn out to be of marginal interest in relation to the dominion of the buffered self. It is argued later that, owing to the construction of algorithm-intense environments, the sun may altogether be setting on the buffered self. We are not necessarily returning to something that is identical to a pre-modern porous self, but we may be moving towards something rather similar.

Now that one way in which a religious worldview could be rendered more or less modern has been described, something similar could be done with regard to a political worldview. Let us consider a conservative worldview. Conservatism can, of course, mean different things. Political conservatism, according to Immanuel Wallerstein, gained traction as a self-conscious ideology, or as a programme for political action, as a reaction to the French Revolution (Wallerstein, 2011, pp. 2–6).[125] Whereas conservatives can marshal support for their approach from pre-modern thinkers such as Plato, the conservatism of Edmund Burke aligns with the notion that a society can progress over time. Burke, although modern-minded, is anti-revolutionary. For Burke, progress represents time-tested and hard-earned gains that nevertheless rest on fragile foundations: a society must therefore proceed empirically and slowly in its endeavours to ameliorate its shortcomings (Burke, 2004). On the other hand, a conservative ideology overly concerned with geographically situated and/or historically grounded institutions that are assumed to be prerequisites for peculiar ways of life and ways of human flourishing will no doubt tend to distance itself from modern ideals. Moreover, if sacredness of such institutions happens to be emphasised, the ideology will certainly tilt toward the pre-modern. Conservatism writ large is compatible with both modern and pre-modern worldviews.

[125] The understanding of ideology as a programme for political action draws on Wallerstein's characterisation of 'ideology': 'Ideologies are not simply ways of viewing the world. They are more than mere prejudices and presuppositions. Ideologies are political meta-strategies, and as such are required only in a world where political change is considered normal and not aberrant. It was precisely such a world that the capitalist world-economy had become under the cultural upheaval of the revolutionary-Napoleonic period' (Wallerstein, 2011, p. 1).

Here the key variables at play are different from the ones in the example of Christianity. Suppose that a conservative understanding is incorporated into a wider Plato-inspired understanding that implies that it is in the nature of things in the experienced world, in contrast to ideas in the higher realm, to degenerate. If we embraced such an understanding, it may well be the destiny of aristocracy to devolve into oligarchy, of oligarchy to devolve into democracy, and of democracy to devolve into tyranny.[126] This kind of outlook does not seem to fit within the family of modern worldviews. Suppose instead that it affirms the belief in *slow* time-tested progress, enabled to an important extent by means of science and technology, but that it also problematises the complexity of culture understood as a prerequisite for progress. Thus constituted, political conservatism would fit within the family of modern worldviews. In the modern context, it represents a counterpoint to modern revolutionary movements.

Revolutionary worldviews could be understood in a similar way. They, too, can be modern or non-modern.[127] But the point has been sufficiently illustrated. Mainline modern worldviews can incorporate intuitions from both conservative and revolutionary mindsets.

How about some of the more romantic views that have been expressed in the context of the Enlightenment and modernity? Could they fit within the family of modern worldviews? Jean-Jacques Rousseau understands most civilisational structures to be impediments to human flourishing. The 'noble savage', elevated as an ideal by Michel de Montaigne, François-René de Chateaubriand, and Rousseau, becomes an ideal.[128] Yet Rousseau's outlook fits in the context of modern worldviews, for it is geared towards progress through a natural sort of reason unspoilt by the impeding conventions imposed by a corrupt society. On Burke's view, conventions, when they are in good order, are prerequisite enablers of human flourishing. Rousseau does not consider his contemporary conventions to be in good order. On his view, all or most conventions are sick, which is to say that the prevailing culture is sick; yet human nature contains the seed in itself to overcome this sickness. Rousseau and other romantics demonstrate that the evaluative aspect of structures such as the reconstructed mythology is unstable. It can be turned on its head, and thus elevate the scorned ancient primitive as an ideal.

We can tentatively conclude that more extreme conservative and revolutionary views, in their respective ways, can represent fringe positions of modernity as defined by Guardini. Followers of Burke may choose to emphasise

---

[126] See, for instance, Plato's *Republic*, book VIII.

[127] Do they, for instance, affirm the notion of human progress and the elevated status of the means of reason, science, and/or technology? This is hardly the case for all revolutionary movements.

[128] Montaigne broaches this theme in *Essais*, specifically in the essay titled 'Des cannibales' ('Of Cannibals'). Chateaubriand elevates the Northern American native as an ideal in *Atala* and *René*. Rousseau treats the corrupting influence of traditional education in *Émile, ou, De l'éducation* (*Emile, or On Education*) and asserts humanity's innate goodness in works such as *Confessions*.

the importance of time-tested culture and conventions, but still believe in and aim for progress. Followers of Rousseau, disqualifying contemporary culture as a corrupting influence, may instead choose to emphasise an unspoilt human nature as the primary means towards progress. Through the exercise of unspoilt human nature, humans could then presumably begin to construct a reasonable culture.

Romanticism does not necessarily imply any ideal of a noble savage. Frazer's writings exercised an important influence over his generation, and the archaic states that he portrays are savage and brutal. Some modernist writers, such as D. H. Lawrence, emphasise and value the rawer passions over and above reason.[129] Fredrich Nietzsche, hardly a modernist, elevates the will to power and heroism over and above reason. Here, no doubt, we are approaching the outer borders of modern outlooks. Some of the shoots that spring up from modern soils could be interpreted as being in open revolt against modernity. Still, as the work of Sigmund Freud demonstrates, a philosophical anthropology that grants important sway to passions and unconscious drives is not incompatible with scientific progress. Such human specimens, too, can be scientifically studied, comprehended, and, based on scientific understanding, be reformed in a way that better enables their flourishing. On Freud's understanding, it is the scientific method that makes possible the equilibration of human passions.

It is often argued – and no doubt rightly so – that modern times are coextensive with individualism. One should not therefore make the mistake of taking the figure of *a* mind controlling matters in time and space as a template for modernity. It is merely one of the ideal images that can emerge in the context of modern times. On a larger scale it must be admitted that the degrees to which humans have managed to control matter in time and space have depended on collective effort. High-modern conditions, based primarily on Guardini's account, appear to have invited a more eccentric individualism compared with the more contemporary late-modern conditions described by Illich, Ellul, and Bauman. Guardini argues that technological societies require that humans become more communal in order to be able to sustain increasing complexity. Even so, it is often taken for granted that, in some sense, most subjects in contemporary techno-modern societies understand themselves as individuals rather than as members of a class, profession, or tribe.

Regardless of how individualistic contemporary societies may or may not be, one often notices an ambivalence in how individualism is viewed. Rampant individualism is often understood as a threat to the modern condition that fosters individuals. It is telling that Pinker, a contemporary modern thinker, stresses the importance of modern institutions. Institutions are collective ventures, a kind of multi-agent organisation that possibly anchors modernity in

[129] Although considered a modernist, D. H. Lawrence certainly represents an anti-modern and romantic temperament.

more ancient institutional praxis. As pre-modern institutions provided frameworks for the collaborative efforts of pre-modern humans to achieve their goals, so do modern institutions. Whether they will continue to do so will depend on the developments that modern institutions will follow. Will they, as they have done previously, continue to provide the framework for the collaborative efforts of modern citizens? Will we live to see a new kind of human–AI partnership in the institutions of tomorrow? Or will institutions undergo increasing hybrid transformation to the effect that the need for humans is squarely reduced?[130]

If we take into account the discussions from part II, then our times will appear to be, in some sense, much less individualistic than is commonly assumed. The individualism that we discern in modern ways of life could just as well be described in terms of different forms of collective structures. Individuals do indeed experience looser bonds to traditional family and regional structures; but, as has been argued, they become more dependent on other collective structures. Here we could refer to these structures as modern institutions. This dependence is not always perceived or recognised, in which case we could perceive rampant individualism in contrast to how we perceive traditional ways of life. But this picture does not seem to be entirely justified.

Many have argued that modernity tends to impose conformity on populations. Concepts such as 'mass man', 'mass existence', and 'mass society' are frequently applied by Ellul and other 20th century existentialist philosophers, including Karl Jaspers. Mass man, simply put, is a sort of average, atomised and alienated product of mass society, the experience of which is a lack of individuality and responsibility. According to André Munro, mass society stands for a society that is at once homogenised and disaggregated. The individuals – mass people – who compose such societies are atomised: 'Mass society theory was based on the thesis that modernity had severely eroded the social fabric. In mass society, individuals are at once subsumed in the social totality and estranged from one another' (Munro, 2017). Towards the end of the 20th century mass society theories, according to Munro, began to be critiqued by thinkers who argued that:

> they relied on a romantic and inaccurate representation of premodern communities. Moreover, the idea that individuals in modern societies are uprooted and atomized seemed to be refuted by studies showing the persistent relevance of interpersonal relationships, intermediary groups and associations, and social networks. (Munro, 2017)

Peter-Paul Verbeek, critiquing the approach of Jaspers, argues that the mass existence notion 'was part and parcel of the historical time in which Jaspers

[130] Modern institutions are considered here from the point of view of those who still have faith in modern institutions. For many postmodernists, modern institutions have no doubt already ceased to serve any useful purpose.

wrote' (Verbeek, 2005, p. 35). A society shaped entirely by the assembly line mode of production can plausibly be understood to threaten human existence by closing off venues for human flourishing. However, the assembly line does not tell us all that is worth knowing about contemporary technologies. In more current contexts, Verbeek informs us, dehumanising aspects of technology are less evident. Meanwhile, information technologies such as cell phones and email 'have made possible moments of contact between human beings that are genuine and personal, and not merely functional' (Verbeek, 2005, p. 36). As information technology structures complement industrial structures, some argue that societies are becoming more diversified. Munro nevertheless informs us that, since the 1990s, some of the themes of mass society theory have been revived. He references Robert D. Putnam, who argues that a weakened state of civil society represents a threat to democracy (Munro, 2017).

If mass society, or mass existence, more aptly describes a society that is defined by the type of industrial structures that were prevalent during the times of Jaspers and Ellul, the analyses of Zuboff suggest that this type of concept can still be useful in interpreting the workings of contemporary societies. Much have changed. Human workers are no longer typically mere extensions of machines at assembly lines. At first glance, it appears as if individuals are freer than ever to express their individual identities. But, at least in some important contexts, contemporary populations seem to be undergoing a form of behavioural modification or behavioural automation of a different order. Information technologies such as cell phones and e-mail, and now smart phones and social media, indeed invite users to engage in behaviours that can be interpreted as genuine and personal, and not merely functional. But we have learnt, from Zuboff and others, that underlying structures and dynamics often are significantly different from perceptions perceived by users who interact with interfaces. If contemporary societies are still understood as individualistic, then the individualism in question should perhaps be represented as a watered-down form of individualism, one that approximates a form of atomism, or, to use Bauman's terminology, 'liquidity'.[131] In addition to the analyses of Zuboff there are many other analyses that suggest that the individualistic diversity that supposedly exists in algorithm-intense environments is in fact shallower than it seems.[132]

The typical contemporary individual, understood as liquid and atomised, does not quite fit under the category of mass man, an average product of the society as a whole. In contemporary contexts it is nevertheless possible to discern a pattern of multiple average products, average products of aggregates of

[131] Atoms are originally conceived of as clearly identifiable separate entities. Although separate, they cannot be imagined as idiosyncratic. They are similar and, as similar, interchangeable. Like the liquid property of water, *liquid* individualism really represents nothing but an illusion of individualism.

[132] See, for instance, Sherry Turkle (2011), Nicholas Carr (2011), Paul Roberts (2015), and Susan Greenfield (2015).

people, that are subject to predictive analysis and various forms of behavioural modification techniques. If predictability and conformity were required of workers at industrial conveyor belts, then, based on what we have learnt from previous analyses, something similar could hold for contemporary human agents vis-à-vis machinery embedded in human environments. Algorithmic machinery will in many instances be efficient to the extent that it is able to apply accurate predictive models of its environment; to the extent that its environment is composed of human beings, there is strong incentive to make human beings more predictable. The dynamics considered in part II can be understood to increase the predictability of people by means of producing conformity not in the population as a whole but in the multiple sub-populations that are aggregated in various algorithmics contexts.

Possible complications in algorithm-intense environments

According to the analyses presented above, it is reasoned that actual human experience, in late modern conditions, appears to be distancing itself increasingly from that which is suggested by modern ideals. If this is indeed the case, then the mythology of modernity may begin to be experienced increasingly as a fictional narrative (a *myth* in the more popular and pejorative sense) that has nothing to do with life as it is actually lived. This process, which, it seems, began before current iterations of algorithmic technologies, can be expected to accelerate in the near future. The themes discussed in part II could be interpreted as signs that we are entering borderline conditions – conditions that may continue to evolve towards something entirely new, something simply incompatible with modern narratives and modern worldviews.

We have been presented with at least two radically different conditions that both, in some sense, could be labelled as modern. We first came across the cog-in-the-wheel condition as characterised by Illich, Ellul and Bauman. We then familiarised ourselves with modernity as a realm for autonomous, mature, and enlightened individuals. If we posit a modern mythology-worldview complex along the lines of the narratives proposed here, then we can hypothesise that individuals inhabiting it should want to create life conditions that enable enlightened individuals to become responsible, to take charge of their affairs, and to flourish *qua* autonomous thinkers and decision-makers. Instead, it seems, we often get the cog-in-the-wheel conditions, which could be interpreted as the antithesis of important modern ideals. What if modernity as a whole – its material conditions, its intellectual ideals, its structural narratives – is lived out in creative tension between these opposite poles? The cog-in-the-wheel condition, as long as the ideal is alive, beckons modernity's subject to pay renewed attention to the ideal. As long as this is the case, we could speak of borderline conditions. When the ideals no longer make any sense, then some more profoundly disruptive events might be under way.

The inferred mythological narrative can be used to explain why, contradictory views on other matters notwithstanding, a certain accord with certain types of ambition and pursuit is still manifest in contemporary technological societies. Even if we find Ellul's analysis plausible, and are convinced that changes in technological life-milieus actually follow their own autonomous or semi-autonomous evolutionary processes relatively independently from what humans might want or desire, technological changes are always marketed to us as if they would enable our becoming more in control, more in charge of our own affairs. *The further advancement of technoscience will be good for us.* In any case, *opportunities outweigh risks*. In other epochs, other types of assertion must have had equal force, such as *the proper maintenance of holy rites will be good for us*. Technophile discourses intended to persuade are not composed of gimmicks invented on the spot; rather, they are instances of an echo that resonates through a millennial mythology. It is from this deep source that they derive their persuasive force. Although there are always people who question the wisdom of salvation by means of technoscience, their views remain marginal. The arguments that align with the epoch-defining mythology usually win the day – at least, until they cease to do so, which would in and of itself be a sign that something has gone amiss in the mythology-world/reality relationship.

