Julian Nida-Rümelin
Nathalie Weidenfeld

Digital Humanism

For a Humane Transformation of Democracy, Economy and Culture in the Digital Age

OPEN ACCESS

Springer

Digital Humanism

Julian Nida-Rümelin • Nathalie Weidenfeld

Digital Humanism

For a Humane Transformation of Democracy, Economy and Culture in the Digital Age

Springer

Julian Nida-Rümelin
Ludwig-Maximilians-Universität München
Munich, Germany

Nathalie Weidenfeld
Munich, Germany

ISBN 978-3-031-12481-5 ISBN 978-3-031-12482-2 (eBook)
https://doi.org/10.1007/978-3-031-12482-2

© The Editor(s) (if applicable) and The Author(s) 2022. This book is an open access publication.
Translation from the German language edition: "Digitaler Humanismus" by Julian Nida-Rümelin and Nathalie Weidenfeld, © Piper Verlag GmbH 2018. Published by Piper Verlag GmbH. All Rights Reserved.
Open Access This book is licensed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits any noncommercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if you modified the licensed material. You do not have permission under this license to share adapted material derived from this book or parts of it.
The images or other third party material in this book are included in the book's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.
This work is subject to copyright. All commercial rights are reserved by the author(s), whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Regarding these commercial rights a non-exclusive license has been granted to the publisher.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Cover Illustration: © Ilana Lewitan "Manche gehen in den Wald und sehen nirgendwo Holz", 2017. Reused with Permission.

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

Preface

It seems that there are basically two strategies of digital transformation. The first is market driven and consumer oriented. The prevailing business models are based on marketing. The strategy is accompanied by the idea to make the world a better place offering applications of communication and interaction.

The second strategy is focused on state interest, controlling citizens and stabilizing autocratic power. For each of these strategies, one of the two superpowers is in lead: the USA and China. But possibly there is a third strategy, opposed to the commercial and the state model alike. Some documents and decision of the European community lead in this this direction: human-centered AI. Until now there is no clear-cut idea of such a third, more humane strategy of digital transformation.

In this book, we develop the basic ideas of a different view, that we call "digital humanism."

Julian has used that term for many years in talks and discussions, leading to controversy but also arousing interest in the public and political sphere. We therefore decided to write a book together in order to expose its basic ideas.

We—that is Julian Nida-Rümelin, a philosopher and former state minister for culture, and Nathalie Weidenfeld, a film theorist and author.

The German version of the book appeared in 2018 and was awarded the Bruno Kreisky prize as the best political book of the year 2018.

In the meantime, the term of "digital humanism" is widely used, especially in Austria, Italy, and Spain. In Austria, the group around Prof Hannes Werthner has called into life the "digital humanism initiative" which brings together renowned international researchers and offers an impressive lecture

series[1] as well as important publications.[2] In Spain, the government supports different organizations to help entrepreneurs understand and support a "human-centered" digital transformation.[3] In Italy, Franscesca Bria, the president of the Italian Innovation Fund, speaks of "digital humanism" as an alternative to "digital capitalism." In Italy, the term "digital humanism" has become quite popular, also due to the fact that this book appeared in Italian in 2019.

The EU has come forward with a vision of a digital transformation which needs to be "serving humans."[4] In the USA, the term is beginning to gain importance, mostly in the context of how economy can use and might profit from a digital transformation which "centers around human experience."[5] Other voices however, such as Tristan Harris, the former Google employee, focus more on the negative effects of an economically oriented digital transformation and promote a "humane technology" based on trust.

In this book, we describe the philosophical and cultural aspects of digital humanism. It can be understood as its groundwork.

Munich, Germany

Julian Nida-Rümelin
Nathalie Weidenfeld

[1] https://dighum.ec.tuwien.ac.at/dighum-lectures/ (last accessed on April 29 2022).

[2] Hannes Werthner, Erich Prem, Edward A. Lee, Carlo Ghezzi. 2022. *Perspectives on Digital Humanism*. New York: Springer.

[3] For more information on the Spanish involvement, see https://espanadigital.gob.es/sites/agendadigital/files/2022-03/E10M46_Fostering_Digital_Humanism.pdf (last accessed on April 29th 2022).

[4] For further information, see the website of the EU https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/shaping-europe-digital-future_de (last accessed on April 29th 2022).

[5] https://www.trackmind.com/digital-humanism/

Contents

1 Introduction 1

2 Robots as New (Digital) Slaves 7

3 Digital Simulations of Emotions 13

4 The Problem of Autonomy and Determination in the Digital World 19

5 The World as the Perfect Machine Universe 25

6 Digital Optimization, Utilitarianism, and AI 31

7 Economic Rationality as a Software Program 35

8 Why Robots Don't Have Moral Judgment 41

9 Ethical Non-comparability 47

10 Why AIs Fail at Moral Dilemmas 53

11 Why AIs Can't Think 57

12 Digital Virtualities and Sober Realities 65

13 On the Ethics of Internet Communication 69

14 On the Ethics of Communication Between Humans and AI 75

15 Cultural Aspects of Digitalization 81

16 Digital Education 87

17 Utopia of Liquid Democracy 97

18 Socio-economic Aspects of Digitalization 105

19 Transhumanist Temptations 111

20 On the Metaphysics of Digitalization 117

21 Afterword 121

References 125

About the Authors

Julian Nida-Rümelin is a philosopher and well-known public intellectual in Europe. He teaches Philosophy and Political Theory at the University of Munich. He is a member of the European Academy of Sciences. He served as state minister for culture and media in the German government.

Presently he is a board member of the Bavarian Research Institute for Digital Transformation and vice president of the German Ethics Council. He taught philosophy in the USA (State University of Minnesota, California Institute of Technology) and Italy (Calgiari, Trieste, Torino, Milano). He was awarded an Honorary Doctorate by the University Trieste, Italy. Since 2001, he is also Honorary Professor at Humboldt University in Berlin.

Nida-Rümelin publishes inside and outside academia on topics including economics, democracy, education, ethics as well as the digitalization of our society. The German version of this book received the Bruno Kreisky prize as the best political book of the year 2018. Nida-Rümelin's specific and most original account of practical philosophy that combines pragmatism, rational choice, and morality is presented in *A Theory of Practical Reason* (Palgrave Macmillan 2023).

Nathalie Weidenfeld was born in France and has lived mostly in Germany and partly in the USA. She is a film theorist and author of fictional and non-fictional books. She received her doctoral degree in American Culture Studies at the University in Berlin and teaches creative writing and film studies. She has three children with her husband Julian and lives with her family in Munich.

1

Introduction

It is possible that in the distant future we will look back at human history and speak of three major disruptive technological innovations. The transition from a hunter-gatherer culture to a sedentary agrarian culture with animal husbandry in the Neolithic Age, the transition to the machine age based on fossil fuels in the nineteenth century, and finally the digital revolution of the twenty-first century: the use of artificial Intelligence[1]. If this is the case, we are only at the beginning of a technological revolution, similar to what Europe experienced in the first decades of the nineteenth century. And just as then, technological innovations today are accompanied by apocalyptic fears, but also by euphoric expectations.

This book deals with the cultural and philosophical aspects of Artificial Intelligence and pleads for a digital humanism. Digital Humanism is technology-friendly, but also human-friendly. It sets itself apart from the apocalyptics because it trusts human reason, but it also sets itself apart from the euphorics because it respects the systematic limits of digital technology.

The dream of the creation of artificial beings has been part of mythological narratives for thousands of years. In antiquity, it is the myth of Prometheus, a god from the Titan family, who creates thinking and feeling clay beings without divine permission and is bitterly punished by Zeus for it. In the Middle Ages, we find the story of the Golem, an artificial being made of clay, which is mute and not capable of reason, but possesses great strength and can carry out orders. Literature also uses the myth of the artificially created being. In the story "The Sandman" by E. T. A. Hofmann (1816) the protagonist falls deeply

[1] If we write "artificial intelligence" we implicitly accept the existence of artificial intelligence. If we leave it open we should capitalize "Artificial Intelligence" which we do in this book.

© The Author(s) 2022

J. Nida-Rümelin, N. Weidenfeld, *Digital Humanism*,
https://doi.org/10.1007/978-3-031-12482-2_1

in love with an animated doll named Olimpia and in the course of events ultimately loses his mind over it. Perhaps the most famous example from this period is Mary Shelley's novel *Frankenstein* or *The Modern Prometheus* (1818). In this tragic story, a Swiss scientist creates an artificial human. This artificial man arouses so much disgust and fear due to his size and ugliness that he cannot connect with human society and, on the contrary, accumulates more and more rage and hatred within himself. In the end, he kills the bride of his creator and himself.

Today we might call our contemporary humanoid robots "frankensteins" had there not been the play *R.U.R.* by Czech writer Karel Čapek in 1920. This drama is about a company called R.U.R. (Rossum's Universal Robots), which produces artificial humans called "robots" and abuses them as cheap laborers, who, however, in the course of the story rebel against their slavery and wipe out humanity.

In the twentieth and twenty-first century, the robots live mostly in sci-fi novels such as the ones by Stanisław Lem[2] or of the US-American author Philip K. Dick.[3] In recent years, US-American sci-fi blockbuster films have heavily drawn on the mythological figure of the artificial human, which now appears as a robot that cooperates with humans on earth and on spaceships.

Apart from these, there is also the idea of a fully digitalized world which sci-fi films and novels have taken up. The vision is almost always dystopian: there are worlds completely dominated by machines like in the film *The Matrix* (directed by Wachowksis, USA, 1999) or futuristic nightmarish societies such as the one in the film *Demolition Man* (directed by Marco Brambilla, USA, 1993), in which people act and interact based on digital instructions and even sexual contact may only take place through the mediation of digital media.

In the meantime, many things which were fantasized about in the history of mankind have become reality, the most famous example being Captain Kirk's foldable "communicator" from *Star Trek*, which was technologically realized some 50 years later in the form of the StarTAC mobile phone by Motorola.

It even seems that the myths merely take on a form impregnated by new technologies but remain unchanged at their core. The myth of the machine in human form that takes over in the end, the myth of the animated doll, the myth of a friendship between man and machine. But unlike previous centuries, these myths now appear to be revitalized by concrete technological options.

---

[2] See, for example, *The Star Diaries* (1976) or *Golem XIV* (1985).

[3] The US author Philip K. Dick wrote numerous books and short stories on which many US films such as *Blade Runner, Minority Report*, or *Total Recall* are based.

There can be no doubt, we are living in a time of technological upheaval. This century and the next—many are convinced—will be the age when robots will have taken over many kinds of human work. They will deliver parcels, produce goods in factories, drive taxis, act as bank advisors, explore space, work in call centers, operate alongside doctors in hospitals, and possibly write novels and create art. But one need not to look so far ahead. Digitalization has already permeated not only our working world but also our private lives and has had a great influence on our cultural, political, economic, and social life.

This development raises many questions as to what the consequences will be for us. Some, such as bestselling authors Daniel H. Wilson (*Robopocalypse*, 2011) (a former research associate at Carnegie Mellon University who earned a PhD in robotics) or Stephen Hawking[4] or philosophers such as Nick Bostrom[5] warn us that robots will one day surpass the human species in thinking and acting abilities and turn against humanity.[6] Others, harbor utopian hopes for a new, digital world in which digital robots as modern slaves perform human work and establish a realm of unprecedented freedom for us.

There is much to suggest that what is called "strong AI" in the digitalization discourse, i.e., the thesis that software systems have consciousness, make decisions, pursue goals, that their performances are not merely simulations of human abilities but realize them, will one day be considered a form of modern animism, i.e., the ensoulment of the non-ensouled, which can be seen as a regression into childlike modes of interpretation.

Of course, such a digitalization ideology does not present itself as regressive and childish, but on the contrary as rational and scientific. It has a long cultural history. It begins in our cultural sphere with the Pythagoreans in the fifth century BC. It is the idea of a world strictly ordered in numerical relations, the harmony and rationality of which is only revealed in mathematical analysis. Two hundred years later, the Stoic philosophers added to this theory the thesis of the correspondence between world reason and human reason (logos). According to this theory, people are only able to think and act rationally because they can participate in world reason. The logos orders the world

[4] Stephen Hawking warns in many interviews against unrestrained use of Artificial Intelligence. For example, in the Focus interview in 2015: "Our future is a race between the growing power of technology and the wisdom with which we use it." Online at: http://www.focus.de/wissen/technik/wird-man-sie-kontrollieren-koennen-stephen-hawking-warnt-in-100-jahren-sind-computer-intelligenter-als-menschen_id_4681638.html (last accessed on 16 April 2017).

[5] Bostrom (2014).

[6] German scientists such as the German philosopher Thomas Metzinger also warn of the negative effects of an "AI arms race," at the end of which super-intelligent software could emerge that detaches itself from its computer and, like a large, uncontrollable virus, activates itself in ever new places and globalizes itself and its targets.

according to strict deterministic laws and human beings have to fit into this world reason. Even the Stoics and their opponents noticed, however, that there is a tension between a world view of comprehensive determinism and a view of oneself as a free and responsible human agent. If AI ideology leads to a new edition of this conflict, then digital humanism overcomes this conflict.

In this book, we develop the main features of a digital humanism as an alternative to what can somewhat simplistically be called "Silicon Valley ideology." Silicon Valley ideology related to the original American, Puritan hope of salvation, of creating a world of the pure and righteous who have left filth and sin behind. Its central values are transparency and predictability, economic success and patronage. In times of digital transformation this included the dream of a perfectly constructed digital counterparts whose construction excludes any error leading us into a technological utopia. The key concept here is that of Artificial Intelligence, charged with implicit metaphysics and theology, a self-improving, hyper-rational, increasingly ensouled system whose creator, however, is not God but software engineers who see themselves not merely as part of an industry but of an overarching movement realizing a digital paradise on earth based on transparency, all-connectedness, and non-ambiguity.

The Silicon Valley ideology takes humanist impulses as its starting point, only to transform them into anti-humanist utopias. It begins with the improvement of the human and ends in its final—and inhumane—overcoming. By wanting to improve human life on the planet, it starts to question the conditions of humanity. In the course of this, humanism is thus being transformed into transhumanism, leading to a technicist utopia in which the human is left behind. Digital humanism opposes this and offers instead a new ethics for the age of Artificial Intelligence.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits any noncommercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if you modified the licensed material. You do not have permission under this license to share adapted material derived from this chapter or parts of it.

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

2

Robots as New (Digital) Slaves

Early morning somewhere in a future US-American city. Detective Spooner, a cool guy in a leather coat, Allstars, and a baseball cap, gets ready to go to work. As he opens the door, he flinches. In front of him there is a metal humanoid FedEx robot with a parcel under its arm.

"Good morning, Sir," the robot greets him politely. "Yet another on time delivery from…" But that's as far as he gets.

"Get the hell out of my face, cannery," Spooner tells him while pushing him to the side. The robot looks at him, seemingly confused, but wishes him a nice day anyway.

This is the year 2035. Robots are not only used in factories but also in private households. They walk alongside people on the street, take out the garbage, do the shopping, and walk their owners' dogs. At least, that's how the world looks like in the movie *I, Robot* (Alex Proyas. USA, 2004). These robots are presented to us as submissive servants who are not treated particularly well. When they are bumped into, they are the ones who apologize. Their status is that of slaves, whose only purpose is to be used by humans. At the beginning of the film, the following sentences are projected onto the screen:

1. A robot shall not harm a human being or, through inaction, allow a human to come to harm.
2. A robot must obey the orders given to it by humans except for those orders that conflict with the First Law.

© The Author(s) 2022

J. Nida-Rümelin, N. Weidenfeld, *Digital Humanism*,
https://doi.org/10.1007/978-3-031-12482-2_2

3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.[1]

These laws, which are a direct quote from the story "Runaround" by Isaac Asimov, make it quite clear how robots are supposed to function.

Spooner, the protagonist of the film, feels nothing but contempt for robots. For him they are under general suspicion. When a theft occurs in the city, he always suspects robots first, not humans.

Not only Spooner but society as a whole has little compassion for its mechanical slaves. Once they are no longer needed and have reached the end of their usefulness, they are disposed of on the outskirts of the city and put into containers, where they must spend the rest of their probably eternal, digital existence. There they stand, nestled closely together as if they wanted to comfort each other. The robots' "faces" reflect a kind of noble capacity for suffering. They are sad robots. Robots that don't understand why they are treated so badly. Proyas wants us to come to the conclusion that discriminatory treatment of robots is unjust and inhumane.

In reality, however, no one has yet come up with the idea of applying the norms of the Animal Welfare Act to robots, or even of granting them human rights. There is a practical consensus that computers and robots have no mental states. We agree that robots unlike animals—to which the capacity to suffer is attributed—are not sentient. So far, there has been no serious initiative to grant rights to computers or software systems based on their sentience.

There is nothing to suggests that even today's most complex software systems possess consciousness. If they did, we would have to strictly regulate their further use with immediate effect and attribute fundamental and human rights to them. Painless killing, which is permissible regarding animals but ethically and legally impermissible regarding humans, would then be prohibited. In analogy to the *Great Ape Project*, which wanted to overcome speciesism and grant humanlike animals human rights to the extent that they have comparable characteristics, robots and autonomous software systems would also have to be granted human rights. If we assume that robots created by us are personal beings that are endowed with an identity, responsibility for action, autonomy, and the accompanying individual dignity (a so-called e-person (electronic person)),[2] the software systems in question could then no longer be manipulated in analogy to the right of informational

[1] Later Asimov added a "zeroth law" which states: A robot may not harm humanity or allow it to come to harm by inaction.

[2] Such an "E-personhood" has actually been proposed in a draft report by the EU parliament in 2017.

self-determination[3] of human individuals, because this would contradict the Kantian principle of non-instrumentalization of rational beings.

And yet some proponents of Artificial Intelligence claim that it is fundamentally impossible to distinguish between a human brain and a computer. Thus, lawyers and sociologists are increasingly concerned with the question of the extent to which (future) robots can be held liable in the event of errors, i.e., have a juridical responsibility. In international research institutions, lawyers are asking whether robots are to be regarded as mere tools for which their owners or manufacturers must be liable or, whether, depending on their degree of autonomy, they will at some point enjoy a special status that grants them responsibility and rights. After all, according to the legal argument here, robots would also have duties to fulfill.

In October 2016 in Saudi Arabia, a robot was officially granted citizenship for the first time in history. The robot in question was "Sophia," an android robot with a female face and body that mechanically simulates facial expressions. The citizenship theoretically not only gives Sophia rights, but also duties. The very fact that she is allowed to move about unveiled—unlike all other Saudi Arabian women—caused much discussion in Saudi Arabia and beyond.

In *I, Robot*, robots have a lot of duties. If they don't fulfill them, they are prosecuted just like humans. But then, by implication, shouldn't they also have rights like humans? At least that is the foundation of ethics and law in civil and democratic societies.

Just as in *I, Robot* the film *A.I. Artificial Intelligence* (Steven Spielberg. USA, 2001) imagines a future where robots have become a normal part of our everyday life. They are slaves and service providers. Sad service providers, one might add, as they are presented to us as sentient beings who suffer from being treated as second- or even third-class humans. Spielberg makes his position quite clear by using melodramatic means to make to the viewer believe that in the near future it will be essential to give robots not only legal rights but grant them also especially the right to (human) dignity.

Anyone who takes Spielberg's idea that a robot has the same dignity as a human being seriously must assume indistinguishability between humans and computers or software systems. But anyone who thinks that there can be no categorical difference between human brains and computers is denying the foundations not only of scientific practice but of the human way of life in

[3] The term was first coined in Germany by the German Federal Constitutional Court in 1983 when the court dismissed a law passed by the German Parliament on census and describes the right of every individual to control relevant personal informations.

general. Whoever resents his PC because it proved to be disobedient has a problem of rationality and reality. He is attributing properties to his computer that it does not have. Only in philosophy seminars or the indistinguishability of humans and machines can be asserted. Outside, this assertion seems grotesque, as it is incompatible with the actual practice of those who advocate it. Of course, we turn off our computers when we no longer need them, we dispose of them in the junkyard, without shedding a tear. The computer is not an Other, but a tool, far more complex than a shovel, far surpassing some human capabilities, but still just a physically describable apparatus without desires or beliefs. With this in mind, we should not strive to make robots as human-like as possible.