In what, then, does the accord allegedly produced by this mythology consist? The mythology locates its inhabitants and the inhabitants' various worldview-beliefs in a structure replete with narrative purpose and meaning; it frames its inhabitants to take a stand vis-à-vis a vaguely suggested human destiny, mission, or purpose. The human purpose or mission that is suggested, in its vaguest form, is that we are to engage in the pursuit of a form of happiness or the betterment of our human condition. The means to this end are also suggested. It is to be understood that the end is gained by means of empowerment over natural and parochial limitations, and that that empowerment is achieved by means of reason, science, and technology. The mythology calls upon its inhabitants to assent to and become instruments of human progress; but it also dares them to rebel and to become exceptions that prove the rule. *The accord produced consists, then, in the elevation of individual reason, a de-traditionalised and universalistic science-informed culture, and advanced technology as the primary enablers of human progress, and the identification of superstition and archaic or traditional pre-scientific culture as the primary impediments to human progress.*

Even as these supposed narrative structures continue to provide the motivation for technological innovation and transformation even further beyond the horizons of contemporary AI, they are at the same time brought into question by the very reconstructed technological life-milieus that they nudge us to produce. This, at least, will be the case to the extent that the actual *life conditions* being produced diverge more and more from anything that could be labelled *modern condition*. It is argued below that, as we move on, the modern

understanding of who we are, what we are capable of, and what we ought to do is likely to come under increasing pressure.

Let us revisit Guardini's consideration of the problem of power. Guardini's predicament is a result of humankind having extensively mastered the immediate forces of nature while not having sufficiently mastered the mediate forces by means of which the immediate forces are mastered. C. S. Lewis, in *The Abolition of Man*, provides a different angle on the same predicament.

> Let us consider three typical examples: the aeroplane, the wireless, and the contraceptive. In a civilized community, in peace-time, anyone who can pay for them may use these things. But it cannot strictly be said that when he does so he is exercising his own proper or individual power over Nature. If I pay you to carry me, I am not therefore myself a strong man. Any or all of the three things I have mentioned can be withheld from some men by other men […] What we call Man's power is, in reality, a power possessed by some men which they may, or may not, allow other men to profit by. (Lewis, 2006, p. 54)

This paragraph highlights a key predicament of the 20th century consumer. In the days ahead, things are likely to be even more problematic. As we enter an age of increasingly complex hybrid multi-agent systems and organisations, we must enquire about the extent to which power is really held by human beings at all. More and more, power inheres in complex and opaque hybrid multi-agent structures. Can power, then, be withheld without the interference of any human being or any group of human beings? This scenario is already somewhat familiar from legal and bureaucratic contexts. If, then, power were to inhere increasingly in structures that were much less transparent than present-day legal codes and bureaucracies, then the magnitude of an already profound problem, it seems, would widen. If individual power indeed implies control over one's own life and one's environment, we – from our various points of view – must consider and weigh the extent to which, by means of the technologies we consent to use, we are in fact increasing or undermining our power. Again, that which may seem to increase our power in the short term might disempower us in the long term.

The economy critiqued by Zuboff situates most human agents in contexts that appear to be very far from any imaginable modern ideal of individual autonomy. Under constant surveillance, we are manipulated by opaque agentic structures. Our behaviour is instrumentalised in the service of ends that we ignore. One could object that this depends on how we look on the matter. Many will be of the view that the minority who own and/or control key algorithmic infrastructure are in positions in which they are free to exercise power not only over their own lives, but over a very large extent of their environments. But even people at the lower ends of technical hierarchies may have the *impression* of freely exercising their individual will. Yes, they may well be under surveillance and be manipulated on one level of analysis; but on the level of their own experienced life conditions, they may still be, or at least

have the impression of being, free and autonomous. This possibly represents a bifurcation of life conditions as experienced by humans at different levels of technical hierarchies: one in which humans at the higher ends of technical hierarchies are increasingly able to exercise more and more power on both phenomenological and comprehensive realist levels of analysis, whereas the scope for expressing agency at the lower ends of technical hierarchies, on realist levels of analysis, becomes increasingly restricted.

At the same time there is a general awareness that we can no longer trust structures that are critical to our ways of life, that our limited cognitive capacities are not quite up to the task of comprehending complex and opaque critical algorithmic structures, and that, although we may feel perfectly free in our online behaviour, we are really inhabiting structures that have been deliberately designed by specialists to elicit responses and behaviours to which we may not ordinarily consent. It is still possible to have the impression of inhabiting a modern condition; however, from the point of view of the mythology of modernity, it is equally possible to interpret these new conditions as something utterly alien and anti-modern. This ambiguity is a sign of a borderline condition. The posited bifurcation of experienced life conditions represents a peculiar power asymmetry. We are now facing scenarios that used to belong to science fiction. Increasingly humans can be immersed in life-worlds that have been engineered for them by others – other humans and/or AIs – life-worlds that, in various respects, may be experienced as more pleasant and even more ideal for the exercise of human autonomy than conventional non-digital spaces. But such life-worlds are the product of the deliberate use of manipulation techniques. They offer possibilities for experiences that, in various respects, could be perceived as *better than reality*. Immersion in this kind of life-world in fact removes us from *life*.

The structures analysed by Zuboff, in which opaque forces engage in continual manipulation of human subjects, instrumentalising their cognition for the benefit of ends that are alien to the individuals so instrumentalised, could strike us as exceptionally wicked. Admittedly, far from all applications of algorithmic technologies have such nefarious effects. It is nevertheless the case that algorithmic structures impinge more and more on the social and personal spheres of human beings. If this interference continues to play out in accordance with contemporary prevalent patterns, then it will become more and more implausible to conceive of ourselves as individually autonomous and/or in charge of our own individual lives. At least, this will appear to be the case for those of us who are relegated to being controlled by rather than being in control of critical technical infrastructure. The few who are situated in the upper echelons of technical hierarchies may instead perceive themselves to be supremely in control.[133]

[133] To the extent that we take the analyses conducted in part II into account, the extent to which 'they' will be in control in any meaningful sense is questionable.

Future forms of AI may of course turn out to be very different from the contemporary structures critiqued by Zuboff. The type of partnership that Kissinger *et al*. envisage represents a potential way forward in which the human partner may not only retain a degree of control over things, but also, possibly, experience an extension of knowledge and know-how. But even if such partnerships developed in accordance with the hopes of these authors, the privileged position of the human agent would likely be put into question. Would the experience of the human partner be one of being equal with the artificial partner? 'Equal' is here used in the very basic sense of possessing similar cognitive qualities and of appearing to be able, at least in theory, to achieve similar degrees of knowledge and know-how. It should be obvious that this is unlikely to be the case as long as human nature remains intact and the performance of algorithmic systems continues to increase. Moreover, Kissinger *et al*. envisage not only the increased performance of contemporary algorithmic systems, but something like artificial general intelligence or superintelligence, in which case there would hardly be any semblance of equality between human and artificial agents. Difference in capabilities does not entail an inferior position of the human partner. Complementarity, it could be argued, constitutes the very glue that makes partnerships meaningful and useful. But meaningful complementarity will manifest only as long as the human partners possess some important quality or qualities that the artificial partners do not. Today, humans alone can consider which overarching goals ought to be pursued. Humans can also, uniquely, reconsider the pursuit-worthiness of existing overarching goals.[134] Even if human partners were to continue to possess unique qualities, it would be possible, depending on how artificial partners were constituted, that human experience would increasingly become one of being inferior to the excellence displayed by artificial partners. Embedded in increasingly artificial hybrid multi-agent structures, in which the capability and power of artificial agents continued to reach new heights, humans would likely perceive non-human agents as superior.

Based on the totality of the analyses presented so far, there is a non-ideology-, non-power-, and non-greed-related sense in which algorithmic systems applied to social environments *must* know a great deal about humans and *must* manipulate humans. Any advanced agent – including animal agents – must, in order to be functional in its environment, have some equivalent of a self- and world-model that is adequate in predicting the self's interaction with its environment. If the environment of artificial agents is a complex social environment that, to an important extent, is composed of complex and interacting human agents, then it seems to follow that, quite regardless of any incitements related to surveillance capitalist objectives, we should expect many of the

[134] One could, of course, make the argument that humans have no free will and, therefore, that they inhabit a condition more akin to the condition of algorithmic systems than it would appear to the common-sense observer.

phenomena that Zuboff criticises and attributes to surveillance capitalist ideological causes. While some of the phenomena that she observes may well be mitigated if certain ideological patterns are reduced, the disruptive and revolutionary potential of the algorithmic structures she observes transcends her ideological categories. The merger of the social realm of humans and the algorithmic realm of machines is likely, therefore, to produce something that can have a positive or a negative spin: the positive spin is articulated by Edward Lee in terms of obligate symbiosis and co-evolution (Lee, 2019); the negative spin has been articulated by Ellul and Illich as the loss of previously valued experiential aspects of life, such as human autonomy. While the labels evoke different emotions and future expectations, one underlying understanding is similar: the days of the self-reliant and autonomous human agent appear to be over.

When we contemplate contemporary changes and imagine future scenarios, will it still make sense to interpret the world through modern worldviews? Not without further qualifications, some of which are considered in the next section. At least some category in the worldview-human-environment complex, it is argued, must adapt in order to preserve the heuristic viability of the whole.

Adaptive measures

We now have available a model for a generic worldview animated by a generic modern mythological narrative. Any modern worldview, it has been argued, implies a generic modern philosophical anthropology. We have considered some key factors in the relationship between worldviews and environments, a relationship in which humans fulfil the function of being acting agents. We are now in a better position to evaluate how the embedding of human agents in algorithm-intense environments is likely to challenge modern conceptions in general. In this concluding section, we specifically consider how the modern philosophical anthropology is likely to fit into algorithm-intense environments and, assuming that there is friction, the potential effects on meaning-making, on the philosophical anthropology itself, on the physical human agent, and on the ongoing technical reconstruction of human environments.

First, we consider 1) the extent to which human embedding in algorithm-intense environments is likely to challenge modern philosophical anthropologies. Then, assuming that modern philosophical anthropologies are in fact likely to be challenged, at least to some extent, we explore some of the likely consequences, namely, 2) meaning crisis, which in turn, it is argued, is likely to incentivise one or several of the following adaptations: 3) of the philosophical anthropology, and/or 4) of the physical human agent, and/or 5) of the pace and intensity of environmental reconstructions.

(1) In anticipating the effects of the implementation of new technologies, stakeholders with vested interests are prone to emphasise human empowerment. There is no doubt that more advanced current and future forms of

algorithmic technologies must be understood not as small but rather as major empowerments. If such empowerment could be generalised as socially unproblematic, then the modern philosophical anthropology would no doubt be boosted rather than challenged, since the modern philosophical anthropology extols the ideal of human empowerment.

Suppose that I, the author, subcontracted the completion of this final chapter to my favourite generative AI: it would immediately empower me to do other things during the time it would take me to write it in the conventional way. From my point of view, this course of action could indeed represent a human–AI-partnership in which I, the human, am in charge. Algorithm-intense environments, then, could imply a mere extension of the previous patterns explored in this treatise – a cyber-modernity, in which the modern philosophical anthropology is still relevant. However, throughout this treatise appearances have been questioned from various angles. We should, of course, continue to ask: Would many individual instances of practising such human–AI partnerships translate into some general social goods? Would short-term conveniences translate into long-term empowerment?

If we accept Guardini's arguments, we must also recognise that such questions are very difficult to answer. We have not yet devised any reliable method to wield power over power. Mere empowerment in the abstract means nothing. In part II we saw how the empowerment of some agents or of some multi-agent organisations would often imply the disempowerment or even the obsolescence of other agents or multi-agent organisations. In any multi-agent organisation, the empowerment of agents at some level of the technical hierarchy may also imply the disempowerment, subservience, or obsolescence of agents at lower ends of the technical hierarchy.

The very name of the science pioneered by Norbert Wiener points to the focal point of the predicament in which modernity finds itself. Wiener referred to the science of control and communications as 'cybernetics', which he derived from the Greek word *kubernétes*, meaning 'steersman'[135] (Wiener, 1988, p. 15). The image of steersman evokes the idea of a controller of a complex vessel. The exercise of control over complex systems – nature itself and

[135] *Encyclopaedia Britannica* introduces cybernetics as 'control theory as it is applied to complex systems'. In the technical work *Cybernetics: or Control and Communication in the Animal and the Machine*, Wiener, according to the editors of *Encyclopaedia Britannica*, defines cybernetics simply as 'the science of control and communications in the animal and the machine' (The Editors of Encyclopaedia Britannica, 1998). The section in part I concerned with the problems raised by Wiener provides good examples of questions that could arise in the context of control theory. Cybernetics, as Wiener defines it, is concerned at once with communication and control: 'It is the thesis of this book that society can only be understood through a study of the messages and the communication facilities which belong to it; and that in the future development of these messages and communication facilities, messages between man and machines, between machines and man, and between machine and machine, are destined to play an ever-increasing part' (Wiener, 1988, p. 16). This gives an idea of the intellectual context in which the timing problem, the goal-intention problem, and the mechanical slave problem were explored.

artefacts invented by humans – can be understood as one of the core endeavours of the modern adventure. The experience of being a *kubernétes*, therefore, is related to the experience of being an ideal modern agent. The answer to the extent to which the modern philosophical anthropology will be challenged at the later stages of modernity, therefore, is tied to that which – *who* or *what* – is perceived or understood to be in the position of steersman. Who or What is *kubernétes*?

We have considered some forms of complex algorithmic systems and interpreted them in terms of multi-agent systems and multi-agent organisations, and we have considered the consequences in respect of short-, medium-, and long-term effects. We have, of course, not considered all forms of existing complex algorithmic systems, and we cannot know how future iterations of algorithmic systems will be constituted and applied. It is certainly possible to envisage decentralising developments that could empower individual users in the long term at all levels of society in meaningful ways. Even if things kept evolving in accordance with the patterns described by Zuboff, we must recognise that the lower ends of technical hierarchies might still afford the experience of empowerment and liberty. By applying algorithmic systems, new realities could be engineered, custom-made for living out individual urges. Compared with a more and more cumbersome reality outside the premises of such engineered realities, new *algorithmic worlds* might be experienced as more compatible with modern worldviews and with the modern philosophical anthropology than the plain old world.