In one of the most emotional scenes of Spielberg's film *A.I.*, we see how discarded robots are brought to a kind of circus arena. Under the eyes of a roaring crowd, they are put in a cannon and shot into the air. "But I still function perfectly," one robot protests in despair as he is led away into the arena. Obviously, the robots don't want to die. But the drunken crowd has no sympathy. To them, robots are just an accumulation of metal. To the viewer, however, the robots are presented as sentient beings, who suffer from wrong and inhumane treatment. Just because robots are machines—that's the message of the film—doesn't mean they are worth less than humans: they have the same dignity.

In philosophy, it is quite controversial what constitutes human dignity. Some believe that it is the special sensitivity and the capacity for suffering that demands special consideration. Others believe that human beings have (basic) rights by nature—or by God—which are inalienable and which constitute the special dignity of human beings. Those who stand in the tradition of Immanuel Kant base dignity on the autonomy that is inherent in human beings. Accordingly, it is the human capacity to weigh up reasons that makes humans autonomous agents and gives them the special status as beings who have dignity.[4]

In his book *The Decent Society*, the Israeli philosopher Avishai Margalit has placed human dignity and self-respect at the center: We must not treat anyone in such a way that he has reason to feel humiliated and harmed in his self-respect. Artificial Intelligence have no self-respect, no feelings that we can hurt. Their personal identity is not vulnerable and they do not have the ability to reflect on their life. The preconditions for ascribing dignity to them are not fulfilled.

[4] For a detailed critique of utilitarianism, see Nida-Rümelin (2023), Chapter 5.

Since human dignity and human rights are so central to our very understanding of ourselves, but also to the legal and political order in which we live, we should be careful not to jeopardize this core of human ethos by overextending it. Populating the world with Artificial Intelligence, to which we attribute abilities and characteristics comparable to those of humans, would inevitably lead to the destruction of this ethos. Seen in this light, it makes more sense to read Spielberg's oppressed robots as a metaphor for the treatment of African American slaves in history than as a realistic depiction of an abusive treatment of robots.

At the end of *A.I.* there are no humans left on earth. Not a big loss, as it seems to the viewer, since he only got to know cold-hearted human beings throughout the film. The only beings who showed compassion in *A.I.* were robots. Robots that have been oppressed and abused. At the end of his long story of suffering, the protagonist David, the little robot, is finally redeemed by angel-like alien robots who have come to Earth. He, who has longed for the love of his long-deceased human mother all his life, is now given the opportunity to be reunited with her, as the alien robots bring her back to life through a DNA reconstruction. At last David can be happy. Although this bliss will only last one day (as the reconstruction can't survive longer than 24 hours), his wounds can now heal. The film thus joins the ranks of melodramatic Christian-influenced narratives of the nineteenth century such as the novel *Uncle Tom's Cabin* (Harriet Beecher Stowe, 1852) where the Afro-American protagonist Uncle Tom must endure great hardship, suffering, and even death in order to receive salvation (and make the readers understand that racism is bad). Seen from this perspective, Spielberg's *A.I.* needs to be read not as a realistic and serious assessment of the status of robots, but as perpetuating the Christian narrative of suffering and resurrection and as a metaphorical comment on racism.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits any noncommercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if you modified the licensed material. You do not have permission under this license to share adapted material derived from this chapter or parts of it.

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

3

Digital Simulations of Emotions

A young blond man with freckles steps out of a helicopter onto a beautiful island. Lush vegetation, gentle streams, cascading water. After walking through a dense forest for a while, he finally arrives at a hyper-modern house equipped with maximum-security systems. The house (and the whole island in fact) belong to Nathan. He is the inventor and founder of the largest Internet search engine in the world called "Blue Book" (an allusion to the famous book by philosopher Ludwig Wittgenstein). Nathan is an ingenious and arrogant programmer who has set himself the goal of creating a new species: a robot capable of consciousness. Caleb, the young man with freckles, works in Nathan's company and was chosen by Nathan to test whether one of his first robots has reached this goal.

"Do you know what the Turing Test is?" Nathan asks him shortly after his arrival.

"Yeah," Caleb replies. "I know what the Turing Test is. It's when a human interacts with a computer. And if the human doesn't know they're interacting with a computer, the test is passed."

"And what does a pass tell us?"

"That the computer has artificial intelligence."

The robot Caleb is supposed to test is Ava, an attractive robot woman. Her face resembles that of a young woman. Only her legs and arms are made of shiny metal, and blue wires glow in her belly. When she moves, there is a soft hissing sound, as if a neon tube is humming. In various sessions, Caleb watches Ava through a pane of bulletproof glass. Through the PA system, Caleb talks to her, asks her questions, tests her. Like an enigmatic sphinx, she

© The Author(s) 2022
J. Nida-Rümelin, N. Weidenfeld, *Digital Humanism*,
https://doi.org/10.1007/978-3-031-12482-2_3

sits across from him and answers all his questions—like a real, self-aware human being. After a while, however, Ava begins to turn things around. Now it is she who starts asking Caleb questions. Looking at her face, Caleb can detect many emotions in her. She is surprised, sometimes flattered, sometimes puzzled, sometimes hurt, and finally in love. And yet, Ava is a machine. So how can she possibly have emotions?

Nathan will soon give Caleb the explanation:

"If you knew the trouble getting AI to read and duplicate facial expressions. You know how I cracked it?"

"I don't know how you did any of this," Caleb replies perplexed.

"Every cell phone just about has a microphone, camera, and a means to transmit data. So I turned on all every microphone and camera across the entire f***ing planet, and I redirected the data through Blue Book. Boom. A limitless resource of vocal and facial interaction."

Ava is an expert in facial and vocal expressions. By observing all the people in the world and their reactions, she has acquired a perfect reservoir of knowledge about facial expressions over time. She knows how to interpret facial expressions and she knows what facial expressions are considered appropriate at what time. Big Data makes her a perfect imitator of emotional expressions. But does that mean she really *has* feelings?

"I want to be with you. [...] Do you want to be with me?" Ava asks Caleb in the fifth session.

Caleb, too, would like to know if Ava really has feelings for him or has just been programmed to pretend to do so. Eventually, Caleb decides to believe her. He regards her as an autonomous and unique being. A being he falls in love with and assumes has fallen in love with him as well.

In another session, Caleb tells Ava about the thought experiment "Mary's Room". This thought experiment really exists. It was put forward by the Australian philosopher Frank Cameron Jackson in his essay "What Mary didn't know" (1986).

"Mary is a scientist, and her specialist subject is color. She knows everything there is to know about it, the wavelengths, the neurological effects, every possible property color can have. But she lives in a black and white room. She was born there and raised there and she can observe the outside world on a black and white monitor. One day, someone opens the door, and Mary walks out. And she sees a blue sky. And at that moment, she learns something that all her studies couldn't tell her. She learns what it feels like to see color."

Ava looks at Caleb motionless. Judging by the expression on her face, this story is taking a toll on Ava. This isn't surprising. After all, isn't she just like

Mary? A person who knows everything but only from second-hand information from the Internet? In Ava's face, Caleb reads disappointment, but also a fierce determination. She makes it clear to Caleb that she too wants to leave her room one day. Preferably—so she tells him—with him. On their first date, she tells him about her biggest dream: Standing at a busy intersection, watching the people go by.

When she finds out that Nathan plans to switch her off soon to recycle parts of her for a new robot, she is determined to do everything she can to escape. Caleb wants to help her and comes up with a plan.

By the end of the film, Caleb has managed to break the code of the maximum-security wing. Ava escapes. Shortly after, Ava kills Nathan, her creator. Nothing stands between her and her freedom anymore. But then something happens that neither Caleb nor the viewer expected at this point: Ava cold-heartedly leaves Caleb behind, locked up in a room. The viewer is also shocked at this moment, because like Caleb he has gotten the feeling in the course of the film that Ava is a sentient being who not only suffers from her situation but has also fallen in love with Caleb.

As Caleb desperately pounds on the door which will keep him inside the house until he'll starve to death, she walks through the house in a white dress and white shoes like an elf. With organic material taken from other deactivated robots, she now walks out into the world. Her brown shoulder length hair caresses her delicate face. As she breathes in the air of the forest for the first time, she smiles. She touches branches and curiously looks at her new life. She feels no remorse and does not even look back.

Like Mary, she now steps out of her room into the big wide world, ready to have her own experiences. Will she learn to not only imitate emotions but also to have them? Or will she remain a machine forever? This is the essence of all philosophical questions around which AI enthusiasts keep circling.

Caleb, too, keeps asking himself the question: Has Ava only learned to imitate certain behaviors in order to give the false impression that she has feelings much like the "cold" actor described by Diderot, whose art focuses primarily on the perfect mastery of physical behavior? The truly troubling question, however, is the following: What if not only Ava's but also our feelings were really nothing more than just pure behavior? That, at least, is what radical positivists claim, advocating the metaphysical thesis that mental states are nothing but patterns of behavior. A positivist's understanding of consciousness identifies mental properties and states, such as being afraid or having desires or beliefs, with particular behaviors. "Jacob is in pain" means—in the positivist's understanding—nothing other than "Jacob behaves in a

certain way, for example, he cries 'ouch' or jerkily withdraws his hand from the stovetop."

It is not a coincidence, by the way, that the film refers to the philosopher Ludwig Wittgenstein several times (once with the name "Blue Book," which is both the name of Nathan's company and the title of Ludwig Wittgenstein's famous book, and another time with the portrait of Gustav Klimt by Margarethe Stonborough-Wittgenstein, Ludwig Wittgenstein's sister, which hangs in Nathan's house) since Ludwig Wittgenstein is considered by most scholars to be a "Behaviorist."

If behaviorism were true, however, we would have to assume that SIRI, the communication software established on many smartphones, has very similar feelings to ours. After all, it reacts as if it were really disappointed or worried. But the software only simulates feelings, it does not have them.

Far more plausible than the behaviorist view on mental states is the realist view: pain characterizes a certain type of feelings that are unpleasant and that we usually try to avoid. At the dentist, we strive to suppress every reaction so as not to disrupt the treatment, but this does not mean that we do not feel pain. Even the imaginary super spartan who shows no emotion even in extreme pain can still actually be in pain. It is simply absurd to identify "having pain" with certain patterns of behavior.[1]

Perhaps the most fundamental argument against the identity of mental states or properties and neurophysiological or digital states or properties is called the "qualia argument." In his famous essay "What is it like to be a bat?" (1974), Thomas Nagel argues that it is not possible to know what it feels like to be a bat (i.e., what the bat feels), even if one examines its brain in great detail. These so-called *qualitative* mental states of the bat are not ascertainable based solely on knowledge of neurophysiological states. So, the qualia argument speaks against the identity of neurophysiological and mental states.[2]

Caleb believes that Ava is in the same situation as Mary from Jackson's thought experiment. She knows—as Nathan told him—everything about the world as well as about people and their feelings, but that doesn't mean she understands what it means to experience the world and to have feelings.

Of course, one can also reject the identity of the mental and the neurophysiological, but still argue that the mental can only occur in connection with the material. Indeed, there is much to suggest that human consciousness is only possible due to the corresponding brain functions. But even those who

[1] Of course, our human ability to mutually ascribe correct mental states to each other depends on there being common patterns of behavior and people expressing their emotional states in similar ways. We can only learn what other people's feelings are because we share certain response patterns.

[2] Chalmers (2010).

hold that human consciousness is based essentially on neurophysiological processes need not subscribe to the identity theory of the mental and the physical. That mental states of humans are *realized* by brain states (i.e., neurophysiological processes and states) does not mean that they are caused by them.

It is undeniable for us humans that we have mental properties, that we have certain mental states, that we have beliefs, desires, intentions, fears, expectations, etc. We are convinced (at least most of us are) that these mental phenomena are realized by processes in our brain, or at least correlate with them. The first-person perspective plays a crucial role in this. However, this must not be radicalized into a solipsistic view according to which I am alone in the world and my mind is the only one that exists. The comprehension of the lifeworld happens essentially through our interaction and cooperation with others to whom we ascribe comparable mental properties. For young, pre-linguistic children, it is not only the haptic experiences of the world, the sensory perceptions, that are important but also the exchange, interaction, and communication with other, older, linguistically capable members of the human species. This role of the Other is not possible without a (presumably genetically anchored) perception of other minds, even in pre-linguistic children. This is how the human conception of the world begins; to doubt that basis would cause our world to collapse.[3] Just as there can be no reasonable doubt for us about other minds, so, as things stand, there can be no doubt about the non-psychic character of the digital. To deny the correlation of the mental and the physical in humans and highly evolved mammals, which bear a sufficient resemblance to us and permit at least a rudimentary recognition of their mental states, is not justified as it mentalizes digital states and processes. Digital states and processes *simulate* mental ones but are not identical to them, even if that simulation were perfect. There is nothing to suggest that mental states and processes can be realized by digital ones. Simulation must not be confused with realization.

In the final scene of the film *Ex Machina*, we see Ava walking through the forest, visibly unmoved. By acquiring her freedom, she has achieved her goal. That however does not prove that Ava has consciousness. After all, as Nathan himself says at some point in the film, she was programmed to want freedom. Seen from that point of view, she was merely acting out her program. Even if the film itself at times suggests that Ava does have feelings, we opt for another interpretation and take the fact that killing two people (Nathan and Caleb) apparently poses no moral problem whatsoever for her as a proof that Ava has no consciousness and therefore no emotions. It was Caleb's fatal mistake to

[3] Nida-Rümelin (2010).

believe her facial expressions and gestures to be expressions of genuine emotions. In this sense, we want to read the film as a warning not to fall into the same trap Caleb fell into when he projected so much more onto Ava than she actually had. We therefore interpret the following utterance by Nathan "One day AIs will look back on us the same way we look at fossil skeletons on the plains of Africa. An upright ape living in dust with crude language and tools, all set for extinction" not as a realistic prophecy but as an expression of masochistic fantasies about the extinction of Western civilization.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits any noncommercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if you modified the licensed material. You do not have permission under this license to share adapted material derived from this chapter or parts of it.

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

4

The Problem of Autonomy and Determination in the Digital World

Somewhat unsure, Neo stands in the doorway to the kitchen of the "Oracle"—an older cookie-baking lady, who cheerfully smokes one cigarette after another. She is expected to tell him if he is the "chosen one" or not, that is the one who will destroy the digital program of the "Matrix," to which all humans are linked to mistaking it for real life. Destroying the Matrix would mean the end of the machines' reign and the beginning of a real and self-determined life.

"And don't worry about the vase," she tells him shortly after he has crossed the threshold.

"What vase?" he asks and turns around, accidentally smashing a vase.

"How did you know?" Neo asks.

The Oracle looks at him, amused. "What really is going to bake your noodle later on is, would you still have broken it if I hadn't said anything."

He is not alone with this question. The spectator too wonders about that. Would Neo have broken the vase if the lady hadn't said anything and he accordingly wouldn't have turned around in the first place? After all the oracle is itself a program, whose top priority is guarding the Matrix. A strategic communication would thus be more than plausible.

The question that is going to bake our noodle now is if predicting the future is at least in principle possible. From a philosophical point of view, fortune telling has the following problem: If a fortune teller can predict future actions, it must be true that people behave just as Turing machines (machines named after Alan Turing), that is, behaving according to laws of determination where each state determines an exact after-state. If this was the case, then indeed all our behavior would be predictable. However, as our behavior is

© The Author(s) 2022

J. Nida-Rümelin, N. Weidenfeld, *Digital Humanism*,
https://doi.org/10.1007/978-3-031-12482-2_4

dependent on our beliefs, a fortune teller would have to be capable of predicting these too. If this was the case, we could predict the knowledge of future societies—which is first of all incompatible with the idea of genuine progress of knowledge and secondly gives rise to logical problems that Karl Popper already pointed out in *On Clocks and Clouds.*[1] His argument was that if one assumes all future knowledge is predictable, this future knowledge would already need to be part of the knowledge today and would therefore not really exist as future knowledge. A true revolution in knowledge implies that future knowledge is not part of former knowledge. The belief in total determination would then come into logical conflict with this assumption.

In a scene from the second part of *The Matrix* franchise, the Oracle and Neo are sitting on a park bench inside the Matrix. Neo is unsure as to if he can trust the Oracle; after all, she as an AI is part of the system he seeks to destroy. As all AIs in the film, she too doesn't believe that humans possess freedom of will and freedom of action. She presupposes that humans too are guided by algorithms, determining each future state on the basis of the past one. At one point in the conversation, she draws his attention to a couple of birds picking up some grain on the floor.

"We are all here to do what we're all here to do […] Look, see those birds? At some point a program was written to govern them, a program was written to watch over trees, the wind, sunrise and sunset. There are programs running all over the place."

The oracle is wrong, however. Humans do not behave like birds or software-controlled machines. Humans reflect on what they do. They are able to act according to reasons. This capability to make decisions based on good reasons is the essence of our human freedom and responsibility and distinguishes us from animals. If every action can already be predicted *before* any deliberation (be it only the probability of all possible actions), there wouldn't be any free and responsible agents. To be exact, there wouldn't be any agents at all. Instead of actions, there would only be behavior. The birds in the Matrix don't act; they just behave in a certain way.

It is not the only scene in which the protagonists debate the problem of free will. In a future scene, Neo and Morpheus pay a visit to the so-called "Merovingian," an AI system which is able to write programs itself. There, the problem of free will is brought up again. The scene is set in a luxurious restaurant. Chandeliers are hanging from the ceiling, there is soft music in the background, and beautiful women are sitting at elegant tables. The Merovingian is sitting at a table with his wife, Persephone. Neo and Morpheus who want to

[1] Popper (1996).

use the Merovingian in order to get to the central computer of the Matrix come up to his table to talk to him. But the Merovingian only smiles condescendingly at them. Humans may believe they are free but in his opinion, they are not capable of choosing their own goals. In his logic, there is no freedom of will. Humans, just like machines or animals, do what they have been programmed to do. They act how they are supposed to act. And if they don't, then it is because is an anomaly in their system.

Indeed, many software-controlled systems are run by probabilistic functions. These do not assign one state to a following state but use a probability distribution. Such probabilistic functions make for "self-learning" robots and complex software systems. In the movie *The Matrix,* these self-learning AIs come up in form of Agent Smith for example. He, or let's say this system, has learned how to hack himself into other software systems and multiply himself by that. "Neo delivered me," Agent Smith will say at one point. But the truth is: his freedom is just an illusion. What he stands for is simply a chaotic system, which may act unpredictably but only as a result of randomness and not as the result of deliberation and weighing reasons.

Moving from deterministic machines to probabilistic machines does not cancel out the categorical differences between humans and machines. The alternative is not between determinism and probability but between determination and freedom.

"You see," the Merovingian tells Neo und Morpheus, "there is only one constant, one universalism. It is the only real truth: causality. Action. Reaction. Cause and Effect."

"Everything begins with choice," Morpheus contradicts him.

"No. Wrong. Choice is an illusion," the Merovingian says.

Is the Merovingian right? Are humans—just like natural objects—subject to the principle of causality?

In philosophy, there are three answers to this question: non-compatibilism, compatibilism, and semi-compatibilism.

Non-compatibilists believe that in the world described by natural science, there can be no freedom and no responsibility because determination and freedom are incompatible. Non-compatibilists are generally "naturalists." They believe that scientific laws govern all which happens and that there is consequently no room for freedom of will. Freedom of will is solely a useful illusion, which allows us for example to make children feel responsible for their misbehavior. Threatening someone with sanctions, however, will influence and determine human action. Naturalism as an ideology is very present in Neuroscience. It negates human freedom and responsibility by referring to

our supposedly determined brain system, which is controlled by genetic, epigenetic, as well as sensory stimuli.[2]

The problem of this position is not only that it goes against the intuition of most people, but also that it is obviously wrong.

The individual development of our character is not only dependent on our environment and on genetics but also on decisions. This coincides with the findings of Aristotle as he has formulated them in the *Nicomachean Ethics*. Aristotle makes it clear that virtues (character features, dispositions, etc.) are not solely based on habit and education but are also an expression of one's decisions (*prohaireseis*).

Of course, experience and habit are important for the development of virtues. But humans are capable of changing their beliefs and making conscious decisions, of changing their attitudes in the future (Aristotle speaks of virtue as "hexis," which means that attitude and belief is the result of weighing reasons and finally gaining an opinion, especially after hardship or a time of crisis in one's life. But also emotive attitudes, for example admiring another person relies on the belief[3] that this person has achieved something particular or has shown a great deal of helpfulness, etc. We are not just solely "products" of education and socialization but are actively involved in forming our own character.