Here we may want to distinguish between an encompassing ontological context and phenomena that are being experienced, and argue, as previously, that short-term empowerment, convenience, or pleasure experienced in engineered algorithmic realities are being experienced at the cost of potential medium- and long-term disempowerments and miseries that will undoubtedly turn out to be considerable. Here the 'encompassing ontological context' stands for the underlying unseen structures that afford given phenomenological experiences at the level of computer interfaces. But the relationship between such ontological contexts and the phenomena experienced is not necessarily similar to how the relationship between ontology and the experienced world tended to be understood in previous times: in algorithmic instances, that which is (ontology) can be devised to the effect that that which is experienced (phenomenology) serves the interests of those who are in the process of constructing and reconstructing that which is (ontology). Given that humans or, sometimes, human 'sub-creations' (algorithmic systems) are in charge, that which *is* can no doubt also be poorly devised, to the effect that that which is experienced may serve no worthwhile long-term ends at all. In both of these cases, human engineers could be understood as embodying the role of the

demiurge – the imperfect god-creator of flawed worlds for human souls to inhabit.[136]

However, even if we were to face this extremely dystopic scenario, the modern philosophical anthropology would not necessarily be directly challenged, for humans at all levels of technical hierarchies could, theoretically, on the level of interfaces, experience that they were *kubernétes*. Human agents at the higher ends of technical hierarchies might then be understood to be ontologically and phenomenologically in control, whereas human agents at the lower ends could be understood as having only the phenomenological experience of being in control; but in order to grasp the difference, those who have merely the phenomenological experience of being in control would have to break out of the Matrix and attain a higher understanding.

Ongoing improvements of virtual immersive realities notwithstanding, there are reasons to doubt the persuasive power of phenomenological experience in engineered realities. Dystopic popular culture provides evidence that there is at least a vague popular awareness that things are not quite as they seem behind the glittery surfaces of innovations marketed to the masses. And in regard to algorithmically engineered realities, there is a more or less vague awareness that they tend to be deceptive to various extents. They are not yet experienced as real in the same sense as the old world is experienced as real.

The above also applies to how I choose to view any generative AI that could supposedly write this treatise for me. Not long ago, one could often hear that the future belongs to those who know programming. One still hears it, by the way. Yet we are increasingly presented with complex programmed algorithmic systems that can be used, with ease, without *any* knowledge of programming. The hypothetical generative-AI-affordance of finishing this chapter for me presents me with a temptation. Yet I am aware that, if I myself am not able to understand and, by means of programming, to adapt the algorithmic 'tool' that I use for my own purposes, I am, in a sense, being programmed by someone or something else. I am also aware that, by using generative AI, I may very well contribute to the potential elimination of all social needs for my own skill sets. No, if I choose to 'use' generative AI, I am not choosing to use a 'tool' for some given purposes; I am choosing to interact with new and uncharted algorithmic realities – realities replete with opaque and unfamiliar agencies. This, at least, is my articulated experience of thinking about 'using' generative AI: ignorant as I am of programming, it would be dis-empowering and would undermine the modern philosophical anthropology, for I would

[136] In Jewish, Christian, and Islamic monotheistic theology, that which *is* (ontology) is typically divided into the transcendent, timeless, and infinite realm of God on the one hand, and on the other that which is created by God and henceforth exists in time and space. In the contexts under discussion, that which creates and re-creates in time and space is itself contingent in time and space. The human creator is more to be pitied than the gnostic demiurge (an imperfect god-creator), for, inextricably bound up in time and space, the human architect must sooner or later suffer the consequences of his or her own 'creations'.

sense the presence of an alien *kubernétes*.[137] We could classify this example under Wiener's mechanical slave problem.

On the other hand, humans who know programming and who are in positions where they can meaningfully adapt the algorithmic systems that they use for their own purposes – it would then be appropriate to understand the algorithmic system more as a 'tool' – would be in a different 'ontological' situation. Here, the ideal of a human *kubernétes* might resonate with both phenomenological experience and the ontological condition. The experience of self in an algorithm-intense environment would then likely be more compatible with the ideal of the modern philosophical anthropology.[138] That much admitted, in a really algorithm-intense environment, the program-literate person would be unlikely to be able meaningfully to adapt all, most, or even a fraction of the totality of algorithmic systems that happened to be in the process of affecting that person's life. In possession of a more in-depth knowledge of the workings of the constituents of algorithm-intense environments, such a person would nevertheless, in comparison with typically modern conditions, be in a radically different condition – a condition in which powerful transpersonal multi-agent systems and organisations, the majority of which would be beyond the ability of an individual programmer to control or even understand, proliferated and intervened in human affairs. Even to the initiated who are more versed in the lore of computer science than the uninitiated, highly algorithm-intense environments would still represent a life-milieu based and centred less on individual human agents and more on a novel and largely impenetrable kind of transpersonal agent.

The modern philosophical anthropology implies an ambition at least to strive towards the position of being *kubernétes* – of our own bodies, to be sure, but also of the bodies that govern the affairs of our life-milieus. The latter positions can only be held, if at all, by a minority of individuals. Modernity has tended to solve the problem of selecting those who are to hold them by means of meritocracy. As for important positions where most individuals of a population cannot themselves be *kubernétes*, they want the positions to be held by their most competent and trustworthy representatives. If we accept the analysis of Susskind as plausible, then, in future iterations of algorithm-intense

[137] Suppose that I do it nonetheless: what would the treatise represent to you who read it? What do generative AIs generate? Knowledge? Valuable insights? Novel perspectives? Information? If researchers were to engage customarily in the practice of generating research output by means of generative AI, the collective output would likely be apprehended, at least by humans, as a confusing information glut rather than as something that could be meaningfully integrated into a body of understanding and knowledge.

[138] To refer to algorithmic systems as 'tools' may seem to run counter to the spirit of this treatise. It must be recognised that algorithmic systems can fill many functions in human contexts. Some serve very narrow and direct purposes in the context of human practices. An algorithmic system used for diagnostics of diseases, for instance, may appropriately be referred to as 'tool'; even more so if the users of the system participate in the programming and adaptation of the system for their purposes.

environments it is likely that most if not all important work-positions will be better suited to artificial agents than to human agents. But any condition in which humans experience themselves to be subjected to impenetrable animistic-like forces is, at first glance, antithetical to modern worldviews. If, moreover, we agree with the analysis presented in parts II and III, then any engagement with algorithmic systems on the part of the 'uninitiated', even though such engagement usually affords short-term conveniences and benefits, risks leading to long-term dependence and, arguably, to subjugation to alien transpersonal agencies. If this were indeed to be the case, then the ambition to become *kubernétes* in algorithm-intense environments must imply at the very least an ambition to become initiated into the lore that governs those environments – that is, an ambition to learn and master the art of programming.

Future patterns of human–algorithmic dynamics could play out in different ways. If we envisage *cybernetic humans* that are to use complex algorithmic systems as *tools*, then the cognitive requirements on humans, as Wiener argued, will increase rather than decrease. The same arguably holds for human–AI partnerships, if the aim is that the human agent ought to occupy the position of *kubernétes*. On the other hand, *it is likely that the cognitive requirements for being able to interact with algorithmic systems capable of emulating human language and behaviour will continue to decrease*. This would have potential important repercussions not only for work-contexts but also for other types of context in which, for instance, novel kinds of partnerships may be formed in order to sooth experiences of loneliness, meaninglessness, and lack of purpose. Any relational dynamic that affords a decrease of human cognitive requirements while the cognitive capabilities of the algorithmic partner increase implies a surrender on the part of the human agent *qua kubernétes*.

The problem, it seems, boils down to who or what is to become an extension of who or what: a human-centred algorithm-intense society will require that humans learn programming, and, probably, much more than programming – it may in fact require a population of cognitive over-achievers and geniuses modelled on the multi-competent Norbert Wiener himself. Potentially, we are about to witness a bifurcation of human experience, subsequent to which one part of the population would still, in some sense, be able to strive meaningfully for the ideal of the modern philosophical anthropology, albeit in rapidly changing environments, but in which the rest would, if at all, only be able to experience the ideal as attainable within the confines of deceptive algorithmically engineered environments. For the first category, the ambition to meet high cognitive requirements is likely to be achieved at a cost: if the level of expert programming is attained, the achievement is likely to come at the expense of a general width of knowledge and culture. (Here and there, exceptional individuals, such as Wiener, will no doubt be able to perform at high levels in multiple and varied disciplines.) The philosophical anthropology of modernity will be challenged for both.

(2) We can posit the onset of a crisis of meaning, or of purpose, if and when the heuristic value of a worldview begins to decrease. We have now become familiar with a context in which the ideal philosophical conception of the human agent appears to be far removed from how humans will actually experience themselves in their environments. A discrepancy between ideal and experience is not in itself a crisis trigger. Such a discrepancy must be understood to be part of normal and borderline conditions, in which the ideal serves to incentivise humans to accomplish ideal-related goals. Any 'natural' discrepancy between ideal and experienced self is now being amplified by the addition to the human ecology of increasingly powerful transpersonal agents that are perceived increasingly and vastly to exceed the limits of human excellence in more and more domains. Even if phenomena may still be experienced to be compatible with the modern philosophical anthropology, any awareness of ongoing ontological reconstructions is likely to spur us to reconsider the role of human agents in algorithm-intense environments. Hence the crisis.

Here the crisis is considered simply as a transitory phase, one that incentivises adaptive measures. Some possible adaptive measures that could be engaged as a consequence of the crisis are now considered. These adaptive measures should not be understood as mutually exclusive. They were distinguished earlier by the conjunction 'and/or' in order to emphasise the unlikeliness that any *one* potential measure would be fully realised to the exclusion of other potential measures, including measures ignored by the present author.

(3) One measure concerns the adaptation of worldviews to new circumstances. Here we consider the adaptation of the modern philosophical anthropology: assuming that it will be adapted, how is it likely to be adapted in order to be more in tune with algorithm-intense environments?

The transition from a porous to a buffered self, it has been argued, is at the root of the development of the modern philosophical anthropology. The modern self-understanding locates all consequential agency, at least in this world, with the human agent. Animal agencies are still operative, but by practising scientific inquiries, in time human agency will penetrate all the remaining mysteries of the world. The human environment will, ideally, become transparent and subject to human control. Cybernetics is concerned with how such control can be exercised in increasingly complex and automated environments.

We can now expect that humans will be increasingly embedded in environments alongside other agentic structures that, to varying degrees, will be experienced as of an ambiguous ontological nature. Algorithmic systems expressing agency are similar to animal agents in that they can be studied and explained by trained scientists. They also have attributes that differentiate them from all the animal agents to which humans are accustomed. The agency of algorithmic systems will often be perceived to elude or transcend, if not the physical world, then at least clearly identifiable physical entities. In many instances where algorithmic agency is brought to bear, the agency will not be

perceived to inhabit any identifiable physical body. In instances in which it may appear that algorithmic agency is incarnated in a physical body, such as in social robots, it will still be unclear to the human perceiver whether or not that which is perceived as agentic is actually 'located' *in* the perceived 'body'. As long as we are even vaguely aware of some interconnected systemic matrix governing the function of algorithmic systems, any anthropomorphic algorithmic embodiment, such as the fictional speaker from the introduction to this treatise, must be perceived as potentially interconnected with and/or under the control of remote and anonymous agencies. Old questions will begin increasingly to be asked anew: who or what is governing the working of things?

Algorithmic agentic systems are also different from the animal agents with which we are familiar, in that algorithmic systems are continually being launched in newer configurations. Throughout his or her life-span, a scientist studying any mammal species will have a relatively unchanging object of study. Evolution will not have time to work its magic on the object of study. Adaptive processes will be more rapid if we move down to the level of microorganisms, bacteria, and viruses. The relationship of the expert to the area of expertise is different in the two cases. In the fields of engineering and computer science, the experts are themselves conceiving, constructing, and adapting the objects being studied. Technical innovation today often manifests as sudden and disruptive, so that, even among trained experts, technical knowledge and know-how tend to be limited to ever narrower fields of expertise and to come with shorter and shorter expiration dates. Algorithm-intense environments as wholes, in consequence, are likely to be experienced as opaque and impenetrable, even by trained experts.

The very relationship between human agents and algorithmic agentic systems, finally, is different from the relationship between typically modern human and animal agents. The modern human agent is accustomed to know and control animal agents, if not personally then at least as a member of a modern culture, so that animal agents are used for human purposes. We have seen that the function of algorithmic systems, especially in the larger contexts of hybrid multi-agent organisations, can be ambiguous. Even if we can sometimes use them for our own purposes, few of us can know their nature in any meaningful way. We are aware that algorithmic systems, including ones with which we voluntarily choose to interact, are used to know and control us in various ways, for purposes that we ignore or about which we can only speculate.

If indeed it becomes increasingly the case that algorithmic systems will learn to know more and more about us, whereas we, or at least most of us, will know less and less about them; that they will increasingly affect the thoughts we think and how we act, whereas we, or at least most of us, will have little power to affect them; and that the agencies expressed by means of algorithmic systems will be perceived to transcend the visible discretely categorised physical world, it will seem quite odd to conceive of the human agent as a buffered self. The environments that are being constructed, it seems, have more in

common with how the moderns imagine that animistic cultures viewed their environments. Therefore, one possible adaptation of the modern philosophical anthropology is to revert to a porous self. One adaptive gain would be to reconstitute a philosophical anthropology that could be experienced as more in tune with the experienced world.

Whereas this would have the effect of profoundly altering modern worldviews, it need not altogether undermine the human *kubernétes* ambition. Ancient pre-modern tribes, supposedly, also controlled their environments to some extent. The methods practised in order to exercise control were typically different from those practised by modern agents. In algorithm-intense environments, humans may need to learn how to placate agencies that are experienced as opaque and alien. For many, this may imply learning 'incantations' that work rather than aiming for something like knowledge of the environment. We may see something analogous to how rite and magic function in premodern tribal contexts manifest anew. Today, for instance, the lay person who learns to work the new environment can be superficially initiated into the mysteries and become a 'prompt engineer', understanding nothing about the inner workings of the multi-agent system-based transpersonal agencies with which he or she interacts, but nonetheless managing, by means of procedures – that is, incantations – that have successful track records, to influence those opaque forces to do his or her bidding. The master programmer, on the other hand, who is initiated into the deeper workings of things, analogously embodies the role of the shaman.

For many, and not least for those still identifying with something like a modern ethos, such changes could be perceived as unfortunate regressions to earlier stages of human development. There is, however, an inherent ambiguity to the mythological modern narrative. A critic of modernity may embrace the same narrative merely by positing, as would many of the ancients, that the modern ideal of becoming a controller of matter in time and space represents hubris rather than virtue. Humans, then, never were the knowers and controllers they imagined themselves to be. That which to the moderns appeared as improvements and developments towards higher states of being were in fact instances of corruption. As evidence in support of such an upside down view of the mythology of modernity, one could cite the almost paradoxical conditions that are in the process of being produced in late modernity: that some of the very structures from which modern science and technology were meant to liberate the modern human subject – namely, supra-human agentic influences deemed to have been the imaginary products of human superstition – are being engineered, by hubristic modern agents, *into* the latest technical systems that are meant to extend human power even further. We are, then, subordinating our human offspring to very real forces that remind us of how we presume that the ancients imagined animistic forces. This, it seems, could be cited as a textbook case of hubris inviting Nemesis.

Whereas this line of adaptation represents one possibility, it must be considered together with other potential adaptive measures. It is far from certain that humans in algorithm-intense environments would tend to abandon the ideal of a buffered self, destined to aim for knowledge and control over matter in time and space. Therefore, let us consider another possibility.

(4) Another potential set of adaptive measures concerns the constitution of the human agent. If the human agent, as it is currently constituted, is deemed to be unfit for the attainment of modern ideals, then the adaptation of its constitution represents one possible way of filling experienced gaps between ideals and reality. This kind of adaptive measure is already undertaken in the guise of experimental brain implants. Nina Bai reports that brain implants have successfully revived cognitive abilities long after traumatic brain injury (Bai, 2023). Meanwhile, the stated mission of Elon Musk's Neuralink is to '[c]reate a generalized brain interface to restore autonomy to those with unmet medical needs today and unlock human potential tomorrow' (Neuralink, 2025).

Whereas brain implants appear to have achieved impressive results in restoring previous levels of performance, transhumanists have many arguments for why it would be a good idea to reconstruct the constitution of human agents so that they become able to perform beyond the limits inherent to in their current constitution.[139] Transhmanist arguments are typically grounded in fundamental and variable attributes, the expression of which we tend already to understand as 'goods', such as health, intelligence, strength, and longevity. If we value these attributes so that we would rather be healthy than unhealthy, more intelligent than less intelligent, stronger than weaker, and live long rather than short lives, then, transhumanists ask, given the technical opportunities, why should we not endeavour to extend these attributes beyond current human limitations?

The analyses presented here suggest, at least to mindsets shaped by modern worldviews and a modern philosophical anthropology, an additional argument in favour of transhumanistic reconstitution. Having engineered highly automated life-environments, being in the process of engineering more and more algorithm-intense environments, and aiming for the even more revolutionary inventions of general artificial intelligence and superintelligence, *human* civilisation is entrenching itself in a dead end. Wiener's analysis surely applies to us, if, that is, *we* – we as individuals relative to the management of our own lives, but also we as collectives relative to the human-run management of public affairs – wish to remain in the driver's seat, if we still hope to be able to reach for the ideal of the human agent *qua kubernétes*. If the aim is to construct intelligences that are far superior to human intelligence, then it stands to reason that humans are likely, in some way, to be subordinated to that which is superior, just as that which has previously been perceived as inferior to human

[139] See, for instance, Bostrom (2001).

superior intelligence have been subordinated to human interests. Reconstruction of the constitution of humans, therefore, could be understood as necessary merely to preserve, in a sense, the status quo: it becomes a question, in other words, of evading existential risk and maintaining the status of humans *qua kubernétes* in radically altered life-environments. Within the figure of the dead end, transhumanist reconstruction theoretically affords a breach of the frontiers of the dead end, enabling future transhumans and post-humans to carry on the modern adventure rather than stopping short or turning back.

In order to be able to maintain a claim to the *kubernétes* status in increasingly algorithm-intense environments where, in more and more domains, algorithmic systems outperform humans, and where general artificial intelligence and superintelligence are imagined to loom on the horizon, human agents must be made to assimilate structures that enable them to become more *like* the new types of artificial transpersonal agent with which they vie for control or supremacy. It is interesting that transhumanists who profess to fear the consequences of superintelligence in general show little interest in slowing down AI development. Many, like Nick Bostrom and Ray Kurzweil, appear to advocate an accelerated approach. Could it be that AI development and transhumanist reconstruction go hand in hand? Loud proclamations of existential and apocalyptic danger in connection with a hypothetical emergence of superintelligence could certainly also serve the purpose of supporting transhumanist arguments.

Judaeo-Christian doctrine holds that human beings are created in God's image. It could be argued that AI developers are endeavouring to create intelligence in our own human image. If this is so, then, since we have no ultimate oversight or control over the potential end results of such endeavour, it could also be argued that humans must recreate themselves in accordance with their own re-created image, and so on, ad infinitum. Instead of venerating a symbol-mediated entity apprehended as perfect, timeless, and unchanging, we would then, in a sense, be venerating contingent and ever-changing human-made structures that at one and the same time served practical or performative functions *and* the symbolic function of an analogous model – read 'icon' – for how we humans *ought* to be.[140]

Given adequate modifications[141] of the human constitution, the viability of the modern ideal human agent could still be preserved in something that resembles a modern worldview. If the ideal philosophical anthropology remained more or less intact, then other worldview categories, such as cosmology and axiology, could be adapted to better fit the evolution of the

[140] This paves the way for a new research project, one that could build further on work that has already been done on technology as metaphor by Barbara Adam (1990), Philip Rieff (2006), Jaron Lanier (2011), and others.

[141] Transhumanists generally refer to modifications as 'improvements', but it is not at all certain that even modifications that would have the effect of improving the performative capacity in various domains would represent an improvement of the human constitution in a general sense.

mythological narrative towards its next post-human level. Even today, thinkers such as Ray Kurzweil view humans as but a step in the evolution of intelligent life. On his view, the axiological purpose, or the heroic mission, of humans now is to accelerate the evolution of intelligent life and to bring into being configurations of intelligence that move beyond human limitations (Kurzweil, 2000). Thus another 'turn' could be added to the mythological narrative of modernity, representing a continuation of the ascent towards a post-human epoch.

An accelerated approach to AI, then, goes hand in hand with transhumanist aspirations. For it is assumed that an AI-intense future would not necessarily define contexts in which human beings as currently constituted would be safe and comfortable. In order to maintain the plausibility of the modern ideal in such a future, it becomes incumbent on us at least to endeavour to increase the cognitive abilities of human agents. Possible alternatives in the apparent dead end of *human* civilisation include developments that could vary in intensity: from the relatively mild neo-animistic adaptations of the philosophical anthropology considered previously, to subjugation to an environment experienced as quasi-automated and controlled by alien forces, to even more extreme developments akin to the scenario of *The Terminator*. But there are also, no doubt, many other adaptive measures that could be engaged. Here we consider one more.

(5) If one drives into a dead-end street, the usual manoeuvre is to turn around and drive back. The dead end was perhaps not a fitting analogy, since adaptive measures that may afford further travels beyond the dead end have been discussed, which, in turn, suggests that there really is no dead end. Moreover, as a solution to the predicament under discussion, the manoeuvre usually undertaken in dead ends appears to be very unlikely to be undertaken under current circumstances. Has modern civilisation ever opted to regress to previous technological stages of development? An alternative and less radical measure would be to slow down the speed of technical development and the application of novel technologies. This too may seem unlikely given current circumstances; but key circumstances might change, and we should perhaps not be too surprised if they did.

The example of the Amish has been used previously in order to illustrate a currently practised way of life that limits the dominion of technique. Limitations are imposed so as not to undermine the overarching social concerns of the Amish. Of course, all cultures do this in some sense. Legislation is routinely used to limit the application of technologies in order to prevent physical, psychological, and social danger. But if technological societies, like all societies, seek to avoid dangers to the extent that it is possible, they typically lack any clear conception of an overarching positive good that effectively limits scientific research and technological development. It could even be argued, as Ellul in effect argues, that scientific research and technological innovation often seem to represent the overarching goods to which many of the other areas

of human practice must be subordinated. The Amish have at least one overarching practical concern, something that they understand as a positive good, and that they value higher than technological innovation: the fostering of community.

At present, in technological civilisation, one easily gets the impression that there is no overarching good that limits innovation and the application of algorithmic technologies. To the extent that research and application are deliberately limited, it appears to be limited by concerns about potential danger and harm. As long as no harm is caused, technological civilisation reasons, the acceleration of innovation and implementation must be a good thing. Wiener's concerns – that important harms may be discovered when it is too late, and that technology should develop in a pace that affords an orderly adaptation of human knowledge and practice – if they are taken into account at all are certainly not driving events. At present, the analysis of Ellul appears to apply: societies that vie for an elevated status among the technically evolved seek to mobilise social and material resources in order to enable and facilitate innovation of the latest technological cutting-edge thing: AI.

However, could not many (or even most) societies be understood to have some positive good that limits the application of technologies? Surely it is too simplistic to assume that danger and harm in technically advanced societies are only considered relative to physical and mental well-being. Harm can be caused to state authority and stability; in societies that identify with a religious tradition, harm can be caused to the religious tradition; and, in societies that identify as democracies, harm can be caused to democratic processes and to whatever fits in the category of 'democratic values'. This is surely the case. If this is so, then science and technique cannot be understood to be uncontested overarching values. Still, in technically advanced societies, democratic values, unlike the value of community translated into practice at local levels in Amish societies, represent something abstract and often contentious: it is not always clear what they are; different factions of a democratic society often argue over what they ought to be; and to the extent that there is a viable consensus, democratic values tend to be lofty and vague, leaving room for many different interpretations. To the extent that technically advanced societies subordinate technique to some positive good, that good will often be so abstract and vague that one rather gets the impression that there is no *meaningful* subordination. Nevertheless, even if we were to accept that the autonomy of technique is stronger than ever, there would no doubt still be room for contention: things that are positively valued in societies but that are not at present perceived to be overarching concerns may rise in the hierarchy of concerns and become overarching. What could such overarching concerns be?

Most likely they would be modelled on things that are valued at present. A change in circumstances could trigger a reprioritisation of concerns. The digital economy and its algorithmic technologies require enormous amounts of electric energy. It is far from inconceivable that future energy shortages might

trigger a re-evaluation. Some societies might then choose to use the energy available to service self-sustaining communities, similar in some respects to the Amish. If we considered this adaptive measure as a means to cope with the challenge to modern ideals, we could imagine such communities as aligning with Illich's convivial modernity.

We should not assume, of course, that energy shortages would inevitably tend to issue in such utopian communal scenarios. If special interest groups that are active today carried on being active during energy shortages, then there would be good reasons to expect that many would argue that AI represented the best hope to solve the energy crisis. AI development could, then, be prioritised above the well-being of demoted consumer-citizens. Some societies could opt to restrict the availability of electricity to human households.

Under current circumstances, deliberate turn-arounds and slowdowns seem unlikely. Circumstances would need first to change so that a reprioritisation of values was triggered. However, remember that all these adaptive measures are meant to represent potential concurrent processes, not exclusively undertaken either/or options. Even today, circumstances frequently change, sometimes so that, for instance, fewer funds are available for research and development. Partial slowdowns, then, occur all the time. Decelerating thrusts in the ongoing processes of AI development should therefore also be expected. A more general and lasting slowdown is still conceivable.

***

There are no doubt many more adaptive measures that could be engaged in in response to challenges to the modern philosophical anthropology. The examples discussed above serve to illustrate how such challenges could be met. They also serve to illustrate how the worldview–world ecology hangs together, and to problematise the key entity of this ecology, namely, the human agent itself, inhabiting both a worldview and a world.

Conclusions Part III

Embedding of humans in algorithm-intense environments will challenge:

1) Modern conceptions of the self.
2) Modern conceptions of the human agent in environments.

This is likely to produce:

3) A meaning crisis.
4) Adaptation of conceptions of the self.
5) And/or transhumanist reconstruction of self.
6) And/or slowdown of environmental reconstructions.

Conclusion

It is time to summarise that which can be concluded on the basis of the reasoning conducted in this treatise.

The very first paragraphs evoked, in dramatised form, a challenge to human identity. The research questions were the following: How will the expression of human agency and autonomy be affected as humans are increasingly embedded in algorithm-intense environments? How are changes in the expression of human agency and autonomy, in turn, likely to affect the viability of dominant worldviews? What noteworthy problems, if any, follow from the changes discussed?

The answers to the first two questions are summarised at the end of this concluding section. Answers to the third question are suggested in the discussions that follow below. The answers to the questions are related to the challenge to human identity in the following way. In parts of the treatise, principally in part II, we have been primarily concerned with likely changes in the expression of human agency and autonomy. We became familiar with some ideals held by some of the cited authors and some of the commented-on societies. Some types of technology, we learnt, are likely to undermine the convivial ideals of Illich and the form of community practised by the Amish. Ellul, moreover, presumes some sort of 'natural' or 'ideal' condition within which human beings are able to exercise free choice; as technological society undermines this condition, and bare functionalism usurps freedom, human beings experience alienation. The answers to research question number two, within the scope of this treatise, presuppose neither a natural human condition nor any objectively valid timeless ideal from which human beings could be alienated. To the extent that human beings experience alienation, given the conceptual apparatus developed here, it should be tied primarily to disharmony between their worldviews and the worlds they actually inhabit. This does not entail that there are no conditions that are constitutionally unfit for the human species. Nor does this entail that there is no natural human condition or that there are no objectively valid timeless ideals. But if there are, then, given the conceptual apparatus and dynamics discussed here, it would still be the case that environmental changes would tend to produce changes in worldviews that bring the humans who inhabit them closer to or further away from natural conditions or objectively valid timeless ideals.