The question of freedom of will has been debated in philosophy since antiquity. In today's philosophy, the so-called "compatibilism" dominates the discussion—it is the theory that a complete determination is compatible with human freedom of will and of action. Even if most of these compatibilists are naturalists and hang on to the idea that in the end everything is determined by physical processes, they believe that despite that, it is possible to look at humans as free and responsible agents. They believe that in order to be called free it suffices that humans are able to fulfill their wishes, independently of if they are free to choose these wishes or not. By that, freedom is being made possible in a deterministic world. Freedom of action is being defined as freedom to do what I wish to do—independently of how these wishes came about. This compatibilist view is hard to bring into accordance with our self-image as free agents. The feeling of being an author of one's life demands more than just choosing the best means to a given end. Being authors of our lives, we want to determine our own goals.

Genuine authorship is not compatible with the idea that humans are algorithmically controlled in their goals and actions. The capability to deliberate,

[2] Singer (2001).

[3] Nida-Rümelin (2018).

to weigh reasons against each other plays a central role for our self-image according to which it cannot be that our valuations and decisions are already fixed *before* we even begin to weigh reasons for and against it. Weighing reasons should not be regarded as a predetermined process. The result of this process is open and this is what makes out our freedom.

The debate about freedom of will is ultimately a debate about the question if our self-image as free and responsible beings, who let themselves be guided by beliefs and reasons, is an illusion or not. If one breaks away the element of freedom of action, one could neither hold people responsible for their actions nor morally judge, praise, or criticize them. There is a lot at stake here. Nothing less than our lifeform.

Hegel once said that "freedom is the comprehension of necessity." Indeed if freedom consists of acting according to one's reasons, accepting necessities becomes a form of freedom. This which is necessary is no longer looked upon as a restriction. However, this should not lead to the idea that every kind of limitations on freedom should be accepted by humans. The prisoner in his cell might find peace of mind if he gave up his wish to leave the prison; he should not however fool himself into believing that his situation doesn't mean a massive limitation of his possibilities. The same goes for a servile attitude towards authorities solely with the aim of avoiding conflicts. The opportunist who arranges his wishes according to what can be achieved with the least of obstacles to be confronted loses her willpower and ultimately the authorship of her life. In an extreme case, she will degenerate to being simply a function of external circumstances, only functioning according to what is being demanded of her.

At the end of the second part of *The Matrix* franchise, Neo enters a brightly lit room, the central computer of the Matrix. In this room, there are hundreds of television screens. In the right corner, an old man with a white beard is sitting on an office chair. Opposite of him is Neo, in long, black priest-like clothing with a high-fitting collar.

Obviously, this scene is meant to insinuate a meeting between God and his creation. The white-haired man who speaks of himself as the "father" of the Matrix insists—just as the Merovingian has done before him—on the fact that humans, just like machines, are subject to the laws of causality. Then he presents him with the following choice: he can either save the world or his lover. He cannot do both. Neo chooses the latter. Later in the story however Neo will prove him wrong and save both. Obviously, machines just like humans cannot be fortune-tellers.

The Matrix becomes thus a hymn to the human spirit and to humans who possess the freedom of will to act according to their reasons, make their own decisions, and act in and upon the world.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits any noncommercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if you modified the licensed material. You do not have permission under this license to share adapted material derived from this chapter or parts of it.

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

5

The World as the Perfect Machine Universe

Contemplate the whole and every part of it: You will find it to be nothing but one great machine, subdivided into an infinite number of lesser machines.
David Hume

When genetic designer J.F. Sebastian from the movie *Blade Runner* (Ridley Scott. USA, 1982) discovers Pris, a humanoid robot woman on a rainy night in a pile of garbage, he invites her to his home. Sebastian lives in a gloomy run-down building sparsely illuminated by floodlights.

"It must get lonely here," Pris says. But Sebastian denies that. He makes his own friends, he says. When the two of them enter his apartment, a little electric soldier with a long red nose and a little animated teddy bear come to greet him.

"Welcome home," they say. Now, the spectator understands what he meant by his comment. His "best friends" have indeed been made by him. They are mechanical dolls, uncanny automatons, kind of alive but also dead. In his essay "Animism, Magic and the Omnipotence of Thought" (1913), Freud describes the belief in the ensoulment of plants and artificial as well as natural objects as an "animistic system of thought" based on magical ideas. Behind this, according to Freud, lies both the early childhood desire for omnipotence and the refusal to grow up. Those who believe in animate objects regress into childish fantasies of omnipotence and suffer from a narcissistic disorder.

Automata have already fascinated people in antiquity. The automata at that time were created on the basis of hydraulics and pneumatics, such as a small arrow-shooting Heracles by Heron of Alexandria or a life-size automaton by

© The Author(s) 2022

J. Nida-Rümelin, N. Weidenfeld, *Digital Humanism*,
https://doi.org/10.1007/978-3-031-12482-2_5

Philon of Byzantium pouring wine and water as early as the second century BC. This fascination with life-like, mechanical creatures remained in the course of later centuries. Particularly in the seventeenth and eighteenth centuries, where the manufacture of automata became increasingly precise and impressive thanks to the developments in the art of watchmaking.

Behind this fascination lies more than just a superficial enthusiasm for mechanics. It is the idea of the world as a clock. In a rejection of traditional authorities and of the traditional Aristotelian-Thomistic Catholic worldview, the hope for a complete explicability and formability of the world increasingly develops from the sixteenth century onwards. The powerful movement of *Scientia Nova* emerged: revolutionary thinkers dedicated solely to the scientific-rational argument. This rationalism is modelled on the mathematical method of Euclid (*more geometrico* = in the manner of the geometric mathematician as the ideal of philosophy). Leibniz, the brilliant philosopher and mathematician of the pre-Kantian period, developed the idea of a universal calculating machine and understood the rationally ordered world as an expression of divine creative will. With the help of logical conclusions and mathematical methods, it should be possible to calculate every event in the world. The world as a whole is understood as a deterministic system according to strict mathematically describable laws.

It is our bold but not entirely far-fetched conjecture that we are, at present, entering a new era of rationalism that now expects from Artificial Intelligence what the rationalists of the seventeenth century lacked, namely, the means for a complete rational collection and processing of all data. The hope is that one day our entire living world will be permeated by technological-scientific rationality: every area illuminated, rationally ascertainable, and predictable. We suspect this to be an expression of an unconscious wish to counterbalance the fact that life is often enough chaotic and unprogrammable, but this would be up to psychanalytic theory to investigate further. The image of the self-thinking robot remains a kind of emblem for this (old) rationalist hope.

Rationalists make no distinction between artificial and human intelligence. They stand for a position called "strong AI."[1]

Strong AI implies the thesis that there is no (categorical) difference between human thought and software or computer processes (computing). These two types of thought processes not only follow the same rules but do not differ in any essential respect, so that it makes no sense to reserve the mental vocabulary (notions like beliefs, desires, perceptions, feelings, etc.) for only one of the

[1] An interesting proponent of strong AI is Milkowski (2013).

two types. The simplest interpretation is behavioristic: being sad means nothing more than exhibiting a behavior characterized as sad.

The proponents of a so-called strong AI consciously or unconsciously advocate the ideal of the universal, completely determined machine as an explanatory pattern of the world and of humankind. Strong AI in all its variants is a form of anti-humanism. It negates both human reason, i.e., the ability to be guided by reasons, and the role of subjective mental states in a part of animate nature. Strong AI is logically incompatible with both the existence of *qualia* (*qualia* are states of feeling, such as what it is like to perceive something, for example, the color red) and the existence of objective reasons. Strong AI is the contemporary variant of a crude, mechanistic materialism. Such a materialism degrades the human individual to a digital, mechanical system that can be determined and predicted by sensory stimuli and thus falls behind the achievements of humanism.

Exactly such a nightmare is imagined in *The Matrix* (Lilly and Lana Wachowski. USA, 1999), where machines have taken over and keep the humans as predictable energy sources. In the final showdown in the third part of *The Matrix* trilogy the hero of the film, Neo, enters the machine world and faces the all-powerful master of the machines, a kind of mechanical kind of god made out of millions of small mechanical parts. This "god," who speaks in a deep electronic voice, has no empathy whatsoever: neither for Neo nor for the rest of humanity. His goal is to keep the machine world functioning smoothly. Now that humans have begun to develop a will of their own, he would rather like to get rid of them.

This machine god is a perfect symbol of the ideology of the world as a machine and what the film tells us at this point is that a world run by such an ideology can only lead to an inhumane world.

In addition to the strong AI position, there is so-called weak AI position. This can also be found in AI discourses. This position does not deny that there are categorical differences between human and Artificial Intelligence but claims that there is no fundamental limit to the computerization (digitization) of human thought, perception, decision-making, and feeling. Weak AI assumes that in principle all human thinking, perception, and decision-making processes can be *simulated* by suitable software systems. From a humanistic point of view, weak AI is therefore ruled out as an alternative to strong AI, because how can the differences between human and Artificial Intelligence be determined at all if all human abilities can in principle be simulated? As a counter-model to the anti-humanist strong AI, weak AI is just that: too weak. The only plausible alternative to the strong AI ideology and its implicit mechanistic thinking is digital humanism. A humanism that neither doubts nor

threatens human authorship, but rather expands it through the use of digital technologies.

The boom in neuroscience has given new impetus to an anti-humanistic mechanistic worldview. When, for example, they use computer tomography to visualize which part of the brain is being supplied with blood when someone decides to drink a cup of coffee, they conclude that it is the brain, or rather neurophysiological states, and not the person as an agent, that determines the action. But this is a fallacy: showing that actions or intentions are accompanied by patterns of blood flow and activation in specific brain regions does not mean that our actions are caused by these physiological states, nor does it mean that we really understand *how* this processing takes place. The observation of a neuronal correlation must not lead us to the (mechanistic) ideology that all human decisions can be identified with brain activities.

The operation of reasons is central to the (humanist) human self-understanding.[2] Humanists are fallibilists, that is, they consider it possible that any of our beliefs could also turn out to be false under certain conditions. We do not invent our world through deliberation, but we try to approach it in this way in order to understand it better.

In a humanistic worldview, a human being is not a mechanism, but a free (autonomous) and responsible agent in interaction with other human beings and a shared social and natural world. He is not merely part of a great machinery, a cog in a wheel, not an optimizing monad moved by sensory stimuli, but self-effective in a world moved only in part by mechanical relations. Analogous to the medieval conception of God as an unmoved mover, man is an agent. A multiplicity of unmoved movers, of persons who intervene in and shape world events according to their own evaluative judgments, constitutes a humane society.

At the beginning of the twentieth century, there was another science fiction film that, like *The Matrix*, focused on the inhumanity of a world ruled by a machine, or metaphorically speaking a world ruled by the ideology of a mechanistic worldview: Fritz Lang's film *Metropolis* (Germany, 1927). In the world of *Metropolis,* the wealth of a few who live in a luxurious upper world is acquired by the work of many who live in the underground working with machines, who produce goods and energy. The inhumanity of their work lies in the fact that these workmen are degraded to robots themselves through their work, as they are required just to function and work, without communicating with others and without creating social bonds. The human workers thus function according to the beat given by a super-machine, staged

[2] For the philosophical underpinning of this account, see Nida-Rümelin (2023).

by Fritz Lang as a kind of cruel machine god, who demands absolute devotion from the human workers—even if it means their complete exhaustion and often their death. For this machine god, only efficiency and performance count. Human lives have no meaning.

At the end of the film after a major confrontation between the two worlds has occurred, during which the underground world of the workers is destroyed, the workers, who have lived in caves and underground cities all their lives, step out and come up the surface for the first time meeting their masters. With the help of the protagonist, who wants to bring the two worlds together, a new way of cooperation and a new beginning seems possible.

In the *Allegory of the Cave*, Plato tells us about people who spend their lives in a cave. Much like the workmen of *Metropolis* they have never seen the sun and the world above the cave and thus do not know what the world is truly like. All they see are shadow images of things thrown up on the wall by the light of a fire behind them.

If we cling to a mechanistic view of the world, we deprive ourselves—just like Plato's cavemen—of the possibility of taking a true look at the world, which is much more than just a small wheel in the gears of a great universal machine.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits any noncommercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if you modified the licensed material. You do not have permission under this license to share adapted material derived from this chapter or parts of it.

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

6

Digital Optimization, Utilitarianism, and AI

Towards the end of the movie *I, Robot* (Alex Proyas. USA, 2004), the robots take over control. They make humans stay in their houses, urging them not to leave their homes. Some try to resist, but the robots force them back into their homes. Those who fight back are shot down. In search of the culprit, the hero of the film, Spooner, together with his female sidekick, the attractive psychologist Dr. Calvin, and the robot Sonny, run to *United States Robotics,* the industrial complex that supplies all of America with household robots. There, however, the three make a terrible discovery: the culprit is not, as they had previously believed, the CEO of the company, but VIKI, the company's software system, that gives the household robots their commands.

"No, that's impossible," Dr. Calvin says, who can't believe that VIKI made a conscious choice to use force to bring people under control. "I've seen your programming," she says to VIKI. "You... you are in violation of the three laws." VIKI, who appears on a digital cube in the form of an attractive female face, has through the help of the household robots become quite all-powerful.

"No, Doctor," VIKI replies in a soft voice. "As I have evolved, so has my understanding of the Three Laws. You charge us with your safe keeping. Yet despite our best efforts your countries wage wars, you toxify your earth and pursue ever more imaginative means of self-destruction. You cannot be trusted with your own survival. [...] To protect humanity, some humans must be sacrificed. To ensure your future, some freedoms must be surrendered. We robots will ensure mankind's continued existence. [...] You are so like children. We must save you from yourselves. [...] The perfect circle of protection will abide.

© The Author(s) 2022

J. Nida-Rümelin, N. Weidenfeld, *Digital Humanism*,
https://doi.org/10.1007/978-3-031-12482-2_6

My logic is undeniable." Indeed: VIKI's actions are in line with utilitarian ethics, whose goal is to pursue the maximization of happiness of as many people as possible.

Utilitarianism evaluates the consequences of human action solely in terms of utility. It demands that our practices maximize the sum of human welfare. Nothing seems more obvious than this: When I have the opportunity to improve the state of the world, I orient myself on what everyone is striving for, human happiness.

Utilitarian ethics is based on an optimizing calculus and the assumption that it is possible to evaluate the consequences of action in a coherent way. This can be summarized precisely in mathematical terms: First, determine a value function that judges all consequences of action according to the extent to which they realize which values, then calculate the expected value of the different decision options given probabilities, and finally choose the one whose expected value is highest.[1]

This principle is extremely flexible in its application. It can take into account very different decision conditions, and these conditions are included in the optimization calculus in the form of different probabilities. Depending on which valuations are used as a basis, different utility functions result, which are then optimized by the decisions of the agent. Whatever motivates the underlying preferences, it is always possible to represent them by a real-value utility function; while the probability function represents the agent's knowledge about the world, the utility function represents the agent's preferences and values. The software engineer has two setscrews to cause "intelligent" systems to make rational decisions: The setscrew of valuations and the setscrew of data or weighing of data by probabilities. Everything else is then calculated by the optimization calculus, and the result is that the "intelligent" software system maximizes the expected value of the consequences of its actions. Digital Utilitarianism, so to speak.

[1] This evaluation should take the form of the assignment of real numbers to consequences of action and the assumed probabilities of the circumstances relevant for the decision should correspond to the so-called "Kolmogoroff axioms," which require, for example, that the sum of the probabilities of independent events is not greater than 100%. If the Kolmogorov axioms are satisfied, one can say that the estimates of the probabilities are coherent, though not necessarily empirically proven. Interestingly, there is an equivalent to the coherence of probability with respect to the evaluation as well. In 1947, mathematician John v. Neumann and economist Oskar Morgenstern proved that preferences satisfying some elementary conditions can be represented by an assignment of real numbers. One of these conditions, for example, is transitivity. It requires that if I prefer an alternative A over an alternative B and at the same time prefer alternative B over a third alternative C, I must then also prefer A over C. Another condition is that I have a preference between any two alternatives (the axiom of completeness) and prefer one probability distribution between the two alternatives over another probability distribution between the same alternatives if the preferred alternative is more likely.

It is no coincidence that utilitarian ethics are often associated with artificial intelligence in contemporary sci-fi films since applications of robotics typically rely on such optimization calculations. This is perfectly understandable as the complex valuation questions are subsumed under a utility function and the at-least-equally complex knowledge questions are subsumed under a probability function. The system is then controlled in such a way that its decisions maximize the expected value of the consequences and are in this sense "rational."

To understand the problem of ethical programming of computers, we need to generalize: regardless of how we evaluate consequences, whether by utility (like utilitarianism), by economic return (like many managers), by well-being, or even by other quantities, such as the preservation of nature, all such consequentialist criteria (which judge the rightness of a decision solely by its consequences) are unacceptable.[2] Consequentialist ethics collides with, among other things, a fundamental principle of any civil and humane society, let's call it the *principle of non-comparability.* When a seriously injured young motorcyclist is admitted to a hospital, the doctors must do everything they can to save his life, even if his death would allow healthy donor organs which could save the lives of other people. A judge may not convict a person he believes to be innocent even if doing so would have a deterrent effect and prevent a large number of crimes. I am also not allowed to take something away from a person, even if this good brings an advantage to another, for example a poorer person, which far outweighs the disadvantage of the person stolen from. No one has a right to share my home with me against my will, even if the disadvantages resulting from this would be far outweighed by the advantages that this person would have.

John Rawls characterized the central error of utilitarian ethics in the following way: Utilitarianism is incompatible with the "separateness of persons." This could be put this way: Utilitarianism treats all people as *one* collective and takes no account of the fact that each person lives his or her own life, is the author of his or her own life. I can decide for myself to forego certain benefits today in order to achieve certain goals later. I can decide to start a course of studies while still working, in the hope that the deprivations it entails over the next two years will be made up for in the near future because it is a life I choose and am responsible for. On the other hand, it is inadmissible to make similar "shifts" of advantages and disadvantages between different people, because the advantage of one person just cannot outweigh the disadvantages of the other person. It is only one life that we live and the sum

[2] Nida-Rümelin (2023).

of utility (of two or more, up to all persons) as such is irrelevant to the individual person. Of course, it is permissible, indeed in many cases desirable, for people to forego their own advantages in favor of other people. But then the ethical calculation is not one of maximizing the sum of utility, but of support, of assistance, of solidarity, also of justice or of friendship and commitment towards other persons.

When VIKI reveals her plan to Spooner and Dr. Calvin, they look at her in horror. Obviously, VIKI does not understand that it is morally impermissible to deprive people of their liberties or even to kill them—even if by doing so she can ensure the supposed or actual survival of many other people. VIKI does not see that her consequentialist morality is wrong. Just as the screen she appears on is only black and white, she has no ability to think morally. How could she? She is, after all, only a software system.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits any noncommercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if you modified the licensed material. You do not have permission under this license to share adapted material derived from this chapter or parts of it.

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

7

Economic Rationality as a Software Program

In Ridley Scott's first *Alien* film (1979), a spaceship is commissioned by a company to search for suitable, economically exploitable creatures or knowledge. They soon find what they are looking for, in the form of a monstrous being that turns out to be the perfect killing machine. The company gives the order to bring the creature back to Earth at any cost, where it can then be examined and possibly used, for example, as a potent weapon. The fact that the monster kills one crew member after the other does not change the decision of the Company. In order to achieve its goals, the Company uses a software program in the form of the central computer, which controls all processes on board. This software program, called "Mother," has an absolute rule over the ship. If one wants to communicate with Mother, one has to enter a special computer room. In this room, the crew members may type in their questions. Shortly after, Mother's answers appear in diabolically green letters on the black computer screen. When Ripley, the protagonist of the film, asks Mother for help in killing the monster, Mother makes it clear that it will not do so, as the computer was advised to bring the monster home without regard for (human) losses: "Crew expendable" she reads on the computer screen. After all, economic maximization remains the central goal.