Even though no natural condition or objectively valid ideal is presupposed, the analyses offered in this treatise imply nevertheless that something important is threatened. It has been explicitly argued that, as humans become embedded in algorithm-intense environments, the mythology of modernity and modern worldviews will come under pressure. Yet it has never been claimed that modernity is a one-sided blessing. Moreover, traits typically inherent in pre-modern societies have been discussed with approval. Still, it has also never been argued that pre-modern societies have been a one-sided blessing. Within the scope of this investigation, that which seems to be threatened is the human ability to express some basic form of human agency – an experienced and genuine scope for thinking and acting freely and spontaneously. It would be *genuine* in the sense that human thinking and acting would not be merely a functional consequence of bureaucratic or algorithmic behavioural manipulation. Keen critical attention has been directed at all processes that could be interpreted as drawing us nearer to algorithmic determinism.

With respect to the three paradigmatic cases predicting the role of human agents in future algorithm-intense environments, the more general conclusion that can be drawn is that all of them probably highlight dynamics that will be at play in algorithm-intense environments. Susskind makes a plausible case for a future with less work for humans. His solution to the consequent challenge to human purpose and meaning, involving leisure policies, is less convincing. The loss of *use* for human beings, it has been argued, would likely be much more consequential than Susskind is willing to recognise, involving, among other things, the removal of human incentives to learn critical skills and widening asymmetric power dynamics between the humans in control of critical algorithmic systems and the humans who are controlled through such systems. Kissinger, Schmidt and Huttenlocher, in their advocacy of human–AI partnership, choose not to dwell on this latter inconvenience. The challenge that this treatise presents to their vision of the future is presented in the series of analyses conducted in part II. To the extent that we care to preserve a basic ability to express human agency for broad human populations and not only for a small elite, the case of Weiner must be deemed the most insightful. Increasingly algorithm-intense environments, then, will raise the cognitive requirements for human agents. To the extent that humans fail to rise to the challenge, it should be expected that dynamics such as those portrayed by Susskind and Kissinger *et al*. will predominate and/or, following Wiener's prediction, that increasing dysfunction and, potentially, disaster will follow.

If we consider the notion of surrendering human control to machines, Susskind and Wiener seem to stand on opposing sides, while Kissinger *et al*. envisage the prospects of an entirely novel partnership. Russell's concept of user-friendly AI represents one way to respond to Wiener's goal-intention problem. This way of responding opens up new areas of problems, notably ones pertaining to the source of that which is to be admitted as desirable or moral: it impels us to inquire into virtues and vices and cultural differences. It

seems unlikely that a generally satisfying solution will be developed in the near future. Wiener's mechanical slave and timing problem are of a different order, and the analyses in parts II and III imply that the dynamics involved therein will remain serious threats to a humane future: even if the goal-intention problem were satisfactorily solved, the scope for expressing human agency in environments defined by quasi-omniscient algorithmic systems operating at super-human speeds could become restricted to sub-human proportions. For these reasons it could be concluded that Wiener's prudential notion ought to be adhered to: 'To be effective in warding off disastrous consequences, our understanding of our man-made machines should in general develop *pari passu* with the performance of the machine' (Wiener, 1999, p. 81). Unfortunately, it could probably also be concluded that speed tends to be too important to powerful interests and that Wiener's recommendation, at least given current circumstances, therefore is unlikely to be heeded.

There is a practical argument in favour of rapid innovation: we must take the lead, or we will suffer at the hands of ruthless competitors that have even fewer scruples than we do. Then, if the notions of modernity – its mythology, worldviews, and conditions – developed here are accurate, it could also be concluded that effective incentives to accelerate technoscientific development are most likely embedded in predominant contemporary worldviews. Modernity, and especially late modernity, harbours a fascination for science and technology. Scientifically engineered technology represents, in a sense, the most powerful, prestigious, and objectively valid means of engagement and control in a world in which the interference of spirits and gods abated long ago. To put limits on technoscientific ingenuity represents retreat and backwardness. Barring any radically disruptive events, or other pattern-breaking changes, such as a boosting of some existing ethical principles or the emergence of new higher ethical principles, so that non-technical concerns begin meaningfully to regulate and limit the dominion of technique, things are likely to proceed in accordance with the patterns previously discussed. Ethical values, then, rather than shaping the world in which we live, play a primarily reactive and adaptive role in the face of the changes brought about by technical innovation.

Much of the reasoning in this treatise has concerned higher principles. In this context 'higher' is a relative notion: either, say, the pursuit of technological excellence is the highest social good, or something else is valued as higher. We have recognised that in technically advanced societies there may well be things that are valued higher than technological pursuits; but it has been argued that such higher values are often so abstract, vague, and subject to dispute that they fail to incentivise much meaningful limitation of the dominion of technique. In contrast, locally situated Amish communities still limit the dominion of technique by subsuming it under their goal of sustaining traditionally structured local communities. Their goal suggests that they ought

to obstruct the technology-induced 'liberation' of the members of their community from the mutual needs that the members have of one another.

The analyses of Illich and Ellul help us to see the extent to which industrial liberation from pre-modern mutual needs implies new industry-related needs and dependencies. Many of modernity's initial impressions in the times of revolutionary change were no doubt formed partially on the basis of ignorance: that which tends to become apparent in the medium and long term is rarely apparent in the short term. Even today, no doubt, it can be tempting for a young Amish person to seek liberation from tradition-imposed duties and limitations in order to be able to explore the agency-venues available on the arenas of modernity – just as it may be equally tempting for a modern agent locked within the functional confines of an increasingly hybrid bureaucratic organisation to seek a simpler and more humane way of life, more in the image of the one practised by the Amish.

Many readers will no doubt think that the Amish way is much too rigid. Even so, the extreme opposite way – one in which novel technologies invented and produced outside local contexts are allowed, without any constraints, continually to upset and revolutionise the status quo – will certainly be beset by its problems. The analyses conducted here do not compel us to conclude that one specific way of life is better than others. They do suggest that we are moving towards the latter extreme, and, given that we seek to avoid losing control, that a more prudential approach is desirable. If there is one lesson that even those who disapprove of Amish tradition should be able to draw from their example, it is to dare to think deeply about the function of technologies without fascination, reverence, and fatalistic acceptance – that is, to exercise human agency and autonomy rationally in the service of some higher ethical aim *other* than the mere increase of technical power.

Unlike the Amish, who have chosen to practise a way of life secluded from the modern world in which they are situated, Illich endeavoured, from within the modern world, to supply foundations for an alternative and more convivial modernity, one of 'higher social effectiveness with lower industrial efficiency' (Illich, 2021, p. 20). Illich recognised that the prowess of modern science could be used either to bring us into a 'hyperindustrial age of electronic cybernetics or to help us develop a wide range of truly modern and yet convivial tools' (Illich, 2021, p. 34). In theory, a convivial modernity seems feasible. Yet, in practice, given the conditions discussed above – competitive dynamics in a ruthless and competitive world and worldview-embedded incentives to accelerate technological innovation – it seems unlikely that convivial patterns will develop to any significant degree in the near future. Since Illich wrote the book cited, in 1973, technological societies have moved further away from conviviality and further towards hyperindustrial electronic cybernetics.

The analyses conducted here have contributed to accentuating the risks of moving in this direction. Evolving organisational bio-technical hybridity enables ever more powerful organisational and systemic structures to function

with ever less human input. Just why this should matter so much becomes clear when we shift the focus on human autonomy from the individual to collaborative organisations. The revolutionary shift describes a movement away from societies composed of individuals who are organised into teams of doers, towards algorithm-intense organisations composed by machinery that is algorithmically organised into teams of doers. Contrary to how Susskind envisages a future with less work, the analyses conducted imply that a world with *less need* for human agents bodes very ill for the prospects of human beings. As humans are no longer needed *qua* workers, it has been argued, the incentives to learn critical skills are likely to weaken. Wiener argues that we need to learn more, not less, in order to be able to cope with increasingly automated environments. It may well be possible to replace much of the cognitive input that is required from human agents in order to make such environments function with input from artificial agents, and thus enable an algorithm-intense 'society' to function without raising the cognitive performance of human agents. It would, however, be a 'society' composed not so much of human beings as of algorithmic systems, possibly managed by a small elite of human agents, managing and husbanding human populations. This is hardly a society in any conventional sense. To the extent that we seek to safeguard an ability to express human agency and autonomy within some tolerable scope of limitations, such prospects should frighten us. If, on the other hand, the AI hype is off target, and the replacement of the required human input with the input of artificial agents fails to produce the desired results *and* humans not only fail to raise their cognitive abilities but actually undergo a cognitive diminishment because of the incentive-removing rhetoric about the potentials of AI, then increasing dysfunction should be expected.

The outcomes discussed above represent future potentials in a continuous process. Ellul argued that, in the absence of any higher principle that limits the dominion of technique, in the service of bare efficiency, technique readapts the way in which human agents express agency. Ellul argued that this was the case long before the development of the algorithmic systems of today. In the contexts that he describes, human agents were valued nevertheless as *the* agents that mattered – there were no other agents that could run and operate industrial modes of production and apply technique to all domains of society. In order to render technological society ever more efficient, human agents needed to be re-educated so that they fitted better with continually evolving technical means. Although Ellul considered such conditions inhuman, human agents were nevertheless still highly valued as the principal doers in society.

Brutal as early industrial contexts may have been for workers, the analysis of Zuboff reveals that technique can be used to re-adapt human agents to functions that are baser than work-related functions in the service of industrial institutions – to re-adapt them into sources of raw material. Put together, the way in which artificial intelligence learns, the envisioning of Susskind, and the analysis of Zuboff suggest prospects that should worry us: environments

in which the mining of human intelligence, knowledge, and attention is *expedient* in view of developing critical technologies, *necessary* to maintain technical functions in contexts where human agents are increasingly excluded, and *profitable* to the agents in control of algorithmic systems.

Meanwhile, it could be argued, new vistas for the free exercise of human agency and autonomy are likely to open up in algorithmically defined virtual realities. The thinkers included in this treatise are perhaps too critical. If others had been included, different conclusions could have been reached. We have nevertheless reasoned about the possible bifurcations of human experience along phenomenological and ontological lines. Even if algorithmically defined realities should be experienced as pleasant, it would still be deeply disturbing if such experience were bought at the price of manipulation on an ontological level, which the analysis of Zuboff strongly implies is a common occurrence even in many current systems.

The gravest moral challenge is located on the *real* level, outside myopic phenomenological experience in algorithmically defined spaces. What people choose to do in virtual realities may be disturbing, potentially beneficial, or potentially debilitating; but if Susskind's predictions came to fruition and there were less and less need for humans *qua* workers, then this would constitute a moral challenge on many levels. Let us revert to the examples of hospitals and concentration camps. With the help of Ellul and Bauman, it has been posited that we live in contexts that could be described as quasi-value-relative dynamics: whereas the objective value of scientifically engineered technologies tends to endure, other values adapt in accordance with changing contingencies. It was argued that the concentration camp and the hospital were structurally similar, that they were both examples of organisation defined by administrative and technical prowess, but that they typically, or in 'normal' circumstances, were organised towards different ends. If the values that designate ends are adaptable, then, it stands to reason, one organisation may in time turn into the other. To turn a hospital into a concentration camp is easier said than done for as long as the operation of the organisation depends on the willing collaboration of many human agents. Consulting the works of Bauman (2018), Ellul (2004), Hanna Arendt (2022), and others, one could nevertheless understand how it could be done. Until today, human agents remain critical in order to sustain the function of either form of organisation. It would require considerable time and effort to re-educate teams of human agents who are accustomed to running a hospital as willing administrators of a concentration camp. Many would typically refuse to participate. In the algorithm-intense hybrid multi-agent organisations of tomorrow, we cannot know the extent to which humans alone will think out and define the values that are to guide the operation of the organisations. Moreover, to the extent that such organisations will not need human agents to carry out their goals, a barrier to the potential accomplishment of ends that many humans would be reluctant to carry out will have been removed. This should worry anyone who happens to

be attached to a human society, composed *of* humans *for* humane human purposes.

It has been posited that algorithm-intense hybrid multi-agent organisations tend to have an internal organisational goal – namely, to make themselves as independent as possible of the human agents that are still recruited to enable their functions. It has not been argued that such a goal frames the development of every single organisation of such a type. It nevertheless seems to describe a general trend: wherever it becomes possible, human task-performing agents are replaced by algorithmic task-performing systems, so that organisations become more algorithm-intense. Such moves supposedly increase profitability and administrative control. As stated above, it also removes potential obstructions, owing to human moral scruples, to the accomplishments of goals.

Nevertheless, even highly algorithm-intense organisations still depend on extensive human interactions, in notable instances on the mining of human intelligence, knowledge, and attention for the purposes of training algorithmic systems and making a profit. In analogy with Bauman's historical description of how internees were recruited by their captors to be administrative agents in concentration camps, it was described how humans could be recruited as agents for the purpose of removing humans from the loop in organisational tasks, and how human instrumental rationality and desire could be harnessed in algorithm-intense environments. The presupposition of these scenarios is a radically altered human environment, one that puts the high-modern and supreme cartesian agent into question. The so-called knower and controller of matter in time and space seems to be busy constructing an environment that is able to know as well as to do. In the words of Zuboff, this 'signals the metamorphosis of the digital infrastructure from *a thing that we have to a thing that has us*' (Zuboff, 2019, p. 203). To the mindset formed by modernity and a modern philosophical anthropology, it also signals the end of the buffered self and a new kind of porosity. We are not about to inhabit the animated nature of our ancestors, worlds defined by recurring natural patterns and cycles, but rather a quasi-animated work-in-progress, an engineering project, a world defined by sudden disruptions, sudden emergences of novel phenomena, and manifestations of powerful agencies that will be more or less opaque to our understanding.