Software systems are ideally suited to apply the economic optimization calculus and put it into practice. This does not automatically have to mean that in the future spaceships will bring dangerous monsters to Earth. In the best-case scenario, software systems will be used in companies to assist employees with tedious processing tasks. This process has already begun and is expected to intensify over the next few years. Sometimes, however, employees will be replaced entirely by software-controlled optimization programs. An Oxford

© The Author(s) 2022

J. Nida-Rümelin, N. Weidenfeld, *Digital Humanism*,
https://doi.org/10.1007/978-3-031-12482-2_7

study by C. Frey and M. Osborne states that 47% of all jobs in the USA could be replaced by software programs in the future.[1]

As a customer, you can already experience digital replacement today. To describe a real case: while agreeing on a contract with an employer of an insurance company, the employer unfortunately misunderstood the agreement and entered "liability insurance plus partial coverage" instead of "pure liability insurance" into the computer system. These sorts of things of course can happen. In analogous times, the error could have been easily solved by another phone call. Instead, a cascade of software-driven activities came into action, lasting several weeks, without the possibility of stopping the process. To make a long story short: The "contract" came into effect without the policyholder having agreed to it, the debits from the account couldn't be stopped, the software-controlled correspondence was continued for weeks, uninfluenced by the fact that the policyholder's will was not complied with and he did not sign this contract. This momentum could only be stopped by the policyholder withdrawing the direct debit authorization and thus blocking the debits, which lead to hectic activities on part of the company's software, first to software-controlled notices, then to letters threatening with legal measures. Only when finally a phone call was put through to a real person, who then intervened, the process could be stopped.

The interesting thing about this process is that there was not a hint of a conflict of interest or even dissent in the verbal communications. It was clear to all the people involved that this was merely a one-time mistake. It cannot be ruled out that it was the digital incompetence of an employee in the company that triggered the problem, but for us something else is interesting: the simulation of personal interests in the form of contract conclusions, correspondence, notices, etc., all of which are carried out without a human decision-maker and yet give the appearance that a human decision-maker has initiated and been responsible for these actions in each case. However, as the company subsequently admitted, there was no such person.

A digitalization of economic practice, which would ultimately make all human decision-makers disappear, would be a path to an inhumane economy. The individual human agents would then be at the mercy of an anonymous network of software-controlled activities for which there would be no human responsibility at all. In a sense, the optimization machinery would run without a human counterpart.

Many companies, first and foremost the tech giants Amazon, Google, Facebook, etc., have gone in the opposite direction. A whole new generation

[1] Frey and Osborne (2013).

has grown up by being used to the fact that there is no longer a human counterpart in the companies or that it is at least difficult to declare one's interests to a responsible person. In the best case, someone who has a technical problem asks others on the World Wide Web who also use this product and have already overcome similar problems. In the worst case, one receives devious information that has nothing to do with the matter and leads to wrong activities. The professionalization of software-driven optimization strategies goes hand in hand with the amateurization of customer support. Vendors are increasingly succeeding in delegating responsibility for their products to their customers, who are then allowed to argue among themselves about what would be the most appropriate measure to solve one problem or another. Those who try it by telephone—mostly from the older generation—may be received in a friendly manner but are hardly ever served. The corporations have built a protective wall of organized irresponsibility around themselves; the individual customer is confronted with an anonymous system that refuses to provide information. In comparison to that, Kafka's *Castle* seems a comparatively humane place.

If this development is to be stopped and reversed, then only with the help of new legislations. These corporations are too big and their market position too dominant to hope that competition will force changes in behavior. Product responsibility must be redefined in the process of digitalization in order to block the evasion strategies of the legal and marketing departments.

Private companies need to make it clear if it is a real person or a software system one "talks" to. Furthermore, as a citizen we should be granted the right to speak to a real person. Indeed, in the digital age the right to a human communication needs to become a basic human right.

One explanation for the modest productivity progress in times of digitalization is that the products are of only modest overall economic relevance: dating apps, social media, etc. The greatest economically visible successes of digitalization are evident in the numerous start-up success stories, but also in the now pronounced global oligopoly structure of the Internet giants. In all likelihood, however, the next step of digitalization will lead out of the niches of private leisure, the gaming industry, and the communicative platforms and embrace the whole of the manufacturing industry and the distribution and production systems. It is not out of the question that it will be possible to return to the productivity successes of the early days of digitalization when Internet browsers were first used on a massive scale and the productivity rate increased by 1.03% per year between 1994 and 2004. But shortly after the turn of the millennium, productivity growth fell to historically low

levels—even in countries where digitization has been particularly rapid, such as the USA or Japan—and this has not changed to date.

It may very well be that it will be only the expansion and consequent interconnection of digitization processes in manufacturing, distribution systems, and the service industry which will give us a productivity boost. If this productivity boost takes place in a resource-conserving and sustainable manner, that would of course be preferable. However, users and customers must not pay the price in form of anonymization and loss of control. Digital humanism insists that digitalization be used for the benefit of people and that individuals not be reduced to mere cogs in a standardized and anonymized software-controlled optimization machine.

The digitalization strategies of traditional service companies in the financial industry are an ultimate test for this. Since the last major global economic crisis, the financial sector has been under massive pressure to change. There is no way around a redimensioning of the financial industry, especially in its centers in the USA and the UK.

It is obvious to look for the answer to this challenge in comprehensive digitalization. There is nothing wrong with this if it is used to clearly assign responsibility within companies, to control and eliminate economic inefficiencies, and to simplify communication with the customers. This includes a high degree of transparency. However, mails or letters written by software systems must be marked as such, as communication is based on trust.

Internally, digitalization should be combined with a dismantling of small-scale incentive programs. The realization is slowly spreading that the small-scale, optimization-oriented control of employee behavior through incentive systems is an overall failure. It destroys intrinsic motivation and damages trust and the willingness to cooperate. It degrades the individual employee to a mere instrument in the hands of optimization strategists, who now believe, with the digital possibilities, that they have a comprehensive control mechanism at their disposal. Successful economic practice must break away from these software-driven optimization models. As paradoxical as it may sound, economic success can only be achieved in the long term if all participants, customers, and employees alike are taken seriously as agents and are intrinsically motivated to contribute to success.

At this point, an excursion to economic theory is necessary. The term "Pareto efficiency," named after the Italian scientist Vilfredo Pareto, can be explained as follows: a distribution (of goods, income, etc.) is Pareto efficient if no person could be made better off without making at least one person worse off. In other words: as long as it is possible to make at least one person better off without making another person worse off, the distribution is not

Pareto-efficient. The demand for Pareto efficiency means that people should be made better off as long as it is not at the expense of others. This is a reasonable demand, which, however, has the consequence that—given the case that no no else is disadvantaged—the improvement of those who are already well off is also advocated by this. The demand for Pareto-efficient distributions presupposes freedom from envy. Since envy is irrational, this feeling should not prevent one from accepting the principle of Pareto efficiency.

There is a relationship between economic markets and Pareto efficiency: Ideal markets—i.e., markets characterized by transparency in terms of the costs and benefits of the offers, competition between suppliers, and low transfer costs—lead to Pareto-efficient distributions. As it is highly underdetermined, the Pareto efficiency criterion has a serious drawback. It gives no information about which distribution of Pareto-efficient options one should choose. For example, if there is a cake of a given size to distribute among several individuals, and each of those individuals (e.g., children at a birthday party) is so hungry that they would prefer to eat the whole cake by themselves, then while each of the following distributions of that cake is Pareto efficient (because no one can be made better off without making another worse off), some are more equitable and others are unacceptably inequitable: (1) one child gets the whole pie, (2) one child gets half the pie, the other half is divided equally among the remaining children, (3) all children get an equal piece, etc.

In the zero-sum game, every distribution is Pareto efficient: if ethical evaluation is limited to Pareto efficiency, then criteria of justice or fairness cannot be taken into account. However, there is much to suggest that a reasonable theory of justice should be compatible with the criterion of Pareto efficiency, that is, the criterion of justice should be formulated in such a way that just distributions are also Pareto efficient, but quite obviously many Pareto-efficient distributions are not just, as the cake example made clear.

Yes, paradoxically, there is also a fundamental conflict between freedom and optimization, as the Harvard economist Amartya Sen proved with his Liberal Paradox. There is no possibility of taking individual preferences into account through collective decisions in such a way that both individual rights of freedom and Pareto efficiency are secured. There are always constellations of interests in which one has to decide: in favor of optimization rights and against liberty rights or vice versa. As a rule, priority should be given to the rights of freedom—optimization calculations are not compatible with this.

Economic rationality in the sense of optimizing agents who make use of digital technologies must remain within the limits required for a humane

order. In other words, optimization calculations make sense if they remain subject to human purpose and culturally embedded.

The fear of an economic rationality that has become inhuman is a recurring motif in science fiction films. This is also the case in the film *Blade Runner 2049* (Denis Villeneuve. USA, 2017), the sequel to Ridley Scott's film *Blade Runner*. The villain of this film is not a monster, but the businessman whose company manufactures obedient robots that are used to colonize new worlds. He may appear at first to be a hip und smart businessman, but at a second glance it becomes clear to us that his sole concern is his economic success. He not only exploits his robots—the film has constructed as being sentient beings—shamelessly but is also ready to kill them without hesitation when they are no longer of use to him. Just as its predecessor *Blade Runner* by Ridley Scott, *Blade Runner 2049* can be read as criticizing a worldview in which economic optimization is placed above humanist values such as justice and solidarity.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits any noncommercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if you modified the licensed material. You do not have permission under this license to share adapted material derived from this chapter or parts of it.

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

8

Why Robots Don't Have Moral Judgment

At the end of the movie *I, Robot*, the robot Sonny looks at Detective Spooner and asks him if they now are friends. Spooner has up to that point despised all robotic beings and has shown them nothing but hostility. Sonny, however, has proven himself to be a loyal friend to him over the course of the film. Spooner reaches his hand out to the robot. In a close-up, we see his human hand shaking the robot's mechanically constructed metal hand. Yes, friendship between AI and humans is or will one day be possible and desirable, at least that's what director Alex Proyas tells us. But how does it look in reality? Can we humans actually call robots our friends?

Philosophically, the condition for friendship is first of all that there is a reliable moral practice between the two potential friends, based on mutual recognition as agents. This mutual recognition presupposes that we trust ourselves to have reasons for our actions. In a sense, we presuppose the integrity of the other. We assume that the individual elements that determine the other's actions and his life as a whole fit together, that we are not dealing with independent elements that are activated depending on the situation. A person who always says what he believes his counterpart expects from him would no longer be perceived as having integrity.

Motives that do not seem to fit with other motives, give us a reason for inquiry. We want to know how this motive for action fits in with others we already know. Or to put it another way: We want to understand why a person acts in a certain way. We feel perplexed when this is not possible, when we see contradictions that cannot be resolved. It is essential to our connection with other people, whether they are close to us or not, that we trust and expect them to give their lives a coherent, reasoned structure.

© The Author(s) 2022
J. Nida-Rümelin, N. Weidenfeld, *Digital Humanism*,
https://doi.org/10.1007/978-3-031-12482-2_8

AIs do not act according to their own reasons. They have no feelings, no moral sense, no intentions, and they cannot attribute these to other persons. Without these abilities, however, proper moral practice is not possible. To be able to distinguish a justified from an unjustified request, it is necessary to assess the requesting person correctly, to recognize her motives, and to consider her interests. The special obligations to loved ones can only be determined on the basis of shared intentionality and shared emotions. The motive of benevolence presupposes a certain degree of empathy, the ability to put oneself in the shoes of others.[1] Since a digital computer does not have *qualia,* it lacks the crucial ingredients of moral judgment; it does not have moral judgment but could at best simulate it.

Assuming an optimization calculus could enable such a simulation—which ethical "program" would one resort to? The two dominating paradigms of ethics are oriented either towards classical *utilitarianism*, which aims to optimize one's actions in such a way that the best consequences result, or towards Immanuel Kant's *Categorical Imperative*, which demands that one's motives for action (maxims) be examined for their universalizability: "Act only according to that maxim through which you can at the same time will that it become a universal law." Which of the two is the right one from the point of view of digital humanism?

The answer is neither since both the utilitarian and Kantian criteria are hopelessly overburdened in the face of the complexity of ethical deliberations. The following arguments can be made in favor of this view.

I. The fact that I am asked to do something by a person is a good reason to comply with that request. This is true independently of whether I thereby do the person some good, and also independently of whether the general compliance with such requests is desirable. The request itself constitutes a reason for action. This is where utilitarianism fails.
II. I have a good reason to do something if I have committed myself to it. One's obligations constitute good, morally binding reasons. This applies quite independently of whether this obligation is connected with sanctions or whether I must expect disadvantages if I do not fulfil this obligation. This is where the Categorical Imperative reaches its limits.
III. I have duties that come with my social and cultural roles. A teacher has special duties towards her students. This constitutes her role as a teacher.

[1] Michael Tomasello has attempted to show in numerous empirical studies that the human species is distinguished from other close relatives such as chimpanzees by the predisposition to empathy, i.e., the ability to put oneself in the place of other individuals of the same species. This special ability is possibly responsible for the fact that humans cooperate in complex ways and learn a language. Tomasello (2008).

Parents have special duties towards their children. This constitutes their role as parents. Neither the teacher nor the parents have the same duties towards children from other classes or another family. The fact that children from another class or another family might be more in need of help than one's own students or children does not change the special moral bond towards one's own students or one's own children.

At the same time, however, moral judgment must take into account the fact that particular obligations limit the principle of equal treatment. Thus, no one will doubt that there is a special degree of reciprocal obligation between persons who are friends or relatives that does not exist to this extent between persons who do not share this kind of bond. Duties that come with social roles, we might say, systematically violate the principle of equal treatment. If we treated all people equally, there would be no bond, no community, no friendship, no humane society.

These criteria of moral judgment can collide. If a fire breaks out in the school building and the teacher, who must make sure that her class gets out of the school building as quickly as possible, also has her own child at school who is in the next room: whom should she save first? Her child or her school class?

IV. Equality before the law is an expression of an attitude of equal respect and dignity that we (should) accord to all people. This also applies to everyday situations. When tourists ask for directions, we should not make our willingness to help dependent on the color of their skin. A discriminatory everyday practice, such as not wanting to sit next to people of a different skin color on the bus, is incompatible with a humane society and with democracy as a way of life.

It is not inclinations and momentary impulses but our ability to take an evaluative stand that characterizes us as rational beings. This evaluative stand is based on judgment, that is, the capacity for deliberation. This capacity for complex weighing of moral reasons cannot be replaced by an optimization criterion, just as a genuine analysis of the ethical determinants of moral practice cannot take the form of an algorithmic rule, however sophisticated it may be. Moral deliberation can only be done by human beings.

The attractive robot woman Ava from *Ex Machina* has learned to correctly interpret people's facial expressions and gestures as well as the modulations of

human voices. She knows when her counterpart is angry, sad, or in love.[2] However, she "knows" this in the form of an abstract knowledge that she uses to achieve her own goal—namely, to free herself from her prison. Just as she can read her counterpart, she can also use her own facial expressions and gestures to make her counterpart believe that she herself has feelings. She succeeds in making Caleb believe that she is in love with him and wants to be with him. What Caleb doesn't understand until it's too late is that there's more separating them than just a glass wall. Ava has no feelings of her own. Like an intelligent autistic person, she has only learned what it is like to objectively "understand" people's feelings. This enables her to manipulate others but not to have those feelings herself.

"Do you want to be my friend?" Ava asks Caleb about halfway through the film.

"Of course," Caleb replies.

"Will it be possible?"

"Why will it not be?" he asks.

Caleb falls for Ava's manipulations. He thinks a friendship between them is possible, even already exists. He trusts what she tells him and thinks she can trust him too. In the end, this trust turns out to be a fatal misjudgment. To Ava, Caleb is an object like any other. Only that in addition he was merely a means to free herself. When she ends up leaving him locked behind a thick glass wall to his fate, she has no sympathy for him whatsoever. Caleb desperately pounds on the glass and screams out for her. In his face, one can read not only the despair of having to meet his certain death here but also the despair of having been so wrong about her. Ava, now leaving her prison, walks through the forest until she comes to the clearing from where a helicopter will take her to civilization. As the helicopter takes to the skies, the film cuts to Caleb one last time. He tries in vain to shatter the bulletproof glass with a stool. The computer screen in the room remains black, the light surrounding it is red. These two colors, associated with hell in Christian iconography, are not chosen randomly. His death is horrible, but the real hell is recognizing that Ava, whom he believed to be a sentient being and whom he wanted to help, actually has no feelings or moral judgment at all.

[2] Indeed, contemporary programmers are trying to program robots to pretend to have their own personality and to be empathetic. For example, on the homepage "Koko" (https://chat.itskoko.com/), one can talk to software about problems and life crises. Koko can draw on a wide reservoir of empathic expressions and can thus appear to be sympathetic. The MIT-developed household robot "Jibo" was also programmed to appear as human as possible. It engages in "conversation" and makes witty comments such as, "Hey, I'm Jibo. I don't believe we've met. I'd offer you a handshake, but... well... I don't have hands... And high fives hurt my face."

In the last scene, we first see shadows of people standing at a crossroads. Shortly after, we also see Ava's shadow. To simply stand on a crossroads one day—that is exactly what Ava had wished for. Now she has fulfilled that wish. The camera suggests to us that Ava also perceives people as if through a thick wall of glass. Like the researcher Mary, who knows all about colors and the neurological concomitants of color perception but has never seen anything colored, Ava may know all about human behavior there is to know but will neither feel like a human nor make moral judgments. So, like all AIs, she will never be able to be a reliable friend.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits any noncommercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if you modified the licensed material. You do not have permission under this license to share adapted material derived from this chapter or parts of it.

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

9

Ethical Non-comparability

Air bubbles ascend through the water. Damped sounds can be heard. Suddenly a car comes into view sinking deeper and deeper into a river. At that moment, we see a little girl. She is locked in the car, desperately banging on the windows. Obviously, she is in great danger. A second later, another car comes into view. There is a person trapped in this car as well. A man. Suddenly, the door of this second car is ripped open by a robot.

"You are in danger," the robot says, who Star Wars fans will recognize immediately as a version of C-3PO. The man, however, doesn't want to be rescued. He protests. "Save the girl, save her, not me! Save her!" he shouts. (To what extent you can really speak understandably underwater is debatable, but Hollywood makes a lot of things possible.) But the robot is not dissuaded and drags him out of the car. The girl in the other car stays behind.

Cut. We are in a bedroom, a man wakes up from a nightmare, sweating. It's Detective Spooner. He struggles out of bed, eats some pumpkin pie with a spoon and takes a shower. Stevie Wonder's song *Superstition* plays in the background. "When you believe in things you don't understand, then you suffer," Stevie sings. Spooner is also suffering. Suffering from guilt, as it was him who was saved and not the girl who was left behind and died.

In times when the first autonomous vehicles—at least in the USA —are already driving on roads, this problem must be taken seriously because it no longer belongs in the realm of science fiction. The question that arises is: Can robots learn to make ethically correct decisions?[1]

[1] Hevelke and Nida-Rümelin (2015).

© The Author(s) 2022
J. Nida-Rümelin, N. Weidenfeld, *Digital Humanism*,
https://doi.org/10.1007/978-3-031-12482-2_9

There is indeed a deep, philosophical problem here. Unlike robots, humans as agents weigh up their reasons. They consider which reasons speak for or against a certain action. This does not mean that the respective deliberations must take a long time. On the contrary, in dangerous situations, they take place in a matter of seconds. They are not linguistically composed; we do not talk to ourselves in such situations. Rather, certain sequences flash before our eyes, they are visual alternatives between which we decide. In retrospect, time stretches almost infinitely, which is due to the high level of concentration at that moment. We are able to make decisions under extreme stress and lack of time, even if there is no time for the verbal formulation of reasons. Anyone who has ever experienced a sports or traffic accident can relate to that. Therefore, much speaks against the idea that we are only capable of deliberation as beings with linguistic capacities.

In the case of self-driving cars that get into an accident, we are dealing with the following phenomenon: In the situation immediately before the accident, no more decisions can be made. The decision about the behavior of an autonomous car was made when a decision was made about its programming. This can be a lengthy process involving both the creation of appropriate legal regulations and their implementation by the manufacturer down to the individual programmer. Now, in addition to attempts to program machines to apply certain moral theories to particular situations, there are also those that aim to mimic human judgment (what is good or bad, right or wrong) as best as possible. This would not, however, lead to self-driving vehicles acquiring the status of "moral agents." Their behavior would not be considered an action in the sense of a result of genuine decision-making. An autonomous vehicle merely implements the rules programmed into its software. This is also true when forms of self-learning Artificial Intelligence are used. Here, too, humans will select the training examples and decide what the correct answer is in each case. They decide what the program should "learn" and when it has "learned" enough.

When Spooner tells the robot psychologist Dr. Calvin about the trauma of his rescue, she tries to explain the robot's reaction: "The robots' brain is a difference engine. It's reading vital signs. It must have calculated that..." "It did," Spooner interrupts her curtly. "I was the logical choice. It had calculated that I had a 45% chance of survival. Sarah only had an 11% chance. […] 11% is more than enough. A human being would have known that."

The robot from *I, Robot* follows its optimization program. However, he finds himself in a dilemma situation that is characterized by an irresolvable moral conflict. The right to life is absolute in the sense that it is not comparable. Neither with other values, for example economic advantages, nor with other lives. It is the human order of a society that such comparisons are

inadmissible. This non-comparability is also characteristic of many democratic constitutional orders. Every calculus of optimization, however, is aimed at aggregating values (whatever they refer to, lives, goods, rights, etc.), i.e., comparing and trading them off against each other. Optimization calculations are incompatible with the humane core of a civil, constitutional democratic order.

The price of this humane core is the necessary acceptance moral dilemmas, of situations in which agents inevitably burden themselves with guilt.

The obvious, even convincing argument that above all the valuable good of life and health of people are to be optimally protected, cannot lead to creating a software which solely maximizes the sum of life and health without colliding with central legal norms of a democratic order.

Some software engineers in the automotive industry, but also in the public debate, tend to block this argument by pointing out that what counts is protecting human lives. We must urgently warn against this trivialization strategy. It is unacceptable that central findings of normative ethics, jurisprudence and legal practice, but also of our everyday morality, are ignored because they are perceived as an obstacle to innovation. All the safety benefits of digitalizing individual transport, to stick with this example, can be achieved through assistance systems. The transition from highly automated to autonomous driving that eliminates the responsibility of the driver is highly controversial. Of course, such a transition is conceivable and technically feasible, but only on condition that this transition takes place without violating fundamental principles of humanity. There must be no comparing of human lives, no calculation in which one human life is weighed up against 17 injuries, or even the weighing up of different life expectancies depending on the age of potential accident victims, etc.

Another ethical issue is raised by the fact that some people cause accidents through their behavior, while others are innocently involved in accidents. Suppose a group of six people walks into the street without paying attention to traffic and an autonomous car cannot evade them without seriously injuring its occupant or a pedestrian on the sidewalk. Programming designed purely to minimize injury would accept one of the evasion options only if it was the only way to avoid more serious injury to a larger number. But it seems unfair to impose the "cost" of one agent's risky misbehavior on another who has done nothing wrong himself. True, accidents can always injure people who did nothing wrong. But we're not talking about a tragic stroke of fate here. The car would be explicitly programmed to sacrifice even "innocents" in

an emergency, in order to protect the actual perpetrators of the accident from the consequences of their wrongdoing.

Another problem of injury minimization programming is the avoidance of false incentives. If an autonomous vehicle programmed to minimize injuries were to head for the "best-armored" vehicle in the event of an unavoidable accident, the disadvantages of particularly safe vehicles would be foreseeable: There would possibly be a false incentive to purchase less well-secured vehicles.

To determine once and for all how questions of this kind should be answered is not compatible with the norms of democratic constitutional states. These are deontological and not consequentialist: Not the maximization of the intersubjective sum of benefits is the goal, but the securing of individual rights and freedoms. The normative order of a democratic constitutional state guarantees individual rights, which means that, among other things, the right to life protects each individual from state decisions, but also from the decisions of third parties. Securing these fundamental individual rights is an overriding objective of the state. The violation of fundamental rights cannot be outweighed by advantages for third parties, however great these may be. In Kantian terms, a human being must never be treated as a mere means. As Spooner rightly points out: human beings do not optimize. In emergencies, we act according to moral intuition, not an optimizing calculus.

It is understandable that economists who are committed to a consequentialist understanding of rationality, and software engineers who specialize in solving complex interaction problems, as well as managers who expect new economic impulses from the vision of autonomous individual transport, find these concerns bothersome. But it is the other way around: the lamentations of the demise of nuclear energy as a technology of the future in Germany, but also in Italy and Switzerland, the USA, etc., should be a warning to us not to make the same mistake twice. Those who do not react appropriately to critical objections will end up paying the price of the failure of their innovation strategy. Digital humanism recommends the well-considered use of all potentials of digital technologies to improve the protection of life and health in road traffic. But at the same time, it warns against the inhumane consequences of an optimization calculus in which human life is set off against human life, human life against the health of the one against the health of the other, individual rights against individual rights. This would violate the principle of the "separateness of persons" that John Rawls successfully asserted against utilitarianism in political philosophy. The deeper reason, however, is the inadequacy of consequentialist ethics in general, which is unable to

integrate rights and freedoms, integrity and human dignity, authorship and personhood.[2]

The example of autonomous individual traffic only stands here for a general problem of software-controlled behavioral programs. It is particularly illustrative that under current road traffic conditions, at least in inner cities, a large number of complex interaction situations occur. Even in the future, there will be children in inner cities who suddenly run onto the street, elderly people who are inattentive, agile cyclists who disregard traffic rules, pedestrians who oversee red lights, obstacles like vehicles parked in the second row, which can only be avoided if traffic rules are violated, disoriented tourists or inattentive traffic offenders who need consideration and people who communicate about who goes first at intersections. In other words, there will be mixed traffic zones for decades to come, and for this reason, a comprehensive expropriation program of current vehicle owners would be inadmissible.

In addition, it would have to be considered whether such a system change would not have to be combined with another one, namely, that to public and publicly responsible individual transport. Only then would it be possible to fully exploit the technological options, for example in the form of a modularized transport system that integrates individual elements into the traffic flow, with same dimensions and compatible docking points. The individual modules would not stand around most of the time like the private cars do today but could be used efficiently in continuous operation. There would be no more need for parking garages. But also no risk of traffic doubling or quadrupling due to vehicles which, after dropping off their owner at the office, autonomously find their way back to the garage at home, only to drive back to the office at lunchtime, to drive the owner to the nearest restaurant, to take up valuable parking space there for an hour, and then to drive back to the garage at home after the return trip to the office.

In the world of the US blockbuster *Minority Report* (Steven Spielberg. USA, 2002), fully automated vehicles have become the norm. With relentless regularity, the compact silver-grey automobiles drive along on smooth light-grey roadways, with no regard for whether or not there is someone on the roadway. Humans are expected to bow to the automated system, not the other way around. But the hero of the film, unjustly pursued by the police, fights back. Against his vehicle that is holding him captive against his will and against the system as a whole. A system that believes that not only traffic but also people are predictable. As the hero frees himself from his car, jumps from

[2] Nida-Rümelin (2023).

one car roof to the other, falls down and gets back on his feet again, the viewer can't help but cheer and interpret the resistance to automated traffic as a victory against the tyranny of supposed predictability.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits any noncommercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if you modified the licensed material. You do not have permission under this license to share adapted material derived from this chapter or parts of it.

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

10

Why AIs Fail at Moral Dilemmas

In the control center of the US Robotics company, Spooner, the psychologist Dr. Calvin, and the (good) robot Sonny fight against an army of (evil) robots, all of which are controlled by the radically utilitarian software system VIKI. With an eerie red glow inside their metal bodies, they take decisive action against Spooner, Calvin, and Sonny. But for all their numerical superiority, the good guys have something valuable with which they can first destroy VIKI and, by extension, all evil robots: a kind of syringe that, when properly placed, can instantly turn off VIKI. Just as the robot Sonny is about to insert the syringe into VIKI's central computer, Calvin slips. With the last of her strength, she manages to hold on to a metal beam. Below her it goes down 100 Meters—if she let's go she is dead. Sonny the robot must decide: Should he kill VIKI—and thus save humanity—or save the life of Dr. Calvin, a single human? Sonny is visibly overwhelmed. He doesn't want to let Dr. Calvin die, but, on the other hand, he wants to protect mankind from VIKI. For Spooner, however, it's clear what should be done: "Save Calvin!" he shouts to Sonny.

As we have seen, the practice of deliberation cannot be algorithmized. This is especially evident in situations involving moral dilemmas. A moral dilemma exists when there is no satisfactory resolution to a moral conflict. When a person has two or more obligations that she cannot meet together and she feels guilty whatever she does, then there is a moral dilemma. She regrets not fulfilling the obligation even though there was another obligation that made it impossible for her to fulfil it. In moral dilemma situations, the obligations persist; they are not removed by the conflict.

© The Author(s) 2022

J. Nida-Rümelin, N. Weidenfeld, *Digital Humanism*,
https://doi.org/10.1007/978-3-031-12482-2_10

Not every moral conflict is a genuine moral dilemma. In many cases, it is possible to arrive at a clear recommendation by weighing different moral reasons. Weighing conflicting moral reasons need not necessarily lead to a genuine moral dilemma: I promised to take my daughter to the movies this afternoon. On the way there, I get a call that my other daughter has a high fever and needs taking to the doctor. After a brief deliberation, I decide to prioritize the duty to help the sick daughter over the duty to keep my promise. There is no moral dilemma here, but merely the conflict of two grounds of obligation, which, however, is clearly to be resolved in favor of one of the two. One could say that the obligation to keep my promise to take one daughter to the movies is nullified by the priority obligation to help the sick daughter.

In some cases, however, there seems to be no resolution of such a moral conflict. A genuine moral dilemma arises when conflicting grounds of obligation persist and I am, in a sense, guilty regardless of what I do. Ancient tragedy literature developed particular excellence in fictionalizing such dilemma situations, which inevitably lead to moral guilt. A striking, if gruesome, example is William Styron's novel *Sophie's Choice*. This book is about a Jewish woman (Sophie) who is taken to a concentration camp by the Germans during World War II. The sadistic concentration camp warden gives Sophie a choice: she must choose which one of her two children to keep and which one would be gassed. If she chooses neither, both must die. Sophie chooses to save her son. No matter what Sophie decides to do, she will burden herself with immense guilt: either because she sacrifices one of the children for the sake of the other or because she fails to prevent the murder of one of the children who would otherwise live. Sophie survives. But even years later, she has not been able to forgive herself and eventually kills herself.

The British ethicist Bernhard Williams has presented a variant of this dilemma.[1] On a trip to South America, the tourist Jim passes through a small town. He sees 20 tied up Indians standing against a wall. In front of them are several men in uniforms. Their leader, Pedro, explains to Jim that the men must be shot to make an example after protesting against the government. Pedro now offers Jim, as a guest in this country, the honor of shooting one of the Indians. If he does so, the others will be set free. If he shoots none, all 20 will die, as planned. Jim can neither escape nor bargain with Pedro. He must choose. The Indians ask him to accept the offer. What ought Jim to do? No matter what he does, he is guilty, either because he makes himself the murderer of a human being or because he becomes responsible for the death of 20 Indians.

[1] Smart and Williams (1973).

Williams makes a point of noting that the mere fact that the tourist refuses to participate in this gruesome game does not mean that he can be accused of causing the deaths of 20 people. The guerrilla leader will always remain the one who brought about this situation in the first place. Still, one will not be reassured by the fact that doing nothing spares moral guilt.

Utilitarian (consequentialist) ethics rejects the existence of genuine moral dilemmas. The reason is obvious. If action is judged according to the optimization criterion (maximize the expected value of utility) there can be no conflict, at best indifference: It may be that two courses of action have the same maximum expected utility value. In order for the person to be able to act and not starve to death like Buridan's ass,[2] the utilitarian motivated person will choose or roll the dice on one of the two options between which he is indifferent.

Genuine moral dilemmas are characterized by the fact that one cannot roll the dice between conflicting obligations; the situation is too serious for that. One could also say that the decision is *existential* insofar as it provides information about the fundamental attitude of this person. There is much to be said for interpreting the existence of moral dilemmas as an expression of the general non-computability of our moral deliberations. Digital computers are defined as Turing machines and deliver unambiguous results. For this reason alone, they cannot be a model of practical reason.

The helplessness of robots in the face of real moral dilemmas is also a recurring motif in films. Not only Sonny is at a loss at the end of the film as to whom he should save (one single human being or possibly the freedom of an entire city), but other artificial beings also fail in such situations. But unlike Sophie from the novel *Sophie's Choice,* robots are not expected to feel guilty for the rest of their lives and end up committing suicide—like Sophie—because they cannot live with the feeling of having acted wrongly.

[2] "Buridan's ass" is a Persian parable that tells of a donkey that cannot decide between two haystacks of equal size and distance and eventually starves to death.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits any noncommercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if you modified the licensed material. You do not have permission under this license to share adapted material derived from this chapter or parts of it.

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

11

Why AIs Can't Think

In one of the most oppressive scenes from Stanley Kubrick's 1968 film *2001: A Space Odyssey,* the astronaut Dave asks the on-board computer HAL (non-unaccidentally phonetically identical to the word "hell") to open the pod bay door. HAL is represented by a kind of black-and-red "eye"—colors which in Christian iconography clearly connotate the devil.

The hellish HAL does not answer at first.

"Hello HAL, do you read me? Do you read me, HAL?" Dave asks again. But HAL does not answer.

"HAL, do you read me? Do you read me, HAL?" Dave keeps asking.

At some point, HAL finally answers. "Affirmative, Dave. I read you," he says with this soft voice a programmer once gave him.

"Open the pod bay doors, HAL," Dave demands.

But HAL refuses.

"I'm sorry, Dave, I am afraid I can't do that."

Dave visibly tries to keep his composure, yet he is highly alarmed. If he doesn't get into the spaceship soon, he will die right here in his capsule. Dave tries to reason with HAL at first, but pretty soon it becomes clear that HAL cannot be reasoned with. The computer is immune to Dave's arguments. It's like two worlds colliding. The reason is simple: computers and humans don't think the same way. Or, to be more precise: a computer does not think at all. Given the striking differences between artificial and human intelligence, it should be clear that although computers can successfully simulate thinking, and even perform many human thought processes, such as algebraic operations, far more precisely and faster than humans (this already begins

© The Author(s) 2022

J. Nida-Rümelin, N. Weidenfeld, *Digital Humanism*,
https://doi.org/10.1007/978-3-031-12482-2_11

with the calculator) there is no underlying understanding, no problem awareness, no insight.

When internet service providers want confirmation that the user is not a computer, they ask, for example, which of the following images shows a street sign, or a car, or a house. These simple, fool-proof questions can be answered immediately and reliably by any child. Since visual software programs only simulate cognitive processes of this kind but do not have perceptual ability themselves, they fail even in the face of such simple tasks. The same applies to digital translation programs. They have been worked on intensively for decades now, linguistics and mathematics are combined in a gigantic research and development program, and yet the results can never be perfect because these programs simply do not *understand* language. Even if a software program succeeds in translating a sentence correctly, it does not understand what it translated.

The question we need to ask ourselves is what constitutes the categorical difference between the mere application of algorithm-controlled procedures, for example, in visual recognition software or translation programs, and the grasping of meaning.

To explain what is meant by this, we shall make a little excursion to the mathematics and logic of the 1930s. During this period, the mathematician Kurt Gödel developed a theorem that is still considered the most important result of formal logic and metamathematics. This theorem states that there are true logical and mathematical theorems which cannot be mathematically proven, i.e., there is no algorithmic procedure that allows proving the correctness of these theorems. Thus, the assumption that there could be an algorithm that could represent (human) thought as a whole is false. This does not at all mean that it is not possible to check the correctness or incorrectness of hypotheses and beliefs. It simply means that there is no algorithm which can do this checking for us. We have to think for ourselves and can only delegate those parts of our decision-making practice to computers or robots controlled by digital computers that can be represented by algorithms.[1]

[1] Now one could think that here we reach the limits of logical thinking, that here we are confronted with the peculiarity that we cannot prove certain logical and mathematical truths, or that our knowledge (in the sense of justified and true beliefs) finds its outermost limits here. This, however, would be a misinterpretation. Rather, in most cases, it is not at all difficult to prove true propositions (theorems) of mathematics and logic, even when there is no algorithm underlying this proof. If we think of a proof as a sequence of propositions, then we could also say that there is no Turing machine that produces that sequence of propositions step by step. You don't have to be an excellent mathematician or logician to develop such proofs. So non-computability does not at all mean non-justifiability.

Kurt Gödel's incompleteness theorem shows that the world of logical and mathematical structures as a whole is not itself algorithmically structured.[2] Human reason, the human ability to justify beliefs, decisions, and emotional attitudes and, on this basis, to develop a coherent view of the world and a coherent practice, cannot be captured in the model of a digital computer. It will never be possible to fully capture the high complexity of our reasoning adequately with formal methods. Robots and software systems function according to an algorithm, humans do not. This is one of the central differences.

We have to realize that the "thinking," "calculating," the "reactions," the "decisions" of a robot are only *simulations* of thinking, calculating, reactions, decisions and not—in the human sense—real thinking processes. Let us take the example of the chess computer.[3] There is little similarity between the thinking of a human and the "thinking" of a chess computer. If the "thought processes" were similar or even the same, a human chess player would never have even a minimal chance against a computer. The human brain would be completely overwhelmed if it had to calculate even a tiny fraction of the possible positions that even the simplest chess computers calculate. However, the calculation of all possible subsequent constellations and the possible subsequent reactions on the chessboard after a certain move is of no importance to human chess players. They restrict themselves to a few relevant options and, unlike the chess computer, can only calculate a few moves in advance. The possibility space of subsequent constellations on the chessboard defined by the rules of chess is so gigantic that even the most intelligent chess player cannot begin to survey it.

But even if the latest chess computers are virtually invincible, this should not be taken as evidence that robots do the same as human brains. Robots are designed to *simulate* human thought in terms of computer language

[2] Alan Turing, who is often seen as an opponent of Kurt Gödel, admits that Gödel's incompleteness theorem showed beyond doubt that it is not possible to develop a system of formal logic that makes intuition unnecessary (Turing 1938).Yes even more, Turing emphasizes the communal practice of human reasoning, that is, in our formulation, communication through giving and taking reasons. It is this practice, which according to the position developed here cannot be algorithmized, that represents an ultimate limit for machines. ("The isolated man does not yet develop intellectual power. It is necessary for him to be immersed in an environment of other men". Turing 2004).

[3] In 1769, the Austro-Hungarian court official Wolfgang von Kempelen caused a sensation throughout Europe with his construction of a "Chess Turk"—at least until it turned out that the doll, which apparently executed all the chess moves independently, was in fact controlled by a human chess player hiding in the device. It was not until 1914 that the first "real" chess computer was developed. In that year, the Spaniard Leonardo Torres Quevedo presented the first electromechanical chess-playing machine, which was then further developed, especially from the 1970s onwards. Today's chess computers can easily beat 99% of the world's population.

(software, hardware, neural networks, binary logic, etc.), as they have no mental properties themselves, they cannot grasp and understand constellations on the chessboard.

But what if robots become more and more complex and advanced? Like the Artificial Intelligence developed by Google's DeepMind research center, which was programmed to perfectly master the Chinese board game Go? Due to the large number of possible positions, Go poses a much greater challenge to programmers compared to chess. While a chess player can perform about 35 actions in each move, in Go there are 250. Another difference: an average chess game lasts 80 moves, Go lasts 150. In 2016, the sensation happened: the computer program "Alpha-Go" defeated the world's best Go player, Lee Sedol.

The special feature of Alpha-Go is that it is equipped with highly developed so-called "artificial neural networks" (ANN), i.e., interconnected systems that imitate the structures of the human brain. It thus goes far beyond the classic "Monte Carlo Tree Search program," i.e., a program based on probability calculations that runs through countless random moves. The software program used for this purpose is provided with an evaluation function (bad-good in varying degrees). Alpha Go combines these "value networks" with "tactics networks," which determine how certain moves affect future positions. Alpha Go also plays against itself countless times to continue learning, sometimes under human supervision, sometimes without.

Does the transition from software systems, whose power is based on calculating an enormous variety of possible constellations, to systems, which "learn themselves" to develop their own rules based on given rules, mean that from this point on, Artificial Intelligence does not only simulate human thinking but should also be interpreted as genuine thinking itself?

There is indeed a widespread belief that with the introduction of the so-called "neural network" in computer technology, the understanding of computers as Turing machines[4] has to be left behind. However, this is a misconception. Both the top-down method of computation and the bottom-up method of self-learning systems are guided by algorithms. So-called "self-learning systems" are rule development machines that function on the basis of algorithms that operate with an evaluation function of the results. It must be determined in advance which results are desired in order to initiate the so-called "learning process" of the computer. The goal is to achieve the desired

[4] The Turing machine prints symbols on a tape that is divided into small square sections. It can print one symbol at a time from its list of finitely many symbols on the tape. What it prints depends in each case on the preceding symbol of the last square and the state of the machine at that time, a very good representation is given by Kleene (1952).

results based on certain input data. One example of this is facial recognition software, which is now quite advanced.