The possibility cannot be excluded that humans will find new ways to flourish and feel at home in algorithm-intense environments. One could ask: what is 'home', anyway, if we posit a context in which values and goals are becoming more and more adaptable to the evolutionary processes of the technical means at our disposal? Under such circumstances, there is perhaps no notion of home worth defending and preserving, just as we should not be overly attached to the values that we happen to value today. But this way of reasoning represents a fatalistic surrender to technological determinism. We cannot here conclude exactly which notions of home, which values, or which worldviews are worth defending and preserving. That which can be concluded, on the basis

of the analyses undertaken here, is that, if we allow technological imperatives to reign over and above all other values, then we have reason to fear that our life-milieus will be rendered less humane. To the extent, therefore, that we wish to preserve a basic humanness for posterity, regardless of the specific worldview that we happen to inhabit, it will be useful to be aware of the dynamics described in this treatise in *all* assertions of non-technical values. If humans are to find ways to flourish and feel at home in algorithm-intense environments, this should not be expected to occur by itself as we give free rein to unrestricted technological evolution, but as a result of the assertion of non-technical goods combined with a deeper understanding of the dynamics that frame human interaction with technologies. If, on the other hand, we wish to conserve an already practised way of life for posterity, even as our environments turn increasingly algorithm-intense, we must become aware that that which we wish to preserve is not the result of some abstract tradition or some abstract intellectual activity. It will be better conceived of as a contingent product resulting from many generations' interaction with some given type of environment. As Walsh informs us, the status of environments – including all dynamics that play out in them – will often be of fundamental importance for the conservation of home and other values.

Should we submit to the Matrix? To this rhetorical question the answer is perhaps too obvious. But given the views we have discussed here, we should be careful not to submit to anything in the vicinity of the Matrix, including to the utopia envisaged by Susskind. Possibly the Matrix is an allegory for a contemporary society that is already overly bureaucratised, where humans, in lieu of freely exercising their agency, fill routinised or even predetermined agentic functions. This is the kind of society which Illich – and I – wish to oppose. To break out of the Matrix symbolises freedom reclaimed. Inconveniently, again, no first principles have been stated in the treatise, on the basis of which we could conclude exactly what kind of society we ought to seek. Rather than to argue in favour of one or several ideals, the objective has been to illuminate dynamics, the awareness of which will be useful from the point of view of many different ideals and worldviews. What could breaking out of the Matrix mean, then, if applied to an overly bureaucratised consumer society that is turning increasingly algorithm-intense? It could mean that it behoves each and every one of us to reflect on first principles or important ethical values *other* than bare technological efficiency and expediency, and by means of such reflection to clear a limited space, starting from our own person, within which the hegemony of mass-produced technology is limited to the *service* of some overarching humane principle or value.

Submitting to the Matrix, on the other hand, could represent a mechanisation of human culture, in the sense that Guardini defines culture. To the modern mindset, according to Guardini, culture as a concept includes all spheres of activities that mediate between humans and their natural environment. Modern culture provided larger and larger buffers against the dangers

represented by natural environments. Over time, Guardini informs us, culture itself, ever more powerful, began to represent the greatest danger, since humans had not developed any reliable way to exercise power over their own power. Facing the algorithmic paradigm, we seem to be standing before a culture that increasingly excludes human beings *qua* consequential agents. The elements that used to constitute culture are becoming mechanised. This view offers another perspective on the new porous self: once again, human agents may be facing an environment replete with risks and dangers, but with very feeble mediating means to ward off the dangers.

Non-mechanised culture, it is to be understood, is based on human-intense institutional frameworks. Even in modern contexts deemed to be 'individualistic', individualism only makes good sense in contexts where there are arenas (e.g., institutions) in which human agency can be exercised with some benefit to individuals. In exchange for abandoning social contexts defined by traditional duties and parochial boundedness, modernity offers individuals prospects of professional careers and social and geographic mobility. The algorithm-intense era, at least in the version envisaged by Susskind, appears to imply the removal of consequential institutional arenas *and* continued dependence on institutions. Thereby it excludes humanity, or large chunks of it, if we follow Guardini, from *meaningful* or *consequential* cultural activity. The species that used to supply agents that were useful in innumerable contexts is increasingly understood in terms of populations that must be managed (e.g., leisure policies). Although we cannot exclude the possibility that such new conditions might offer pleasant experiences to the humans living in them, only with great casuistry could such disempowerment be interpreted as liberating.

The human–AI partnership envisaged by Kissinger *et al*., on the other hand, offers a meaningful and consequential cultural role for human agents. This, at least, will appear to be the case at some levels of technical hierarchies. At lower levels of technical hierarchies, it has been argued, that which appears to be affording partnership may be specious, affording short-term conveniences at the cost of medium- and long-term extractive dependence. The scenario of Kissinger *et al*. and that of Susskind could, in fact, be quite complementary. The small elite that will still have meaningful work-related tasks in the case of Susskind could be identical to the humans who form culturally meaningful and consequential partnerships with AI in the case of Kissinger *et al*.; meanwhile, the humans who form partnerships at the lower levels of technical hierarchies could be identical to the populations who are subjected to leisure policies. In both of these cases, it is difficult not to see the shaping of extremely inegalitarian structures.

The case of Wiener, finally, identified a set of critical automation-related challenges to human agency and autonomy. The crucial implication is that human cognitive abilities need to increase in order to ward off disaster. Since we do not know the extent to which that increase could be accomplished, the case of Wiener suggests that, if we continue technological innovation at the

current pace – or even more so if we accelerate innovation – we are moving ever nearer the edges of disaster.

All these cases, in their more developed forms, have moved so far away from anything that could be labelled 'modern conditions' that they are not even borderline conditions: they represent something new, something beyond the anthropocentric notion of modernity. In modern and borderline conditions, it has been posited, human agents have been able live out a sort of creative tension between modern ideals and the cog-in-the-wheel version of modernity. This has often taken the form of heroic struggles for ideals – but in *this* world, which is, or used to be, a common human-run world. We now stand on the brink of a more and more algorithmically defined condition, or set of conditions, the ultimate limitations and affordances of which will be revealed in time. Some affordances, such as those commented on by Zuboff, have already been revealed in our time. We have also considered the type of affordance that enables the creation of algorithmically defined environments, which human agents could experience as worlds in and of themselves. In the more dystopic episodes of the treatise, the prospects of hedonistic virtual worlds in a more and more oppressive real world were discussed, and a bifurcation of human experience was raised as a possibility. Arguably, we are already experiencing such bifurcation, in which pleasure, comfort, and help are sought at the lower levels of technical hierarchies – in virtual worlds, in partnership with artificial agents – whereas human agents at the higher levels still meaningfully and consequentially engage in cultural activity.

This brings us to the consideration of the power dynamics in algorithm-intense environments. Guardini and Lewis considered power dynamics in modern conditions. The inability to control its own means of power, according to Guardini, was at the root of the modern predicament: the cultural means that had been used to ward off dangers posed by the natural environment were increasingly turning into fertile ground for new dangers. Meanwhile, Lewis pointed out the illusory aspects of modern individual empowerments: 'If I pay you to carry me, I am not therefore myself a strong man. […] What we call Man's power is, in reality, a power possessed by some men which they may, or may not, allow other men to profit by' (Lewis, 2006, p. 54). The latter applies in typically modern conditions, within which society's most powerful institutions can be described as purely human multi-agent organisations. In our era of hybrid and algorithm-intense multi-agent systems and organisations, the phenomenon of human power, if there is such a thing, seems to be slipping even further away from human control: the sheer complexity, the prevalent opacity, and the wide applicability range of currently applied and potential future iterations of algorithmic systems add up to a tool of cultural mediation, the medium- and long-term consequences of which must be considered extremely unpredictable. If the ability of an agent – human or artificial – to navigate its environment depends on its ability to adequately predict its

environment, then the inability to predict the consequences of cultural undertakings can hardly be understood as representing human empowerment.

In the short term, many nevertheless seem able to make fairly accurate predictions of the consequences of the application of algorithmic systems. The designers and controllers of algorithmic systems make it their professional task to overlook the whole; meanwhile, 'users' or 'partners' at the lower ends of technical hierarchies practise in order to acquire intuitive predictive notions on the level of the interfaces with which they interact. The former make it their business to acquire a bird's view of the whole in view of short-, medium- and long-term gains; the latter train their perception in the myopic mode for the sake of acquiring short-term benefits. Even if the former *consider* medium- and long-term consequences, we can hardly assume that they have accurate predictive models for what *will* happen in the medium and long term. The algorithm-intense society is an ongoing experiment; the civilisation that depends on its functionality is, in the medium and long term, hanging on a thread, the dimensions and strength of which are unknown.

The potential of experience bifurcation in algorithm-intense environments has implications for the potential viability of modern worldviews and the modern philosophical anthropology that is centred on the notion of the buffered self. Even if things were to continue evolving in accordance with, for instance, Zuboff's surveillance capitalist structures, we could not exclude the possibility that the lower ends of technical hierarchies might still *afford the experience* of empowerment and liberty. Who knows? In some of the algorithmically defined spaces of tomorrow, the modern type of agent might be able to flourish as never before. If this were to become the case, then there would be no challenge to modern worldviews on the level of experience. Yet, if we were to zoom out and adopt a realist view of the ontological whole, we would nevertheless perceive that this type of condition, within which humans are able to live in different worlds, is very different from the conditions that engendered modernity.

Such a new type of condition, one in which human agents live in ongoing engineering projects that are subject to constant reconstruction, is not one in which most humans could attain the ideal of *kubernétes*. Moreover, to the extent that such ongoing engineering projects, in which human agents live their lives, have the ability not only to know but also to act – in some respects algorithmic systems will know more about us than we do about ourselves, and be able to act on levels that exceed our own capacity – the notion of the buffered self will seem more and more ridiculous. We are not facing a return to the porous self of our ancestors. In the ancient contexts, even though nature was opaque to scientific explanations, our ancestors were typically able to observe regular patterns in nature and to build up impressive knowledge about their natural environments. Combined with such knowledge, they practised the arts of incantations and magic in order to effect desirable outcomes. The porous self of algorithm-intense environments may become increasingly

compelled to rely on the equivalent of incantations and magic – prompt-engineering, divorced from any real engineering skills, practised in order to effect desirable outcomes – without the possibility of observing the equivalent of regular patterns in nature and thereby building up impressive knowledge about its artificial environments – not, at least, to the extent that artificial environments will remain ongoing engineering projects.

The surprising irony, then, comes from the notion that emerging algorithmic systems appear to be manifesting some of the qualities that modern agents typically imagine that their ancient ancestors feared in the animistic world, and from which modern agents imagine that they liberated themselves once and for all by means of the culture and heroic feats associated with modernity. Contemporary and secularised humans can *already* invite oracular and agentic intervention – not, of course, by petitioning spirits or gods, but by petitioning chatbots.

In order to maintain the *kubernétes* ideal and a buffered self in a dawning algorithm-intense world, some see a solution in transhumanist reconstructions of the human agent. If this were possible, it could meet Wiener's ultimate challenge to the human agent, the one that would require humans to increase their cognitive capacities. Wiener himself, however, seems to have preferred the *pari passu* way of proceeding – that is, a slowdown. Although the explicit discussion of this adaptive measure was brief, it overlaps with the way practised by the intermittently discussed Amish. It implies not letting technological innovation *drive* events, but subjecting technological innovation to higher ethical goods not only in the short term in relation to that which has become obvious to everyone, but also to medium- and long-term considerations. This would require us to become cognisant of dynamics such as those discussed in this treatise. In order for the modus operandi of technically advanced societies to be changed, our societies would also need to undergo radical change. This does not imply a reversion to pre-modern or pre-industrial societies; it probably does imply, as Illich stated the matter, that we must become fully humane societies, with higher social effectiveness but with lower industrial efficiency.

***

The answers to the second research question have been worked out within the framework of this treatise on the basis of the answers to the first research question: the manner in which the expression of human agency and autonomy is likely to be affected in algorithm-intense environments, then, has repercussions for the viability of modern worldviews. The function of human beings *qua* agents in larger wholes and our ideal notions about human beings – namely, the philosophical anthropology at the core of our worldviews – constitute the main foci of this treatise.

It is now time to summarise the main conclusions that can be drawn from the work undertaken.

If we accept the reasoning concerning so-called late-modern conditions, notably that conventional values tend to be viewed as increasingly relative in relation to the one sector of human activity that is continually proving its power and usefulness, namely, technique, then it is difficult to envisage this trend being reversed in the near future. In general, contemporary subjects are too enmeshed in and dependent on their environments – ongoing engineering projects – in order to be able to fashion technique-subordinating ethics. But if technique-driven developments and changes incentivised adaptations of the ethics of the societies in which they occurred in earlier industrial stages, it seems reasonable to conclude that algorithm-intense societies introduce a new factor into the whole: to the extent that algorithmic systems are consulted on ever wider ranges of issues, including ethics, it becomes even more difficult to envisage large numbers of humans following in the footsteps of Illich or the Amish. Thus, the expression of human agency and autonomy is likely to be even further enmeshed in and defined by technique in algorithm-intense environments.

The analyses in this treatise strongly suggest that we are moving closer to the opposite of the Illich-Amish end of the spectrum, reinforcing instead a condition in which novel technologies invented and produced outside local contexts are allowed, with few constraints, continually to upset and revolutionise the status quo. Even though the analyses do not permit us to conclude that some specific ethics or way of life is necessarily better than others, normative *oughts* can nevertheless be derived on the basis of commonly shared assumptions. If, for instance, we desire to maintain a degree of medium- and long-term human control over culture and society, then the analyses strongly suggest that we ought to adopt attitudes similar to those of Illich and the Amish vis-à-vis our technical milieu – namely, that we ought to dare to think deeply about the functions and affordances of technologies without fascination, reverence, and fatalistic acceptance. As a necessary complement to such an attitude, Wiener's prudential approach should be adopted on a wide personal and social scale, establishing as an aim that our understanding of machines should develop *pari passu* with the performance of machines. This probably implies a considerable slowdown of technological innovation.

The analyses suggest two potential and mutually non-exclusive consequences of a non-prudential accelerated approach: the further strengthening of the autonomy of technique at the cost of further restrictions on the scope within which human agency and autonomy can meaningfully be expressed, and/or systemic dysfunction.

As for how the expression of human agency and autonomy is likely to be shaped in algorithm-intense contexts, the analyses provide important hints. In previous industrial phases, it has been argued, the new conditions triggered a need continually to re-educate human agents in order for them to be able to sustain continually evolving technical means. Although thinkers such as Ellul and Illich tend to interpret such conditions as bordering on the sub-human,

humans *qua* agents are still valued as indispensable in such conditions. If algorithm-intense structures end up needing less and less sustenance from human agents, then it will follow that human agents could be re-adapted for purposes other than sustaining the operative function of technological structures. In contexts where there is very little need for humans *qua* agents in work-related tasks, humans could be re-adapted for whole ranges of purposes, limited only by the imagination of those who – or *that* which – control the means used to control human populations.