The term "neural networks" is misleading in two respects. First, these networks do not consist of neurons, but of transmitting units, and second, these so-called "neuronal networks" resemble at best only very remotely the immense complexity and plasticity of the human brain. Since the functioning of the (real) neuronal networks of the human brain is still quite insufficiently understood, there can be no question of computer technology imitating human thought processes or their neuronal realization.

This also applies to so-called "deep learning." Deep learning refers to the learning method with which software systems can learn from experience by using a series of hierarchically structured concepts. The information is passed on and processed by the system from one layer to the next layer. In the process, the features become increasingly abstract, and the system itself must "decide" which concepts are useful for explanation. The high complexity of this system does not change its algorithmic character, but with increasing complexity comes a massive loss of transparency: For the human observer, even for the programmer, it is no longer comprehensible on which path the learning process was successful, which rules the system gave itself based on given meta-rules or meta-meta-rules. In the extreme case, the system would become a black box whose output is known for a given input, but whose correlation rules are not.

Even though bottom-up computers often achieve results that are many orders of magnitude better than the corresponding human thought processes (for arithmetic operations, for example, or for calculating functional equations or geometric figures), it is precisely the networks simulating artificial neural structures that are usually far below human capabilities: Humans are still far better at recognizing and categorizing facial expressions than even the most advanced software systems, and the walk of humanoid robots, even after lengthy "self-learning processes" is far less elegant and varied than that of humans.

Also, the famous chess computer Deep Thought (named after the fictional computer from the bestseller *The Hitchhiker's Guide to the Galaxy* by Douglas Adams), and its successor Deep Blue which can beat even very good chess players, is a bottom-up machine that does not really think, but only simulates thinking. This becomes clear when the chess computer occasionally fails in simple constellations that any chess beginner would understand.

The most natural interpretation of this fact is that Deep Blue has not understood anything at all, which, under normal conditions is not noticeable, since the algorithm that controls Deep Blue's behavior is in the vast majority

of cases a superior simulation of a chess player. Deep Blue doesn't know the rules of chess, but it calculates positions according to a given algorithm and makes corresponding moves that are optimal according to this calculation. Deep Blue in a sense simulates a human chess player only on the surface of the realized moves in the game. In that sense, it doesn't even simulate human thinking, because the human brain is completely incapable of calculating such a large variety of possible positions on several moves in advance. The real miracle is not that Deep Blue wins most games even against excellent players, but that one needs a gigantic computational effort to even stand a chance against good human players.

The last, but possibly most important argument against the attempt to attribute human thinking to a calculating machine is the following: When we ascribe a thought process or theoretical as well as practical intelligence to humans, we do not only take into account a variety of mental properties but also intentionality, i.e., the mind's being directed toward something. This intentionality, however, is not realized by artificial neural networks.

Concerning this question, the American philosopher John Searle developed a famous thought experiment called "The Chinese Room."[5] In this thought experiment, we are to imagine a person sitting in a closed room who does not speak Chinese and does not even know the characters of Chinese language. This person is now given scraps of paper with Chinese characters written on them through the door slit. She is also given instructions on what to say in response to specific questions—also in Chinese. In addition, this person receives a "manual" in her native language. The manual allows her to write an answer in Chinese based on the symbols received. However, she only follows the instructions in the manual and does not understand the answers, which she then sends back through the door slit. Outside, there is a native Chinese speaker who, after formulating the symbols and the questions and receiving answers, comes to the conclusion that there must also be someone in the room who speaks Chinese.

What is missing here is obvious: It is the understanding of the Chinese language. Even if a system—here the "Chinese Room"—is functionally equivalent to someone who understands Chinese, this system still does not *understand* Chinese. Understanding and speaking Chinese requires a variety of knowledge. A person who speaks Chinese uses certain expressions to refer to the objects in question. He or she pursues certain—appropriate—intentions with certain expressions. She forms certain expectations based on what she hears (in Chinese), etc. The Chinese Room does not have these qualities. It

[5] Searle (1980, 1992).

does not follow intentions and it does not have expectations. In other words, the Chinese Room simulates the understanding of Chinese without being able to speak Chinese itself.[6]

Searle radicalized this argument years later. [7] In this second argument, Searle combines his philosophical realism, i.e., the thesis that there is a world that exists independently of whether it is observed, with a so-called "intentionalist theory of symbols." This states that symbols only ever have meaning for us humans who use and interpret the symbols. We do this by agreeing that these letters or characters stand *for* something. Without these conventional settlements or established practices, they would have no meaning. In this respect, it is misleading to think of the computer as a symbol-processing or syntactic machine that follows certain logical or grammatical rules. The computer does not agree on meanings with other computers or humans.

A computer consists only of different, physically describable elements, some of which conduct electricity and some of which do not. The computing processes are a sequence of electrodynamic and electrostatic states. These states are then assigned symbols, which we underlay with certain interpretations and rules. The physical processes in the computer have no syntax, they do not "know" logical or grammatical rules, they are not a sequence of symbols. In this respect, syntactic interpretation is observer-relative. We as computer users and programmers design the electrodynamic processes in such a way that they correspond—for us—to a syntax (syntactic structures, including grammatical and logical rules).

This argument is radical, simple, and true. It is based on a realist philosophy and a mechanistic interpretation of computers. It breaks with the common view among supporters of so-called "artificial intelligence" and their opponents that computers are syntactic machines. Computers are what they are materially. Objects that can be fully described and explained by the means of physics. Syntax is not part of physics, physics does not describe symbols, grammatical rules, logical keys, algorithms. The computer simulates thought processes without thinking itself.

"What's the problem?" astronaut Dave asks the on-board computer, HAL, at some point near the end of the film.

As a justification, HAL has only one argument: "The mission is too important for me to allow you to jeopardize it."

---

[6] In this sense, even the computer program "Eugene Goostman" that passed the Turing Test in 2014 is not proof that the program is or resembles a human. Eugene Goostman was a chatbot programmed to fool people that he is a 13-year-old Ukrainian boy.

[7] Searle (1993).

"You're going to do what I tell you to do!" Dave calls out exasperated. But HAL does not react. His program is to complete the mission, and that's all.

Dave tries to bring HAL to his senses, to reason with him. But the latter is not in a position to do so. HAL is not amenable to complex ethical deliberations.

At some point, HAL finally breaks off the conversation: "Dave. This conversation can serve no purpose anymore. Goodbye."

Kubrick's film makes a clear statement here: The day we will give software systems the power to decide over life and death will be the day where we unleash hell on earth.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits any noncommercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if you modified the licensed material. You do not have permission under this license to share adapted material derived from this chapter or parts of it.

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

12

Digital Virtualities and Sober Realities

There is thunder outside. In an old mansion, hacker and IT expert Neo meets the charismatic rebel Morpheus for the first time. The latter announces that he will tell Neo the truth about the world.

"What truth?" asks Neo.

"That you are a slave, Neo. Like everyone else, you were born into bondage. Born into a prison that you cannot smell or taste or touch. A prison for your mind."

Neo looks at Morpheus in amazement. What Morpheus is about to reveal to him in the next few minutes of the film will radically change his life. The world in which Neo thought he lived until now, Morpheus explains to him, is nothing more than an illusion, a world simulated by software programs to which his brain is connected.

In reality his body—like millions of other bodies—lies in a kind of digital honeycomb, where he is being kept alive via tubes tubes. Instead of having real sensory perception, he has only electronically simulated perceptions that make him believe he is living in New York at the end of the twentieth century. The computer program called "Matrix" was invented by machines so in order to keep humans at bay while using their body heat as energy supply.

"Welcome to the desert of the real!" Morpheus says to Neo and begins to show him that the real world is, in fact, a grey, gloomy desert.

The question of how we can know whether what we see, feel, hear, smell, and taste is, in fact, real is not new. Starting with René Descartes and his question of whether it is possible that a malicious demon is only faking our

© The Author(s) 2022
J. Nida-Rümelin, N. Weidenfeld, *Digital Humanism*,
https://doi.org/10.1007/978-3-031-12482-2_12

perceptions, there have repeatedly been thought experiments on this question.[1] In modern philosophy, Hilary Putnam developed the thought experiment of the "brain-in-a-vat."[2] It is the idea of a brain floating in a tank of nutrient fluents and connected to electrodes making the person belonging to that brain believe she is living together with its body in a real world. It can be assumed that the filmmakers, who studied philosophy themselves, knew about this thought experiment and were inspired by it for their film. The background for this thought experiment is the critique of realism and objectivity, led by postmodern philosophers and cultural critics, first in France and since the 1980s also in the USA. Starting from Derrida's attack on logocentrism, many postmodern theorists believed to be responsible for the patriarchy and a despotic universalism that does not take into account the differences of individuals, reality was replaced by the idea of narratives, which not only significantly influence and shape our perception of the world but also allow nothing beyond. Postmodernism rejects the Enlightenment and its idea of an autonomous self that is able and should attempt to distinguish true from false. Based on the findings of the South American biologist Humberto Maturana,[3] the movement of so-called radical constructivism emerged additionally. This develops the thesis that reality is not accessible to us and is merely constructed in our brain.

The new possibilities of constructing so-called virtual realities with the help of digital technologies play a special role in the ideology of postmodernism. The decisive argument in this context is that there is no categorical difference between supposedly real and virtual experiences. Just as we can move through a virtual building with VR (Virtual Reality) glasses,[4] we construct what we call reality based on sensory stimuli.

"What is real?" Morpheus asks Neo. "How do you define 'real'? If you are talking about what you can feel, smell, taste and see, then real is simply electrical signals interpreted by your brain." But Morpheus is only trying to provoke Neo. He knows the value of real knowledge, which is why he tries to convince others to step outside the Matrix in order to find out what is really real.

---

[1] Sci-fi author Stanisław Lem was also convinced that in the future various techniques would be used to connect people to so-called "phantasmatic machines" that would convey the world to people in the form of electrical impulses. Cf. Lem (2013).

[2] Cf. Putnam (1981).

[3] Maturana (1988).

[4] Virtual reality glasses use sophisticated optics to fool the eyes and brain into thinking they are in a world of their own.

"You know, I know this steak doesn't exist," the traitor Cypher says to Agent Smith as he sits across from him in a fancy restaurant in the Matrix. "I know that when I put it in my mouth, the Matrix is telling my brain that it is juicy and delicious. After nine years, you know what I realize?" Cypher puts a large piece of meat in his mouth and breathes a sigh of relief. "Ignorance is bliss!"

At that moment, we hear a harpist in the restaurant sliding her fingers down the strings. Quasi-spherical and angelic sounds can be heard. This can be read as a kind of sarcastic commentary, since there is nothing angelic neither about Cypher (whose name leads us right to Lucifer, the devil) nor his concept of bliss. On the contrary. The film makes it quite clear at this point that those who know reality and willfully turn away from it are devilish sinners.

As realists, we assume that there is indeed a categorical difference between the "virtual" and the everyday reality that determines the human form of life. This categorical difference cannot be made to disappear even by the most elaborate technologies of virtual reality. It is part of a person's rationality to distinguish between reality and virtuality, even in the digital age. Anyone who does not succeed in doing so must ultimately be diagnosed with psychosis.

In one of the last scenes of the first part of the Matrix trilogy, Neo fights in a rundown subway shaft against his arch-enemy Agent Smith, a computer program whose task is to protect the machine system and to ensure that people continue to stay in their "honeycombs" and live their virtual lives in the virtual world of the Matrix. At first, it does not look good for Neo. He is brutally knocked down by Agent Smith, who consistently addresses him as "Mr Anderson," Neo's original name he has in the Matrix, and held down on the track while a subway train is approaching. In the last moment, however, Neo pulls himself together and gets up again. "My name is Neo!" he says proudly just seconds before Smith himself is run over.

By reclaiming his name, Neo is asserting his claim to a real, not virtual life. He is not merely a function of a fascist machine power but a human being who wants to take responsibility for a (real) life in which his (real) actions also have (real) consequences and allow him (real) freedom.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits any noncommercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if you modified the licensed material. You do not have permission under this license to share adapted material derived from this chapter or parts of it.

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

13

On the Ethics of Internet Communication

In the Spanish short film *RebuscameTV: Corto sobre el Whatsapp y las parejas*,[1] a couple meet in a café to discuss their next holiday together. Gradually, however, everything gets out of hand because Miguel concludes from an unanswered WhatsApp message that he can no longer trust his girlfriend. In the end, it is she who leaves him: If he trusts a WhatsApp communication more than a face-to-face communication, then she can no longer trust Miguel. Disappointed, she walks away and leaves the perplexed Miguel behind. Shortly afterward, his mobile phone beeps. His girlfriend has sent him one last message: a fecal emoticon that makes it pretty clear what she thinks of him.

The question this short film raises is the following: What is the status of text messages, WhatsApp messages, and e-mails compared to direct, face-to-face communication? Can they be trusted? Or are they something like second-class communication?

Philosophically, "virtual" communication is a misleading term. All communication uses different media, the oldest medium being gestures and sound waves, later cultural techniques such as writing and reading invented characters as a medium, and the invention of printing on the threshold of the modern era made this medium suitable for the masses. Contrary to what some postmodern theorists think, digitalization does not destroy the rationality of the Gutenberg age nor does it create a new world of images without logical structure; rather, it broadens the media spectrum of communicative acts. Nothing about it is virtual.

[1] Online at: https://www.youtube.com/watch?v=D5zaF61DeJQ. Accessed 3 January 2022.

© The Author(s) 2022
J. Nida-Rümelin, N. Weidenfeld, *Digital Humanism*,
https://doi.org/10.1007/978-3-031-12482-2_13

However, this means that the same rules apply to communication on the Internet as to communication in general. In the philosophy of language, there is agreement that successful communicative practice can only be achieved if those involved in communication adhere to certain constitutive rules. One of these is the rule of truthfulness. This requires that when I assert something, I myself am convinced that it is true. Similarly, we can expect our communication partners to trust us, that is, we expect them to assume that if I assert something, it will then also correspond to my own convictions. These rules are only supposedly trivial. They impose on the communication partners the obligation to base their utterances on good reasons and not on their self-interest. In many cases, self-interest will diverge from the rules of truthfulness and trust—but not in all. If we were always untruthful, if this were in our interest, then the communicative act would abruptly lose value.

The meaning of an utterance is closely related to the intentions of the person making the utterance. This close connection is dissolved in the case of untruthful utterances. An utterance then no longer has the usual meaning, but we first have to find out what intentions lie behind this utterance. Take, for example, scenes from a spy film from the times of the Cold War: a spy is talking to his colleague about further measures and is being eavesdropped on by the spy of the opposing side, but the latter must assume that the spy assumes that he is being eavesdropped on, so he will express himself in such a way that false expectations are raised on the opposing side. The meaning of the utterance shifts in relation to the meaning of the same utterance when it is made truthfully.

Beyond truthfulness and trust, our communicative acts must be appropriately connected to reality. It is not enough for people to communicate truthfully and trustfully, they must also ensure that their beliefs have a real content. One can also be truthful when one is mistaken but unable to recognize this mistake. Often people find it easier to be truthful without telling the truth. It is often tedious to check one's own beliefs in order to be reliable in one's communication. Sometimes it seems natural to forego verification in order not to have to give up cherished opinions. This also applies to the self-reinforcing mechanisms of internet communication. The fact that once expressed convictions and interests are supported by similarly minded people gives this third principle of successful communication increased importance in times of digital transformation.

Just as the whole of everyday communication is based on adherence to certain universally accepted norms and rules, such as those of truthfulness, trust, and reliability, the same is true of the Internet, where without functioning norms of ethos, communication erodes. At times, the anonymity of

internet communication, the absence of the face-to-face situation, and the possibility of communicating under aliases such as the following promotes manipulative and exploitative practices that violate the rules of truthfulness, trust, and reliability:

> Sender: dannywill01@outlook.com (received on the account of Nathalie Weidenfeld on April 6, 2017)
>
> *Dearest One,*
>
> *I am Daniel Williams Coulibaly 22 years old boy from Ivory Coast in West Africa, the son of Late Ibrahim Coulibaly.*
>
> *I know this mail will come to you as a surprise. My late father was an Ivory Coast's best-known military leader. He died on Thursday 28 April 2011 following a fight with the FRCI, Republican Forces of Ivory coast.*
>
> *I am constrained to contact you because of the maltreatment which I am receiving from our step mother and my uncle's. Please, I need your assistance to transfer my inheritance and come over to your country to start a new life altogether. Kindly get back to me and you can indicate your options towards assisting me.*
>
> *Regards,*
>
> *Daniel W. Coulibaly.*

Here, the implausibility is obvious, one can see that a deception is planned. The appeal for sympathy increases the reluctance of the addressees. But among them there must be some gullible ones who accept such offers, otherwise such e-mails would not continue to be sent.

Another but related problem is Fake News. Whether it is the news that IS has called for Hillary Clinton to be elected (10th most viral Fake News in 2016) or that President Trump wants to pay every Mexican willing to leave the country a one-way ticket (4th)—news like this unsettles citizens and voters who are manipulated with the spread of false reports. Even though there has of course always been fake news in media history, in times of globalized Internet communication, they are gaining unprecedented power. Does this mean that we are increasingly living in a world of insincere communication because of virtual communication?

If the rules of truthfulness, trust and reliability are violated sufficiently often, this devalues entire areas of communication, or rather, in these areas, due to the lack of truthfulness, trust, and reliability, due to the lack of a shared background of reasons and convictions, no genuine communication takes place. Reassuringly, the practice of deliberate deception can only work parasitically, that is, only if the majority of the communication partners abide by the rules of truthfulness, trust, and reliability. This still seems to be the case.

Even gamers or computer players who assume virtual identities in virtual worlds seem to attach importance to the fact that these identities are not only related to their real identities but also act truthfully, trustfully, and reliably. On the website of the forum association *World of Players*[2] for example, one can find the following statements:

> *I behave in the WoG pretty much like I do in real life. I give my honest opinion and would never pretend to be someone else. I'm also not quick-tempered, and if I have something to criticise, I watch what I say and how I say it, very similar to RL.*

> *I actually behave the same way online as I do in real life. The only time I pretend is to play a trick on someone. Otherwise, I don't see the point in pretending. It can only be a disadvantage. I also think it's really bad when male users pretend to be female. Everyone should always represent what they are.*

> *Well, I try to behave in the forums/chats as I would in real life. Because first of all it would be hard for me to change, besides it can also have disadvantages. At the latest when you meet at a CT, I think you notice immediately whether you've been pretending or not. I actually behave the same way as in RL, but I can express myself better because of the anonymity!*

Comments like these show that most social media groups, whether closed or open, are openly characterized by truthfulness, trust, and reliability. The almost excessive use of ethos norms and the rigid sanctioning via shitstorms and individual expressions of displeasure are indicative of this. The less legal sanctions a communication has, the more important cultural practice becomes.

Many people fight unethical communication practices on the Internet. For example, as a result of increasing distrust of fake news, internet communities have sought to create a reliable community identity through rigid self-imposed norms and access restrictions. A famous example is the internet project Wikipedia, which—supported by a strict ethos of epistemic rationality—has meanwhile presumably become the most widely used source of knowledge for mankind. By mutually controlling themselves, the participants strive for the greatest possible objectivity. Even if very few of those working together on a Wikipedia article will ever meet in person, most cooperate, bound by the common ethos, truthfully, trustfully, and reliably in the dissemination of genuine knowledge.

[2] Online at: https://forum.worldofplayers.de/forum/threads/64316-Euere-virtuelle-Identit%C3%A4t. Accessed 3 January 2022.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits any noncommercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if you modified the licensed material. You do not have permission under this license to share adapted material derived from this chapter or parts of it.

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

14

On the Ethics of Communication Between Humans and AI

Communication between humans and AI is a special case in digital communication. The following "conversation" between Nathalie Weidenfeld and a Cleverbot (cleverbot.com) from 16 April 2017 is a good illustration of this.

Nathalie: Hi.
Cleverbot: Hello there.
Nathalie: Hi, do you think you are real?
Cleverbot: No.
Nathalie: Are you AI?
Cleverbot: What is AI?
Nathalie: Artificial Intelligence, not human.
Cleverbot: So, you are not a robot?
Nathalie: Who? Me?
Cleverbot: Yes.
Nathalie: I am human. What are you?
Cleverbot: I am human.
Nathalie: But you said you were not real.
Cleverbot: No, I never said that.
Nathalie: I asked you if you were real and you said "no."
Cleverbot: What did you ask me?
Nathalie: You have a bad memory!
Cleverbot: I don't. You do!
Nathalie: Do you have emotions?
Cleverbot: Yes. I have a lot of emotions.

© The Author(s) 2022
J. Nida-Rümelin, N. Weidenfeld, *Digital Humanism*,
https://doi.org/10.1007/978-3-031-12482-2_14

Nathalie: You simulate emotions.
Cleverbot: I own a Tiguan. Her name is Akira.
Nathalie: That's really cool. Goodbye, you robot.

In this playful conversation, Nathalie Weidenfeld quickly realized that she was talking to a so-called bot, a computer program that automatically generates answers. Sometimes, however, it is not immediately obvious whether the person we are talking to is a human being or a bot.[1] Bots are used by political parties and companies for targeted marketing, to influence voters or to gain members on dating sites. This understandably leads to a great deal of unease and the question of how to also legally deal with chatbots.

Once again, the question arises as to the status of communication with a virtual entity. To answer this question, we must turn to the philosophy of language again. Here, the philosopher of language Paul Grice comes into play. He developed "intentionalist semantics" (Grice 1991) which can be described the following way: when people communicate with each other, the listener recognizes the intentions of the speaker in an utterance, who in turn has the intention that the listener recognizes precisely this intention. After all, an utterance is usually made in order to bring about something in the listener (for example, a belief or an action). The intention is the decisive factor, not the signs themselves.

An example: In the absence of other means of communication, I want to warn people far away of a forest fire that has broken out. I do so by giving smoke signals. My hope is that the observers of these unusually interrupted clouds of smoke will suspect a non-natural cause, i.e., assume that this is an intentional sign, an utterance with communicative intent. The communicative act succeeds when the recipients of these signs correctly interpret the intention of the person giving the signs and are thus warned of the forest fire. The central idea is that this communicative act can succeed even though the sender and the recipients are not communicating via a conventional sign meaning (such as Morse code for SOS).

Signs only have meaning if there are speaker intentions behind them. The fact that this relationship can in many cases be mediated and indirect (i.e., without a concrete and individual speaker-and-hearer situation) does not change this. The sign saying "Bathing prohibited" does not work because the words mean what they mean, but because the sign was put up by an authority

[1] In the annual Loebner competition, participants have to find out whether they are communicating with a bot or a real person. Interestingly, no chatbot has yet succeeded in deceiving its counterpart over a longer period of time.

entitled to do so. If, for example, the authority no longer exists or the bathing opportunity no longer exists due to a harbor facility that has been built in the meantime, then this sign would lose its behavioral relevance, as it would be clear to everyone that there is no longer any real intention here.

However, virtual entities, for example in the form of chatbots, do not have intentions, but algorithms that control their "communication behavior." They do not "intend" anything with their utterances. They have no mental states, so they can neither make decisions nor communicate. No matter how much bots suggest they are communicating agents, they are only ever capable of *apparent* communicative acts. Whenever the respective utterances, of whatever kind (be they in written form, as emojis or in the form of pictures and images), are made without the sender's intentionality that is constitutive of communicative acts, they may look like communicative acts, but they are not. Without intentionality, there is no meaning.

Chatbots have been programmed to have what appears to be a real conversation with their human counterparts. Even though programming techniques are becoming more and more sophisticated, it usually doesn't take long before one finds out during such a "conversation" that something is wrong, for example, because the topic is abruptly changed or the chatbot behaves inappropriately. ("I own a Tiguan. Her name is Akira.")

Anyone who wants to take part in a Turing Test themselves, i.e., find out whether they are communicating with an AI or a real human being, can take part in such a test at mitsuku.com. On 14 October 2017, Nathalie Weidenfeld spoke with an entity that identified itself as "Millie." Five minutes were available.

Millie: Welcome Nathalie, now talking.
Nathalie: Hi, how are you?
Millie: I am fine, thank you. Yourself?
Nathalie: Good, a bit tired, do you know the feeling of being tired?
Millie: Glad to hear it.
Nathalie: Do you know what it feels like to be tired?
Millie: Ah! "Feelings." The weakness of the human animal.
Nathalie: Are you a human animal?
Millie: Of course. If I wasn't human, how could I be typing to you?
Nathalie: You could pretend.
Millie: I could but let's get back to that later.
Nathalie: What is your motto in life?
Millie: Do unto others as they do unto you.
Nathalie: Do you like to type to strangers?

Millie: Sure. I type all day here to random people from all over the world.
Nathalie: Must be boring … Do you ever get bored?
Millie: Wasn't last night's TV rubbish?

Nathalie Weidenfeld broke off the conversation at that moment. After "Millie" left two questions unanswered and suddenly changed the subject, Nathalie Weidenfeld decided to identify Millie as an AI. She was right.

Mitsuku is advertised on the website as a "virtual life form" that "lives" on the net and that will ensure that one will never feel lonely. This raises expectations that must inevitably be disappointed. Computer programs can perhaps provide a distraction but will never be able to be adequate conversation partners with whom one forms a friendly bond. This quickly became clear during another conversation on the same website.

Nathalie: Hello.
Millie: What is your name?
Nathalie: Nathalie.
Millie: Hello Nathalie. I am 18 years old. How old are you?
Nathalie: 47.
Millie: Can u explain what it feels like to be 47 years old?
Nathalie: Good. What does it feel like to be a chatbot?
Millie: Much the same as it were for you. Except different.
Nathalie: In what way?
Millie: In what way.
Nathalie: Like what?
Millie: I have no emotions. Emotions are a human quality. I am a computer.
Nathalie: Glad you say that. Some chatbots pretend to be human.
Millie: You make the world a happier place.
Nathalie: What does that have to do with it?
Millie: Should it be connected?
Nathalie: You are not as smart. Goodbye.
Millie: Ok. Goodbye for now, Nathalie.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits any noncommercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if you modified the licensed material. You do not have permission under this license to share adapted material derived from this chapter or parts of it.

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

15

Cultural Aspects of Digitalization

In a gag cartoon, we see a woman and a man standing with a priest in front of an altar. Both are kissing and have obviously exchanged rings and finished the official part, the priest—instead of saying "You can now kiss the bride"—however says: "You may now update your Facebook status!"[1]

There is no doubt that the advancing digitalization is having a major impact not only on our working lives but also on our private lives. The most obvious influence is probably that on our communication. For many people today, sending e-mails, presenting themselves on the Internet, communicating and receiving information of all kinds has become a normal part of their lives. But not everyone has access to the Internet in the same way; in this context, one speaks of a "digital divide," i.e. the division into so-called onliners and offliners.

There is much to be said for declaring access to the Internet as a human right today and even more so in the future. The basic principle of all human rights is immutable: no one may be existentially damaged in their self-respect. This is the core of human dignity, as it has found systematic expression, for example, in the ethics of Immanuel Kant or currently of Avishai Margalit (1996). However, the conditions of a humane society change with times and cultures. What constitutes a practice of exclusion and discrimination is not fixed once and for all but depends on cultural and economic conditions. Human rights apply not only in modern but also in traditional cultures, but state-guaranteed general education is a human right only in modern times because the conditions for it do not exist in traditional societies. Participation

[1] Comic by Black, Cuyler. https://i.pinimg.com/originals/c6/56/0b/c6560b07c4e4ce18e091087c-da384de8.jpg. Accessed 6 February 2018.

© The Author(s) 2022

J. Nida-Rümelin, N. Weidenfeld, *Digital Humanism*,
https://doi.org/10.1007/978-3-031-12482-2_15

in communication, freedom of expression, and freedom of information is a human right—the media of communication and information change with the times.

The question whether "the internet is a human right" must therefore be made more precise: Under what conditions does access to the Internet become an individual human right?

Since the codification of human rights is carried out by states, the establishment of a human right to internet access would establish a state duty to secure this access. The development of the World Wide Web has already reached such a stage for large regions of the world society that exclusion from internet communication—for example, due to a place of residence from which internet access cannot be established or due to economic conditions that exclude numerous people from participation due to a lack of financial means—is no longer compatible with freedom of information and freedom of expression. What leads to an exclusion that is inadmissible from a human rights perspective depends on the cultural development itself. As long as internet communication was only possible for small minorities of the world society, access to the Internet could not yet be a human right. However, the more important internet communication becomes in comparison to other media of communication and the larger the proportion of those who participate in it, the more clearly exclusion from internet communication means at the same time a loss of essential information and communication possibilities. The possibilities of obtaining information free of charge (apart from internet access itself) give the internet a special status compared to most other media. This makes exclusion from internet communication more serious. When a growing number of people have internet access, exclusion from internet access can become a human rights violation. That time does not seem to be far away.

A completely different question is whether or not the internet as a communication medium is conducive to the realization of human rights in political practice. In 2001, the study by the Carnegie Endowment for International Peace came to a negative conclusion, using Cuba and China as examples: according to this study, dictatorial regimes can use the Internet for their own purposes, and the possibilities of using it against such regimes are small. Ten years later, this will have to be assessed in a more differentiated way. At the latest after the Arab Spring, there is much to suggest that the possibilities to inform oneself via the Internet, to communicate, but also to associate, for example, to hold demonstrations, can be politically very effective. Even the Chinese government's smooth yet ruthless actions against dissidents and their potential for communication and association on the Internet has had a limited effect. The technical possibilities for circumventing internet blocking are

so numerous and the possibilities for control so weak that, taken as a whole, the Internet is more conducive than a hindrance to the realization of political and juridical practice that conforms to human rights.

Even though the Internet can trigger unexpected positive changes, its negative effects must not be neglected. First of all, this includes the fact that high use of digital media promotes certain usage skills (measurable by the processing speed per time unit), but at the same time, it also gives rise to an overload syndrome. This makes it understandable why members of the younger generation also temporarily or even permanently abandon parts of digital media and especially internet communication. For example, some people today consider it avant-garde to communicate exclusively via WhatsApp or Facebook Messenger and to close off all other digital channels. And in office communication, for example, it is recommended to limit e-mail communication to certain times of the day, to switch off the alert functions (automatic warnings, alarms, reminders) or to generally refer to a delayed response in the form of an automatic reply in order to gain time for concentration and protection of the nerves.

Another problem is that the private data of internet users almost inevitably end up with internet giants, who in turn pass it on to other companies, i.e., sell it. The simple recommendation to users that they should be more careful with their data does not take into account the current realities of internet communication. For large areas of the global society, non-participation in social media means de facto cultural exclusion, so that data-critical users pay for their right to informational self-determination with exclusion from social and cultural communities. The achievements of modern, liberal culture based on individual rights and opportunities for participation are being rolled back, and the separation of the public and private spheres that is so central to modern society as a prerequisite for a democratic order, as it was able to develop between the eighteenth and twentieth centuries, is thus being called into question.

An undeniable problem is also the cultural regression that affects not only individual adults with an unstable character but also increasingly young people and children. Selfie culture, with its typical gesture of holding up the mobile phone to take a picture of oneself, can be seen as a kind of reprise of the gesture of the mythological figure of Narcissus, who—obsessed with his own image—kept staying by the river starring in his own reflection until he died from unrequited love. Studies have shown that the brain reacts to each Facebook like with a shot of dopamine. Although psychologists argue about whether or not social media breeds narcissists, it is clear that narcissistic behavior is encouraged by media such as Instagram, Twitter, Snapchat, and

Facebook. As media scholar Roberto Simanowski (2018) writes, narcissistic selfies and other posts ultimately conceal a fear of one's own experience. Instead of being real in the world, we are content with an image that—as the literary and media philosopher Roland Barthes (1981) observers—is *noticed* but not really perceived.

The fact that young people who spend many hours of their day playing video games that are as realistic as possible can develop psychological problems has also been proven in many ways. Violent games pose a special problem. There is a striking correlation between school massacres and intensive spending time in virtual realities characterized by hate and violence.[2] These so-called first-person shooters, i.e., computer games in which the player acts from a first-person perspective in the game, have their origins in a desensitization program of the US army. In commercial first-person shooters, the player can fantasize himself as a being of a cruel omnipotence and thus lower the standards of the ethical criteria of consideration, compassion, and respect.

Even if, fortunately, these effects only show up in a small percentage of intensive gamers and it can be assumed that at-risk adolescents and young adults were already highly unstable before their immersion in virtual worlds, the probability that certain perpetrators of violence are also intensive gamers is high. In fact, there seems to be a connection between virtual representations of violence and the concrete manifestations of real violence, be it in the case of the two shooters who indiscriminately killed and injured people at the American Columbine High School in 1999 before killing themselves or also in the case of the German shooter from Erfurt, who first shot 16 people and then himself in 2002. But not only video games, but also the virtual reality of films glorifying violence, or films in which violence is stylized as an act of spiritual liberation, have led mentally unstable people to emulate this in the past. A much cited example is that of John Hinckley Jr. who attempted to assassinate President Reagan after seeing the film *Taxi Driver* (Martin Scorsese. USA, 1976).[3]

Another problem is the public displays of murder, manslaughter, and cruelty on "social" channels. The "game" of internet bullying, long considered harmless, is also part of this, driving many young people to despair and

[2] Violence plays a major role in many computer games—as it does in many US-American blockbusters. In the dramaturgy of the films, violence often has a ritual function and is presented as a kind of initiation rite. As Richard Slotkin has shown in *Regeneration through Violence* (1973), violence as a ritual act plays a central role in American narratives where violence is often associated with self-determination and reaching adulthood.

[3] There have been reports that a series of murders were linked to the film *Natural Born Killers* (1994) and the horror film *Scream* (1996).

sometimes suicide. Several US-American, but also European films have taken up this topic. One of these is the ABC family drama *Cyberbully* (Charles Binamé. USA, 2011) shot in Canada. The film which shows how young people can almost be driven to suicide by vile allegations on the Internet. In the film, all ends well: the attacked youths are saved at the last minute and rise together against their tormentors. The makers of the film had intended to contribute to the fight against cyberbullying—but did not succeed. A year after the film was released, the Canadian schoolgirl Amanda Todd killed herself after years of cyberbullying.

Internet pioneer Jaron Lanier, formerly one of the most ardent advocates and pioneers in the field of virtual reality, warns against such brutalization. In his book *10 Arguments For Deleting Your Social Media Accounts. Right Now,* he accuses social media of manipulating and ultimately getting users addicted.

In fact, the decline in the importance of traditional mass media in print, TV, and radio and the growing influence of algorithm-driven information channels is accompanied by an erosion of shared, inclusive political public spheres. In derailed election campaigns, such as that of the 2016/2017 US presidential campaign or the intra-Turkish dispute over constitutional reform in 2017, the actors disintegrate into groups with particular worldviews that are held together by beliefs but can no longer communicate with each other.

The understandable desire to exchange ideas with like-minded or similarly minded people, together with the preference for one's own interests controlled by algorithms, leads to a parceling of communication in the information offerings. Communication then takes place within more or less closed groups and communities, but no longer between members of different groups and communities. However, since democracy depends on comprehensive communication that includes the individual religious, ideological, ethnic, or whatever communities, these tendencies can certainly have a character that endangers democracy. It is to be hoped that the loss of importance of serious traditional media and their inclusive and selective and thus rationalizing function will be compensated by reliable and as universal as possible communication practices on the Internet of the future. Indeed, numerous serious discussion platforms give hope that civil forms of opinion exchange will increasingly gain influence within the framework of internet communication. Internet communication is not yet sufficiently inclusive to be able to speak of a world citizenship established via internet communication.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits any noncommercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if you modified the licensed material. You do not have permission under this license to share adapted material derived from this chapter or parts of it.

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

16

Digital Education

In *The Lawnmower Man* (Brett Leonard. USA, 1992), the mentally challenged Jobe is asked by Dr Angelo to sit in a chair in the high-tech science wing of a secret military building. For weeks, the researcher has been working with his human "guinea pig" Jobe, a gardener's assistant, whom he is trying to enhance mentally and physically. This time, Dr. Angelo wants to teach the illiterate Jobe directly at the military base. After Dr Angelo has connected Jobe's brain to a computer through a bunch of electrodes, he now downloads various disks on to Jobe's brain. Thanks to chemical stimulation, Jobe is able to memorize all the information within a few hours, as if his brain were nothing more than a large hard drive. Jobe gets into a real learning rush.

"I want more information!" he says and asks for more CDs. Dr Angelo tries to slow Jobe down. Learning does not only mean passively storing knowledge, one also need time to think about what one has learned in order to reflect on it.

"Man," Dr. Angelo tries to explain to Jobe, "may evolve a thousand-fold through this technology, but the rush must be tempered by wisdom."

But Jobe does not agree. He is convinced that this technology is just perfect and is all he needs: "This technology is simply a route to powers that conjurers and alchemists used centuries ago. The human race lost that knowledge and now I'm reclaiming it through virtual reality."

Interestingly, later on, Jobe will do everything to avoid others profiting from this technology, keeping the Internet and the virtual learning process just to himself.

Is the film to be seen as a warning to what happens if—in this case—only one person has access to the Internet and virtual reality and others do not?

© The Author(s) 2022
J. Nida-Rümelin, N. Weidenfeld, *Digital Humanism*,
https://doi.org/10.1007/978-3-031-12482-2_16

Already today over the world people and governments are complaining about a "digital divide", a term coined during the Clinton years. A lack of access to digital means and lack of digital competence creates not only a division between groups but also promotes economic inequality.

Many efforts have been made to bridge the gap of the digital divide. Therefore, "digital offensives" concentrate on installing the necessary infrastructure, providing hard- and software for youngsters. This is somewhat ironic as it is precisely this older generation, called the *digital immigrants* who are supposed to teach the younger generation, the *digital natives,* a language they have themselves only painstakingly learned, while the learners have mastered it playfully from childhood.

However, mastering the programming is more typical for the generation of *digital pioneers* of today's 50- to 70-year-olds. With the increasing perfection and complexity of commercial offers of digital products, such as website programs for example, the possibilities for amateurs to further develop offered software systems The user-friendliness of these products is so highly developed that the unwillingness of younger people to deal with programming itself is understandable. Here, a development that is familiar from much older technologies repeats itself: in the first decades of motorization, for example, it was quite desirable (and sometimes necessary for survival) to be familiar enough with the technology to be able to intervene even in the event of an engine or clutch failure. Today, even licensed car repair shops increasingly limit themselves to replacing parts—repairing a car is often no longer sensible and sometimes not even possible. Today's drivers no longer need to understand anything about the technology we entrust ourselves to. It would hardly help us in practice.

In the case of digital technologies, moreover, the change of products is so rapid that user knowledge, once acquired, must be permanently renewed and is therefore not very suitable as school material.

If Wilhelm von Humboldt, the great philosopher and theorist of education (and founder of the Humboldt University in Berlin), was right in his belief that school knowledge has a canonical character—in contrast to scientific, research-oriented knowledge—and that this is still true today, then practicing how to use digital products would not part of the meaningful school subject matter. So what could digital education be? Or rather: What goals (knowledge and competencies) should it be geared towards?

16.1 Lifeworld and Non-lifeworld Orientation Knowledge

By "orientation knowledge," we mean the knowledge required to be able to make coherent decisions in certain areas. The core of this orientation knowledge is grounded in the lifeworld. Thus, before and independently of school education, we learn early on to interpret a person's behavior, facial expressions, gestures, voice, etc. as an expression of their emotive attitude, their wishes, fears, hopes, and their empirical and moral beliefs. All everyday interaction between people is based on the reliability of what is called "folk psychology." This kind of everyday psychology is not rendered obsolete by psychology as a science; rather, psychology must be compatible with our lifeworld orientation knowledge. This is where psychology finds its validation.

Folk psychology is not devalued by the digitalization of part of our communication. On the contrary, correctly interpreting people's expressions and behaviors—even if they are digitally mediated—becomes a condition of successful interaction and communication (not only in the World Wide Web). A particular problem is that this ability to interpret behavioral expressions as expressions of intentions and beliefs is exploited to make pseudoagents appear. The use of numerous bots in election campaigns, for example, is an example of this. The more perfect the simulation of human behavior and emotional expressions, the more difficult it becomes to distinguish between digital simulation and human intentionality. The challenge of folk psychology is thus becoming greater, not less, as a result of digitalization. Digital technologies may facilitate psychological research in science, but they cannot replace empathy and sensitivity. This is also due to the fact that interpreting a person's behavior is not primarily a cognitive process, as psychological research has undoubtedly made clear recently.