In the conditions analysed by Zuboff, humans are increasingly embedded in algorithmically defined spaces. We have become familiar with the notion of algorithmically defined 'worlds', and have considered the potential of a novel and radical bifurcation of human experience. If at earlier stages of human history people at the lower and higher ends of social hierarchies have tended to engage in society in different ways, that which is radical with respect to the higher and lower ends of the emerging technical hierarchies is that, potentially, the lower ends may afford all manner of conveniences and even the experience of seemingly limitless satisfaction of wants and desires, while the higher ends simultaneously afford unprecedented control and power. Short-term conveniences at the lower ends can come with concealed medium- and long-term costs. To the extent that we care to maintain a basic scope for the expression of human agency and autonomy, we should not let ourselves be seduced by the experiences that algorithmically defined spaces afford so that we ignore the wider workings of things.

The more consequential dangers that increasingly autonomous algorithm-intense multi-agent structures represent become manifest only when we consider the broad, real, or ontological context, as opposed to narrow phenomenological experiences that the lower ends of technical hierarchies may afford. One of the graver dangers that can be derived from the analyses undertaken here emerges from the increasing independence of algorithm-intense multi-agent organisations from human oversight and guidance. As fewer humans are needed *qua* agents in order to carry out organisational goals, it becomes easier to pursue goals independently of widely shared human sensibilities. The example discussed was that of changing the organisational goal of a bureaucracy, turning a hospital into a concentration camp: the conventional human-intense organisation would require considerable effort in order to overcome the obstacle of reluctant key human agents in the transformative venture; to the extent that an organisation is liberated from the need of human agents, changing organisational goals becomes easier. This, in turn, implies that the algorithm-intense environment potentially becomes, in some sense, less predictable to human agents.

The sense in which technical elements become less predictable relates to the transformation described by Zuboff, within which technology turns from a thing, an instrumental means that we have at our disposal, to *a thing that has us*. It is not only that technologies have become integral parts of our

environment: technical elements embedded in our environment are increasingly able to perceive, know, and act with higher and higher degrees of autonomy. This transformation is pivotal in provoking a return of the porous self, for that which was used to reinforce the notion of a buffered self, namely, scientifically engineered technology, is turning into a novel quasi-animated milieu, thus raising the prospect of a post-buffered-self. The new porosity in some sense represents a higher degree of vulnerability than both the modern buffered self and the pre-modern porous self: for it is embedded in an ongoing engineering project that is subject to frequent and sudden changes, a quasi-animated environment that, at least in the short to medium term, appears not to be following regularly recurring cyclical patterns.

Another source of vulnerability concerns the status of the more general means used to render the modern subject buffered, namely, the status of culture, in the sense that Guardini uses culture. As culture becomes more algorithm-intense, both in its narrow technological components and in conventionally non-technological spheres of human activity, we stand before the prospect of a culture that excludes human beings *qua* meaningful and consequential participatory agents. If culture, which was used as a means to ward off the dangers of the natural environment, is leaping out of human control, then, in a novel quasi-animated environment, humans risk being without any time-tested means that can be used to ward off the dangers of this new environment. This does not necessarily doom human beings, for new means may be developed; but it does represent a new vulnerability.

This new vulnerability could be explained, in part, by the loss of consequential arenas for human agents. This loss is most clearly illustrated by the scenario presented by Susskind. Humans are to be largely excluded from the executive functions of institutional or work-related arenas. In the terminology of Guardini, such arenas correspond to highly consequential cultural spheres of activity. In Susskind's future, as humans are excluded *qua* consequential agents from institutions, which will be largely animated by AI, humans will become even more dependent on the very same institutions. The human individual, then, turns from a needed cultural animator to a subject that must be managed, the human population from a cultural body of people to populations that must be managed. Exactly what this would imply for most people would depend on the circumstances, whims, or strategies that govern the workings of the system.

The human–AI partnership envisaged by Kissinger, Schmidt and Huttenlocher presents us with more ambiguous prospects. Here a consequential role is still maintained for humans *qua* agents. The question is whether the so-called age of AI will take us towards extremely inegalitarian social outcomes, where people at the lower ends of technical hierarchies are as removed from consequential agentic functions as are the larger portions of the population in the case envisaged by Susskind, or whether it will bring about more ambiguous and unexpected outcomes. If the latter turns out to be the case, then we

may experience surprising modifications of the scope within which human agency and autonomy can be expressed.

The case of Wiener, meanwhile, acts as a potential spoiler to the other two paradigmatic cases: the requirements for human cognition will increase, not decrease, in algorithm-intense environments. The more AI-intense we make our environments, the more capable we humans must become. If Wiener is right, then, to the extent that we are unable to increase the capacity of human cognition, we should expect dramatic dysfunctions in algorithm-intense societies.

In considering the viability of modern worldviews and the modern philosophical anthropology in algorithm-intense environments, we must distinguish between, on the one hand, potential phenomenological experiences on a personal level, and, on the other, algorithm-intense conditions as wholes. It could well be the case that modern worldviews will turn out to be highly viable in some experienced algorithmically defined spaces. However, any wider awareness of novel conditions as wholes, it has been argued, would elicit adaptive measures. Of the three adaptive measures we have discussed, one – namely, adaptation of the pace and intensity of environmental reconstruction – would potentially enable modern worldviews to persevere with minor modifications. Another, transhumanist adaptation of the physical human agent, might, if it could be accomplished successfully, enable large structures of modern worldviews to remain heuristically viable. The agent at the core of the worldview, although no longer strictly human, could still plausibly conceive of itself as buffered and destined to seek control over matter in time and space. The most radical adaptation of modern worldviews is represented by adaptation of its philosophical anthropology, the abandonment of the notion of a buffered self for a new kind of porous self. This is equivalent to forsaking a long-dominant human ideal, one that has spurred innumerable humans to mobilise reason, science, and technique for the purpose of controlling their environment. Such abandonment is likely to occur in steps, parallel to adoption of emergent coping tactics that, more or less successfully, could be levelled against an opaque, capricious, and quasi-animated environment.

Bibliography

Adam, B. (1990) *Time and social theory*. Cambridge: Polity press.

Anscombe, G.E.M. (1985) *Intention*. Second ed. Oxford: B. Blackwell.

Arendt, H. (2022) *Eichmann in Jerusalem: a report on the banality of evil*. London: Penguin Books (Modern classics).

Bai, N. (2023) 'Brain Implants Revive Cognitive Abilities Long After Traumatic Brain Injury', *Stanford Medicine News Center*. Available at: https://med.stanford.edu/news/all-news/2023/12/traumatic-brain-injury-implant.html (Accessed: 29 August 2024).

Barbour, I.G. (1996) 'Response to Critiques of Religion in an Age of Science', *Zygon®*, 31(1), pp. 51–65. Available at: https://doi.org/10.1111/j.1467-9744.1996.tb00007.x.

Bauman, Z. (2000) *Liquid modernity*. Cambridge, UK : Malden, MA: Polity Press ; Blackwell.

Bauman, Z. (2003) *Liquid love: on the frailty of human bonds*. Cambridge, UK : Malden, MA USA: Polity Press ; Distributed in the USA by Blackwell Pub.

Bauman, Z. (2007) *Liquid times: living in an age of uncertainty*. Cambridge: Polity Press.

Bauman, Z. (2008) *Modernity and the Holocaust*. Repr. Cambridge: Polity Press.

Bernays, E.L. (2011) *Crystallizing public opinion*. Brooklyn, N.Y: Ig Pub.

Bloch, M. (2008) 'Why religion is nothing special but is central', *Philosophical Transactions of the Royal Society B: Biological Sciences*.

Edited by C. Renfrew, C. Frith, and L. Malafouris, 363(1499), pp. 2055–2061. Available at: https://doi.org/10.1098/rstb.2008.0007.

Bostrom, N. (2003) 'Ethical Issues in Advanced Artificial Intelligence'. Available at: https://www.fhi.ox.ac.uk/wp-content/uploads/ethical-issues-in-advanced-ai.pdf (Accessed: 21 August 2024).

Bostrom, N. (2014) *Superintelligence: paths, dangers, strategies*. First edition. Oxford, United Kingdom: Oxford University Press.

Burke, E. (2004) *Reflections on the Revolution in France and on the proceedings in certain societies in London relative to that event*. London: Penguin Books.

Chapin, B., El Ouardani, C. and Barlow, K. (2016) 'Socialization'. Available at: https://doi.org/10.1093/obo/9780199766567-0133.

Chaudhary, M.Y. (2019) 'Augmented Reality, Artificial Intelligence, and the Re-Enchantment of the World', *Zygon®*, 54(2), pp. 454–478. Available at: https://doi.org/10.1111/zygo.12521.

Chaudhary, M.Y. (2020) 'The Artificialization of Mind and World', *Zygon®*, 55(2), pp. 361–381. Available at: https://doi.org/10.1111/zygo.12597.

Coeckelbergh, M. (2020a) *AI ethics*. Cambridge, MA: The MIT Press (The MIT press essential knowledge series).

Coeckelbergh, M. (2020b) 'Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability', *Science and Engineering Ethics*, 26(4), pp. 2051–2068. Available at: https://doi.org/10.1007/s11948-019-00146-8.

Coeckelbergh, M. (2022) *Self-improvement: technologies of the soul in the age of artificial intelligence*. New York Chichester: Columbia University Press (No limits).

Coeckelbergh, M. (2023) 'Narrative responsibility and artificial intelligence: How AI challenges human responsibility and sense-making', *AI & SOCIETY*, 38(6), pp. 2437–2450. Available at: https://doi.org/10.1007/s00146-021-01375-x.

Davidson, D. (1963) ‘Actions, Reasons, and Causes’, *The Journal of Philosophy*, 60(23), p. 685. Available at: https://doi.org/10.2307/2023177.

Deneen, P.J. (2018) *Why liberalism failed*. New Haven (Conn.): Yale university press (Politics and culture).

Dennett, D.C. (1987) *The Intentional stance*. Cambridge, Mass: MIT press.

Ekot (2024). Sveriges Radio.

Ellul, J. (1990a) *La technique: ou, l’enjeu du siècle*. 2e éd. rev. Paris: Economica (Classiques des sciences sociales).

Ellul, J. (1990b) *Propagandes*. Paris: Economica (Classiques des sciences sociales).

Ellul, J. (2021) *The Technological Society*. New York: Knopf Doubleday Publishing Group.

Ellul, J. and Porquet, J.-L. (2004) *Le système technicien*. Paris: le Cherche midi (Collection Documents).

Farman, A. (2020) *On not dying: secular immortality in the age of technoscience*. Minneapolis: University of Minnesota press.

Floridi, L. (2011) *The Philosophy of Information*. Oxford University Press. Available at: https://doi.org/10.1093/acprof:oso/9780199232383.001.0001.

Frankfurt, H.G. (1971) ‘Freedom of the Will and the Concept of a Person’, *The Journal of Philosophy*, 68(1), p. 5. Available at: https://doi.org/10.2307/2024717.

Frazer, J.G. (1996) *The golden bough: a study in magic and religion*. Abridged ed. London: Penguin Books (Penguin twentieth-century classics).

Gibson, J.J. (1979) *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Ginet, C. (1990) *On action*. Cambridge [England] ; New York: Cambridge University Press (Cambridge studies in philosophy).

Girard, R. (2010) *La violence et le sacré*. Paris: Hachette Littératures.

Girard, R. (2011) *Mensonge romantique et vérité romanesque*. Paris: Pluriel.

Guardini, R. (2019) *The End of the Modern World*. Eastford, USA: Martino Fine Books.

Hassani, B.K. (2021) 'Societal bias reinforcement through machine learning: a credit scoring perspective', *AI and Ethics*, 1(3), pp. 239–247. Available at: https://doi.org/10.1007/s43681-020-00026-z.

Hesiod *et al.* (2006) *Hesiod*. Cambridge, Mass: Harvard University Press (The Loeb classical library, 57, 503).

Hossaini, A. (2019) 'Forum: Artificial Intelligence, Artificial Agency and Artificial Life', *The RUSI Journal*, 164(5–6), pp. 120–144. Available at: https://doi.org/10.1080/03071847.2019.1694264.

Illich, I. (2021) *Tools for conviviality*. London ; New York: Marion Boyars.

Janz, P.D. (2014) 'Transcendence, "Spin," and the Jamesian Open Space', in C.D. Colorado and J.D. Klassen (eds) *Aspiring to Fullness in a Secular Age*. University of Notre Dame Press.

Jaspers, K. (1953) *The Origin and Goal of History*. New Haven and London: Yale Unversity Press.

Kahneman, D. (2013) *Thinking, fast and slow*. First paperback edition. New York: Farrar, Straus and Giroux (Psychology/economics).

Kant, I. (2017) *Groundwork for the Metaphysic of Morals*. In the version by Jonathan Bennett presented at www.earlymoderntexts.com. Available at: https://www.earlymoderntexts.com/assets/pdfs/kant1785.pdf (Accessed: 6 December 2024).

Kissinger, H., Schmidt, E. and Huttenlocher, D.P. (2021) *The age of AI: and our human future*. London: John Murray.

Kuhn, T.S. (1996) *The structure of scientific revolutions*. 3rd ed. Chicago, IL: University of Chicago Press.

Kurzweil, R. (2000) *The age of spiritual machines: when computers exceed human intelligence*. New York, NY: Penguin Books.

Kurzweil, R. (2018) *The singularity is near: when humans transcend biology*. This impression 2018. London: Duckworth.

Lanier, J. (2011) *You are not a gadget: a manifesto*. 1st Vintage Books ed. New York: Vintage Books.

Lanier, J. (2013) *Who owns the future?* London: Allen Lane.

Latour, B. (2010) *Nous n'avons jamais été modernes: essai d'anthropologie symétrique*. Nachdr. Paris: Editions La Découverte [u.a.].

Lee, E.A. (2019) *The coevolution: the entwined futures of humans and machines*. Cambridge, Massachusetts: The MIT Press.

Lewis, C.S. (2006) *The abolition of man, or, Reflections on education with special reference to the teaching of English in the upper forms of schools*. Princeton, N.J.: Recording for the Blind & Dyslexic.

Marett, R.R. and Penniman, T.K. (1932) *Spencer's scientific correspondence with Sir J. G. Frazer and others*. Oxford: The Clarendon Press.

Maushart, S. (2011) *The Winter of Our Disconnect: How Three Totally Wired Teenagers (and a Mother Who Slept with Her IPhone)Pulled the Plug on Their Technology and Lived to Tell the Tale*. East Rutherford: Penguin Publishing Group.

McLuhan, M. (1994) *Understanding media: the extensions of man*. 1st MIT Press ed. Cambridge, Mass: MIT Press.

Milgram, S. (1975) *Obedience to authority: an experimental view*. New York [etc.]: Harper Torchbooks.

Mill, J.S. and Gray, J.N. (2008) *On liberty and other essays*. Reissue. Oxford: Oxford University Press (Oxford world's classics).

Misselhorn, C. (ed.) (2015) *Collective Agency and Cooperation in Natural and Artificial Systems: Explanation, Implementation and Simulation*. 1st ed. 2015. Cham: Springer International Publishing : Imprint: Springer (Philosophical Studies Series, 122). Available at: https://doi.org/10.1007/978-3-319-15515-9.

Mitchell, M. (2020) *Artificial intelligence: a guide for thinking humans*. Published in paperback. London: Pelican, an imprint of Penguin Books (A Pelican book).

Modrell, A.K. and Tadi, P. (2023) 'Primitive Reflexes'. Treasure Island (FL): StatPearls Publishing. Available at: https://www.ncbi.nlm.nih.gov/books/NBK554606/ (Accessed: 12 November 2024).

Mouriquand, D. (2024) '"Miss AI": World's first beauty contest with computer generated women', *Euronews*, 25 April. Available at: https://www.euronews.com/culture/2024/04/25/miss-ai-worlds-first-beauty-contest-with-computer-generated-women (Accessed: 25 April 2024).

Mozorov, E. (2019) 'Capitalism's New Clothes', *The Baffler*. Available at: https://thebaffler.com/latest/capitalisms-new-clothes-morozov (Accessed: 28 June 2024).

Mullins, D.A. *et al.* (2018) 'A Systematic Assessment of "Axial Age" Proposals Using Global Comparative Historical Evidence', *American Sociological Review*, 83(3), pp. 596–626. Available at: https://doi.org/10.1177/0003122418772567.

Mumford, L. and Winner, L. (2010) *Technics and civilization*. The University Of Chicago Pres.

Munro, A. (2017) 'Mass Society', *Encyclopaedia Britannica*. Available at: https://www.britannica.com/topic/mass-society (Accessed: 29 January 2005).

Neuralink (2025) *Neuralink — Pioneering Brain Computer Interfaces*, *Neuralink*. Available at: https://neuralink.com/ (Accessed: 23 January 2025).

O'Connor, T. (2000) *Persons and causes: the metaphysics of free will*. New York: Oxford University Press.

Olafson, F. (1998) 'Philosophical Anthropology'. Available at: https://www.britannica.com/topic/philosophical-anthropology/Modern-science-and-the-demotion-of-mind (Accessed: 13 December 2024).

Ophir, Y., Rosenberg, H. and Tikochinski, R. (2021) 'What are the psychological impacts of children's screen use? A critical review and meta-analysis of the literature underlying the World Health Organization guidelines', *Computers in Human Behavior*, 124, p. 106925. Available at: https://doi.org/10.1016/j.chb.2021.106925.

Oswald, T.K. *et al.* (2020) 'Psychological impacts of "screen time" and "green time" for children and adolescents: A systematic scoping review', *PLOS ONE*. Edited by H.R. Slobodskaya, 15(9), p. e0237725. Available at: https://doi.org/10.1371/journal.pone.0237725.

Pageau, J. (2024) 'A Comment on Alex O'Connor's Conversation with JBP - Do Adam and Eve Die After They Eat the Fruit?' Available at: https://www.youtube.com/watch?v=cnsUT97VSeA (Accessed: 26 November 2024).

Pascal, B. (2022) *Pensées*. Edited by M. Le Guern. Paris: Gallimard (Collection folio classique, 4054. classique).

Pasquinelli, M. (2019) 'How a Machine Learns and Fails – A Grammar of Error for Artificial Intelligence', *Spheres journal for digital culture* [Preprint]. Available at: https://spheres-journal.org/contribution/how-a-machine-learns-and-fails-a-grammar-of-error-for-artificial-intelligence/ (Accessed: 12 March 2024).

Pinker, S. (2018) *Enlightenment now: the case for reason, science, humanism, and progress*. New York, New York: Viking, an imprint of Penguin Random House LLC.

Postman, N. (1993) *Technopoly: the surrender of culture to technology*. 1st Vintage Books ed. New York: Vintage Books.

Ramstead, M.J.D., Veissière, S.P.L. and Kirmayer, L.J. (2016) 'Cultural Affordances: Scaffolding Local Worlds Through Shared

Intentionality and Regimes of Attention', *Frontiers in Psychology*, 7. Available at: https://doi.org/10.3389/fpsyg.2016.01090.

Rieff, P. (2006) *The triumph of the therapeutic: uses of faith after Freud*. 40th anniversary ed. Wilmington (Del.): ISI books.

Rockman, B. (2024) 'Bureaucracy', *Encyclopaedia Britannica*. Available at: https://www.britannica.com/topic/bureaucracy (Accessed: 5 October 2024).

Russell, S.J. (2020) *Human compatible: artificial intelligence and the problem of control*. [London] New York NY: Penguin Books.

Sahlins, M.D. (1972) *Stone age economics*. Chicago: Aldine-Atherton.

Schlosser, M. (2019) *'Agency', The Stanford Encyclopedia of Philosophy*. Winter 2019. Metaphysics Research Lab, Stanford University. Available at: https://plato.stanford.edu/archives/win2019/entries/agency/ (Accessed: 4 March 2024).

Smith, W. (2016) *Culture of Death: The Age of Do Harm Medicine*. Revised Edition. New York, NY: Encounter Books.

Susskind, D. (2020) *A world without work: technology, automation, and how we should respond*. London: Allen Lane, an imprint of Penguin Books.

Taves, A., Asprem, E. and Ihm, E. (2018) 'Psychology, meaning making, and the study of worldviews: Beyond religion and non-religion.', *Psychology of Religion and Spirituality*, 10(3), pp. 207–217. Available at: https://doi.org/10.1037/rel0000201.

Taylor, C. (2018) *A secular age*. First Harvard University Press paperback edition. Cambridge, Massachusetts London, England: The Belknap Press of Harvard University Press.

Tegmark, M. (2018) *Life 3.0: being human in the age of artificial intelligence*. [London] UK [New York, NY] USA [Toronto] Canada: Penguin Books (An Allen Lane book).

The Editors of Encyclopaedia Britannica (1998) ‘Cybernetics’, *Encyclopaedia Britannica*. Available at: https://www.britannica.com/science/cybernetics (Accessed: 13 July 2024).

Thiselton, A.C. (2009) *Hermeneutics: an introduction*. Grand Rapids, Mich: W.B. Eerdmans Pub. Co.

Thomson, A. (2009) *Critical reasoning: a practical introduction*. 3rd ed. London ; New York: Routledge.

Turkle, S. (2011) *Alone together: why we expect more from technology and less from each other*. New York: Basic books.

Turkle, S. (2015) *Reclaiming conversation: the power of talk in a digital age*. New York: Penguin press.

Vedung, E. (1977) *Det rationella politiska samtalet: hur politiska budskap tolkas, ordnas och prövas*. Stockholm: Aldus/Bonnier.

Verbeek, P.-P. (2005) *What things do: Philosophical reflections on technology, agency, and design*. University Park, Pa: Pennsylvania State University Press.

Vervaeke, J. and Ferraro, L. (2013) ‘Relevance Realization and the Neurodynamics and Neuroconnectivity of General Intelligence’, in I. Harvey et al. (eds) *SmartData*. New York, NY: Springer New York, pp. 57–68. Available at: https://doi.org/10.1007/978-1-4614-6409-9_6.

Vervaeke, J., Mastropietro, C. and Miscevic, F. (2017) *Zombies in Western culture: a twenty-first century crisis*. Cambridge, UK: Open Book Publishers.

Wallerstein, I. (2011) *The modern world-system*. Berkeley: University of California Press (The Modern world-system, 4).

Walsh, B. (2006) ‘From housing to homemaking: Worldviews and the shaping of home.’, *Christian Scholar’s Review*, 35(2), pp. 237–257.

Weber, M. *et al.* (2012) *The Protestant ethic and the ‘spirit’ of capitalism and other writings*. 23. print. New York: Penguin Books (Penguin twentieth-century classics).

Weizenbaum, J. (2008) ‘Social and Political Impact of the Long-term History of Computing’, *IEEE Annals of the History of Computing*, 30(3), pp. 40–42. Available at: https://doi.org/10.1109/MAHC.2008.58.

Wiener, N. (1988) *The human use of human beings: cybernetics and society*. New York, N.Y: Da Capo Press (The Da Capo series in science).

Wiener, N. (1990) *God and Golem, Inc: a comment on certain points where cybernetics impinges on religion*. 7. pr. Cambridge: M.I.T. Pr.

Wiener, N. (1999) ‘Some moral and technical consequences of automation’, *Resonance*, 4(1), pp. 80–88. Available at: https://doi.org/10.1007/BF02837160.

Zuboff, S. (2019) *The age of surveillance capitalism: the fight for a human future at the new frontier of power*. London: Profile books.

ACTA UNIVERSITATIS UPSALIENSIS
Uppsala Studies in Philosophy of Religion
Editor: Mikael Stenmark

1. Maria Klasson Sundin, Barnets religionsfrihet – en villkorad rättighet? En filosofisk undersökning utifrån barnkonventionen. 2016.
2. Lotta Knutsson Bråkenhielm, Religion – evolutionens missfoster eller kärleksbarn? Kognitionsvetenskaplig religionsforskning och dess relevans för religiösa trosföreställningars rationalitet. 2016.
3. Johan Eddebo, Death and the Self: A Metaphysical Investigation of the Rationality of Afterlife Beliefs in the Contemporary Intellectual Climate. 2017.
4. Oliver Li, Panentheism, Panpsychism and Neuroscience: In Search of an Alternative Metaphysical Framework in Relation to Neuroscience, Consciousness, Free Will, and Theistic Beliefs. 2018.
5. Filosofiska metoder i praktiken. Mikael Stenmark, Karin Johannesson, Francis Jonbäck & Ulf Zackariasson (red.). 2018.
6. Mikael Sörhuus, En känsla för det heliga: En undersökning av den samtida emotionsforskningens möjliga bidrag till religionsfilosofin. 2020.
7. Ingrid Malm Lindberg, The multifaceted role of imagination in science and religion: A critical examination of its epistemic, creative and meaning-making functions. 2021.
8. Lina Langby, God and the world: Pragmatic and epistemic arguments for panentheistic and pantheistic conceptions of the God–world relationship. 2023.
9. Johan Martícki, Human agency and autonomy in algorithm-intense environments. 2025.

***

Prior to 2002, studies in philosophy of religion from Uppsala university were published in a joint series with Lund university: Studia Philosophiae Religionis. The following books were published in this series:

1. Ulf Hanson, ”Religious experience”: en semantisk studie. 1973.
2. Carl-Reinhold Bråkenhielm, How philosophy shapes theories of religion: an analysis of contemporary philosophies of religion with special regard to the thought of John Wilson, John Hick and D. Z. Phillips. 1975.
3. Bo Hanson, Application of rules in new situations: a hermeneutical study. 1977.
4. Lars Haikola, Religion as language-game: a critical study with special regard to D. Z. Phillips. 1977.
5. Ulf Görman, A good God?: a logical and semantical analysis of the problem of evil. 1977.
6. Eberhard Herrmann, Der religionsphilosophische Standpunkt Bernard Bolzanos unter Berücksichtigung seiner Semantik, Wissenschaftstheorie und Moralphilosophie. 1977.
7. Eberhard Herrmann, Die logische Stellung des ontologischen Gottesbeweises in Charles Hartshornes Prozesstheologie und neoklassischer Metaphysik. 1980.

8. António Barbosa da Silva, The phenomenology of religion as a philosophical problem: an analysis of the theoretical background of the phenomenology of religion, in general, and of M. Eliade´s phenomenological approach, in particular. 1982.
9. John A. Sealey, Religion in schools: a philosophical examination. 1982.
10. Jan Löfberg, Spiritual or human value?: an evaluation-systematical reconstruction and analysis of the preaching of Jesus in the synoptical gospels. 1982.
11. Håkan Thorsén, Peak-Experience, religion and knowledge: a philosophical inquiry into some main themes in the writings of Abraham H. Maslow. 1983.
12. Eberhard Herrmann: Erkenntnisansprüche: eine orientierende Erkenntnistheoretische Untersuchung über Fragen zum Verhältnis zwischen Religion und Wissenschaft. 1984.
13. Christer Norrman, Mystical experiences and scientific method: a study of the possibility of identifying a ”mystical” experience by a scientific method, with special reference to the theory of Walter T. Stace. 1986.
14. Jari Ristiniemi, Experiential dialectics: an inquiry into the epistemological status and the methodological role of the experiential core in Paul Tillich’s systematic thought. 1987.
15. Olof Franck, The criteriologic problem: a critical study with special regard to theories presented by Anthony Flew, D. Z. Phillips, John Hick, Basil Mitchell, Anders Jeffner and Hans Hof. 1989.
16. Martin Holmberg, Narrative, transcendence and meaning: an essay on the question about the meaning of life. 1994.
17. Anders Nordgren, evolutionary thinking: an analysis of rationality, morality and religion from an evolutionary perspective. 1994.
18. Stefan Andersson, In quest of certainty: Bertrand Russell’s search for certainty in religion and mathematics up to The Principles of Mathematics (1903). 1994
19. Dag Hedin, Phenomenology and the making of the world. 1997.
20. Olof Franck, Tro och transcendens i Ulf Ekmans och Kristina Wennergrens författarskap: om teologisk realism och referentiell identifikationsteori i två samtida trosuppfattningar. 1998.
21. Ulf Zackariasson, Forces by which we live: religion and religious experience from the perspective of a pragmatic philosophical anthropology. 2002.
22. Olof Franck, Förtryckets grundvalar: norm och avvikelse i argument om homosexuellas, invandrares och kvinnors rättigheter. 2002.
23. Kirsten Grönlien Zetterqvist, Att vara kroppssubjekt: ett fenomenologiskt bidrag till feministisk teori och religionsfilosofi. 2002.