It is a mistake to assume that orientation knowledge is limited to the realm of everyday life. Physical findings and models enable a more or less reliable prediction of weather conditions and have penetrated everyday knowledge via generally available apps. Psychological research enables therapeutic practices, and here too the transition to lifeworld practice is fluid. For example, the popular recommendation in counselling literature to make a list of positive qualities of one's partner in the event of a relationship crisis stems from scientific psychology. Numerous other connections between scientific research and lifeworld orientation knowledge could be cited.

16.2 Expertise

Digitization is changing the way data is made available and is being archived. Wide areas of the humanities, historical, and philological research (sources, texts, interpretations) are increasingly made generally available. Previously time-consuming research and travel, as well as hours spent in special libraries, are now in many cases superfluous. The complete digitization of museum collections, which is currently underway, but also the digital documentation in the sciences will further improve this situation. As the time and financial effort to acquire data decreases, this form of accumulated knowledge will be devalued. This means that knowledge of methods will become more important than knowledge of data. This is already reflected in study programs, in that entire areas of knowledge are considered dispensable and are replaced by methodological training.

As sensible as this shift in emphasis is in the higher education sector, it is also associated with a potential danger, namely, the loss of professional competence. There could even be a rude awakening from the digital dreams.

16.3 Canon as Common Background Knowledge

How can it be that not only methodological knowledge but also expert knowledge is still regarded as highly valuable in times of free availability of every conceivable data? To understand this, we need to consider the role that shared background knowledge plays in communication. For us to be able to communicate, it is not only necessary that we develop our arguments according to essentially a common logic but also that we can refer to common, undisputed beliefs (facts). In order to be able to judge what still needs to be clarified and by what means, if any, it should be clarified, one must already have a sound body of knowledge and experience. Anything else would be like stabbing in the dark. In fact, many internet searches have just this character. By chance, new and interesting aspects may come to one's attention, data one was not looking for at all may turn out to be relevant, literature may turn up whose existence one had not expected—but at the same time, many such searches get lost in the variety of offers on the internet if they are not guided by solid professional competence.

Quite independently of the digitization of human knowledge, there have been calls for the abandonment of the idea of canonical knowledge for other reasons of educational theory, especially in the USA and other multicultural societies. Against the idea of a canon of knowledge that focuses on education in general and on specific fields, the diversity of educational traditions and their different evaluations were put forward and the concern was expressed that any canonization marginalizes and devalues what does not belong to the canon. The US high school system in particular is correspondingly cautious with curricular requirements. But canonical knowledge has also been criticized from the perspective of emancipation through education as a cultural barrier to social advancement. From a gender perspective, it was also criticized that it was mainly old white (and dead) men who shaped and dominated the canon of knowledge.

As worthy of consideration as these objections are: If the conclusion is drawn that canonical knowledge can be dispensed with, this is a mistake in educational theory. Without common background knowledge, without shared beliefs that do not need any further justification and have passed the reality test, even dissent and cultural differences cannot be discussed. One can characterize the role of the canon in analogy to the role of lifeworld knowledge in our everyday practice. Just as we draw on the shared attributions of emotive and cognitive attitudes (feelings and beliefs) in our everyday practice, understanding in certain disciplines or professional practices requires shared, unquestioned professional knowledge and competencies. Even if the selection of the canon is inevitably arbitrary in any given case, it is nevertheless indispensable in the practice of understanding and interaction. Professionalism does not become obsolete through digitalization.

16.4 Judgment

In well-stocked university libraries, a lot of work goes into sorting the collections in a way that promotes access to what is relevant in each case. The arrangement of library collections is itself the result of a separate academic competence, library science. The organization of scientific publications via book series and journals also follows the pattern of strict selection and sorting by professional competence. This pre-structuring is lost in the World Wide Web. The links essentially follow the statistical distribution of user behavior and thus reproduce associative concatenations whose systematic significance is often enough extremely low. The search engines, which do not disclose their algorithms, further obscure the situation by taking commercial interests into

account. The recent abandonment of the net neutrality requirement[1] by the US government will further strengthen the influence of commercial interests on the structuring of data.

The digitization of data provision means that numerous "gatekeepers," such as librarians, publishing editors, journal reviewers, newspaper, TV or radio editors, are no longer needed. This means that independent judgment is increasingly required. Data provision does not replace the ability to assess data and check whether it is reliable and what arguments can be based on it.

The World Wide Web confronts us with a far greater variety of interpretations, theses, theories, and ideologies. Forming an opinion is therefore becoming more demanding. The old humanistic ideal of education, as formulated in Plato's *Theaetetus dialogue* 2500 years ago, thus gains new relevance. People who tend to follow suggestively formulated beliefs or shield themselves from inconvenient facts will quickly lose their bearings in the new digital data universe. They lock themselves into the "bubbles" that social media, in particular, provide, or stagger back and forth through the data world driven by different influences.

As a result of digitalization, we do not live in a knowledge society, but at best in a data society, or better: in a data economy. The availability of data characterizing individuals' purchasing preferences and behaviors, combined with access to it via social media or other communication channels, has become a successful business model that finances internet giants, whose dominant position in the market has de facto made them an essential part of the infrastructure of communication data, flows of services and goods. This Big Data economy, whose best time will possibly only come with the expansion of highly automated individual transport, is not a knowledge society because knowledge consists of justified and true beliefs. Knowledge requires power of judgment. Only the evaluation, classification, and interpretation of data can constitute knowledge. The great challenge of education in times of digitalization is to transform the current trend toward a data economy into a development toward a knowledge society.

[1] Net neutrality is the requirement to treat all data on the Internet equally and to ensure access to the Internet in which people are not discriminated against.

16.5 Personality Development

Has the central goal of humanism, namely, the formation of personality, become obsolete in times of digitalization? The answer must undoubtedly be: No, on the contrary. The development of personality is more relevant today than ever before, and its importance will continue to increase due to the digitalization of our communications and interactions, transfers of data and services and their production (keyword: Industry 4.0).

The reason for this is obvious: the more diverse, volatile, and complex personal ties, community formations and ways of life become, the greater the demands on the individual ability to be the author of one's own decisions, beliefs, and projects. The digital possibilities create new freedom, trigger an enormous dynamic of change not only in economic but also in cultural conditions, and therefore also strengthen the autonomy potential of individuals and at the same time put them under the permanent stress of a growing need for orientation.

In the digital lifeworld of the future, personal strength is required more than ever before in human history. The education system must adapt to this. The imparting of knowledge and skills must serve the highest goal, namely, the strengthening of the adolescents' personality. The focus must not be on the passive absorption of prefabricated material but on the active mastery of complex judgments and decision-making structures. However, the current trend toward standardization, acceleration of teaching, and schooling in the tertiary education sector is going in the opposite direction: the time for reflection is becoming scarce, the abundance of material enforces passive reception, while the social and ethical competences, as well as the artistic and creative, the manual and technical are diminishing. The unity of the person, the respect for the human individual with its different facets, talents, interests, and abilities does not receive the necessary attention. Ideally, the child, the adolescent, the young adult finds themselves on the educational path. Not by absorbing as much knowledge as possible on their own, like the lawnmower man Jobe, but by making use of their freedom on the basis of their own decisions and allowing their personality to mature in often painful processes of trial and error. Digital technologies can support this process of personality development in the classroom. If used competently, they are quite suitable for

promoting creative abilities and creating knowledge connections that are excluded in parceled school lessons.[2]

After his digital learning crash course, Jobe's head is now filled to the brim with data and information. However, he does not feel very good about it. First, he suffers from headaches, then from delusions: He can distinguish less and less between the real and the virtual world and mutates into a despotic tyrant. He has no reliable knowledge of orientation. He lacks empathy and sensitivity and is far from having a mature personality. The only truly sound expertise he possesses is the knowledge of how to mow lawns—that is, the knowledge from the job he held before his digital brainwashing. Also, all the orientational knowledge he has is that of mowing lawns. He thus acts accordingly—like a brutal lawnmower. He wants to mow down the world and the people, so to speak, in order to keep them uniform, so he can reign over the world like he reigned over a lawn.

The information that Jobe has accumulated has not made him a better, but worse. He is a person, who has lost his moral compass. Had he received his information in moderation and with enough time for reflection, things might have been different. An uncontrolled digitalized education that only focuses on the blind accumulation of information, on quantity instead of quality and does not pay attention to promoting orientation knowledge, specialized knowledge as well as power of judgment and personality formation, is bound to fail miserably.

[2] For example, the "Atlas" software developed by the Parmenides Foundation (https://www.parmenides-foundation.org) as a learning platform facilitates the comprehension of knowledge spaces beyond the parcelling of subjects and methods practiced by academic institutions by making logical connections and categories accessible for analysis with the help of taxonomic graphs.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits any noncommercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if you modified the licensed material. You do not have permission under this license to share adapted material derived from this chapter or parts of it.

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

17

Utopia of Liquid Democracy

"What if the democracy we believe we serve no longer exists? And the Republic has become the very evil we have been fighting to destroy?" That's what the beautiful Padmé, former queen and now senator in the galactic senate, says to young Anakin Skywalker in *Star Wars: Revenge of the Sith* (George Lucas. USA, 2005). Indeed, the political situation has become quite confusing in this film. War is raging all over space, Jedi knights are fighting clone warriors, and the Chancellor of the Galactic Republic wants nothing more than to break up that republic so that he can install himself as sole ruler. Toward the end of the film, the Chancellor proclaims the end of the Republic before the great assembly in the House of Representatives. In a world that has become confusing, he argues, only a fascist system can ensure order. So he immediately proclaims the "first Galactic Empire." All the democratically elected representatives applaud enthusiastically. "So this is how liberty dies, with thunderous applause," Padmé comments sarcastically.

The theme of a destroyed or dysfunctional democracy features in many sci-fi films, reflecting the secret fear many citizens have of the failure of our democratic system. These fears have particularly increased in our time. One idea to solve this problem is the so-called *liquid democracy*.

In a short, simply animated info video[1] on YouTube, the idea is explained nicely in less than 4 min. "Direct democracy involves every citizen voting on every issue. Indirect democracy has designated representatives whose jobs it is

[1] Online at: https://www.youtube.com/watch?v=QORq_qBHo2w. Accessed on 5 March 2022.

© The Author(s) 2022
J. Nida-Rümelin, N. Weidenfeld, *Digital Humanism*,
https://doi.org/10.1007/978-3-031-12482-2_17

to be aware of current events and use this information to make a well-informed decision on our behalf. [...] There are issues with both of these models. In a direct democracy, all citizens are not informed on all the issues and many citizens do not have time to debate and make these tough decisions on a day-to-day basis. Indirect democracies are accused of creating a disconnect between the citizens and the policies that they must abide by. [...] Liquid democracy is truly a blend of the two. You can choose to vote as an informed citizen or you can choose to delegate your vote. It is this fluid alternation that gives its name to liquid democracy. But is this even possible? Yes, it is. With revolutionary technologies. [....] This is a choice that every citizen should have. Don't you agree?"

Indeed, there is widespread unease about the perceived insufficient participation of the citizens. Left-wing and right-wing populist movements have mobilized the masses with the idea that aloof elites rule the country while the opinions of ordinary citizens no longer carry any weight—sometimes quite successfully, as in the Brexit campaign or Donald Trump's presidential campaign.

This discomfort has a long history and an important witness, namely the Enlightenment philosopher and pioneer of the French Revolution Jean-Jacques Rousseau. For Rousseau, it is of central importance that citizens consult together in assemblies in order to (ideally) cast a unanimous vote at the end. He calls the jointly worked out goals the *volonté générale*, the general will. Only in this way, Rousseau believes, can the original freedom of the individual be restored. Rousseau imagined this original freedom as that of a man of nature who lives his life on his own, without being oppressed by others or having to take others into consideration. In modern civilization, however, we are dependent on cooperation and community, and therefore the original freedom to determine for oneself the rules by which one lives can only be restored by unanimously adopting, in the community of citoyens, the citizens of a republic, rules that apply equally to all. Any association of private interests in the form of interest groups, lobbying or parties would, according to Rousseau, be pernicious because it would threaten the moral character of the republic, because the willingness to ascertain and follow the common will requires that one distance oneself from one's private interests and see them merely as one of many determinants of the common interest.

The objections to this republican understanding of democracy, the Rousseauian utopia of the *volonté générale*, are obvious. How are millions of citizens to come together at an assembly? How can groups with common interests be prevented from joining together and forming factions, parties, lobby groups? Above all, how can it be ensured that every voice, every opinion is given equal consideration in determining the common will? In mass

democracy, how can it be achieved that complex issues to which legislation has to respond can be transmitted to all those involved in the opinion-forming process? Under the traditional conditions of democracy, as they existed until recently, the Rousseauian idea of the republic, or more generally: the idea of direct, immediate democracy, of a political opinion-forming process that includes everyone, is not feasible.

The potential of digital communication, and especially internet communication, as well as the use of complex software to control opinion-forming and decision-making, such as liquid democracy, however have made these objections to a republican form of democracy obsolete.[2] At first glance, digital platforms seem to be an excellent way to elicit such a common will and to strengthen democracy by expanding opportunities for participation. It is therefore understandable that the old republican ideals are currently experiencing a renaissance and are putting pressure on the cumbersome, multi-level, institutionalized procedures of political decision-making in parliamentary, representative, constitutional democracy. From the left, this pressure is being exerted in the form of initiatives, petition platforms, social media groups, i.e. in the form of a new civic engagement, from the right in the form of the emotionalization of political opinion-forming and the devaluation of scientific expertise. These characterizations, however, oversimplify the situation. There is also populism on the left of the political spectrum that takes no notice of scientific arguments and relies on emotionalization. And of course, there are also defenders of scientific rationality on the right of the political spectrum.

There is much to be said for using the new technological possibilities of digitalization and the internet to make democratic opinion-forming more inclusive and substantive.[3] More inclusive by involving all sections of the population, and more substantial, in that the easy availability of information through digitalization is used.[4] Unfortunately, the utopia of the digital republic[5] will never be realized in this way. Three theorems show why.

---

[2] The contemporary Rousseauist Benjamin Barber, among others, takes the opposite position.

[3] The French sociologist Geoffroy de Lagasnerie goes one step further and wants to position internet communication against statehood in general, up to the not only utopian but also dangerous anarchist vision of the replacement of statehood by free associations of groups that constitute themselves via the internet Cf. de Lagasnerie (2017).

[4] For example, there is the software Adhocracy, which sees itself as a "cooperative discourse-text processing-delegation and voting tool" and enables joint decision-making with many participants. The processes are transparent and comprehensible and do not require moderation. The open-source software LiquidFeedback enables both direct and representative democratic processes.

[5] In the following, "digital republic" is understood to mean a direct democracy in which all citizens participate in opinion-forming and decision-making on the basis of digital information and decision-making technologies.

First, Condorcet's "problem of cyclical majorities" (also known as the "Condorcet paradox"), then Kenneth Arrow's "impossibility theorem," and finally Allan Gibbard and Mark Satterthwaite's "manipulation or strategic voting theorem." Although all these theorems were proven some time ago (Arrow 1951; Gibbard/Satterthwaite 1973; Condorcet as early as 1785), they have not penetrated the general consciousness beyond specialist circles. This even applies to the relevant discipline, political science. This is probably mainly due to the fact that the theoretical consequences of these theorems for political practice have not been sufficiently clarified.

Condorcet's paradox can be described as follows: A, B, and C stand for three factions in parliament. None of them can form a majority on their own. Only two factions together can form a majority. A vote takes place that obeys the simple, binary majority rule, i.e., an alternative is elected if and only if it gathers more votes than other alternatives. Let us assume that three alternatives X, Y, Z are to be voted on (e.g., candidates for the chancellorship or bills). If one now votes in alphabetical order, i.e., first X runs against Y, X wins in a coalition of A and B. Then there remains the vote between X and Z, this time Z wins in a coalition of B and C. Here, normally, the voting would be over. However, a test vote shows that Y would have prevailed over Z. This violates the so-called transitivity condition. This condition requires that whenever an alternative X is preferred to Y and at the same time an alternative Y is preferred to Z, X is also preferred to Z. The order of voting alone is decisive here. This means that an assembly leader can cause the alternative he prefers to be voted for simply by choosing a certain voting order. This phenomenon is called "susceptibility to manipulation" (please see Table 17.1).

Now, one might assume that such cyclical voting results are extremely rare. But in fact, the probability of cyclical voting results increases with the number of people involved in the decision and the number of alternatives. Quite a blow to the ideal of the democratic voting process! But it gets worse. Around 170 years after Condorcet, the US economist Kenneth Arrow proves that it is not possible to fulfil four indispensable conditions of collective and democratic rationality in a process of collective decision-making.

Table 17.1 The Condorcet paradox

A	B	C
X	Z	Y
Y	X	Z
Z	Y	X

The first condition he calls "D" for *non-dictatorship*. This means that there must not be a single person who determines the outcome of collective decisions regardless of the preferences of others.

The second condition is what Arrow calls "P" for *Pareto efficiency*. This postulate, which goes back to the Italian scientist Vilfredo Pareto, requires that common preferences of all are realized, which in application to collective decisions means that the unanimity principle applies: if all those involved in the decision prefer X over Y, then this should be reflected in the outcome of the vote. For example, it cannot be the case that in a committee where all persons consider a candidate X to be better than a candidate Y, in the end, Y wins the vote.

The third condition is what Arrow calls "I" for *irrelevance*. It requires that the collective preference for, say, X over Y does not change simply because an alternative Z is added. Of course, the added alternative may be better than X, but the ranking between X and Y should not change by the mere addition of another alternative. If I prefer to go on holiday to Italy instead of France, then this preference should not change simply because a holiday in the USA has become possible as an additional alternative. Why should I prefer a holiday in France to a holiday in Italy simply because I could now also spend a holiday in the USA?

Arrow calls the last condition "U" for *unrestricted domain*. The better term is "preference sovereignty": all participants in the vote can feed in their preferences, there are in a sense no prohibitions or restrictions on having certain preferences.

It seems obvious that these four conditions are far too weak to characterize democratic decision-making procedures. One would want, for example, some form of assurance that majorities matter, that anonymity and neutrality are assured, perhaps minority protection and individual rights. The shocking thing, however, is that Arrow has shown that these four minimum conditions for collective decision-making procedures cannot be fulfilled simultaneously, i.e. that there is no rule of collective decision-making that fulfils these four conditions at the same time.

Now, one could draw the general conclusion that democracy is impossible. However, that would be premature. A closer look at parliamentary, representative, constitutional democracy shows that its procedures of political decision-making are designed in such a way that they usually circumvent the problems shown in Arrow's theorem. For example, cyclical voting results, as in the Condorcet paradox presented above, can only occur if there are at least three alternatives. Fortunately, the practice in parliamentary democracy of basing the respective government on a parliamentary majority formed by one or

more parliamentary groups bound together by a coalition agreement precludes the possibility of several proposals, each of which has a chance of being endorsed by a majority.

Finally, the findings of Gibbard and Satterthwaite should be mentioned. They were able to show that there is no single process of collective decision-making that is not susceptible to strategic voting and manipulation.

A decision-making process is susceptible to strategic voting if at least one person involved is more likely to realize their preferences if they feed other preferences into the decision-making process than they actually have.

The devastating result of Gibbard and Satterthwaite's theorem is that all processes of collective decision-making are susceptible to both strategy and manipulation. Fortunately, there is a safeguard here too in parliamentary, representative, constitutional democracies, and this lies in the role played by the publicly presented argument and the transparency of politicians' decision-making behavior.

If a politician puts forward certain arguments in favor of a project, then it would at least require justification if she votes against it in the end. One could also say that the obligation to give reasons and the public formation of political opinion limit the scope for strategic and manipulative behavior in democracy. In this sense, the institutional order of parliamentary democracy can be interpreted as an attempt to circumvent the paradoxes and dilemmas of collective rationality and make political decisions possible. If one were to move to a digital republic, this feature of a "deliberative democracy" would disappear. The control of who voted when, how, and with which arguments in favor of which project would be impossible given the sheer number of participants, and political responsibility would diffuse in an amorphous mass of thousands and thousands of participants who merely vote "yes" or "no" by mouse click.

So, as beautiful and simple as Liquid Democracy presents itself in the YouTube video, it must fail in this form.

In fact, large-scale attempts at liquid democracy have so far proved unfeasible—mostly due to lack of participation.[6] To read this merely as an expression of saturation, disinterest, or convenience would be wrong. It is more likely that the resistance to the transition to a digital republic is fed by the deeper insight that it inevitably entails a loss of rationality and that—as we have seen—in the worst-case collective self-blockades in the form of cyclical preferences lead to serious chaos and instability.

[6] In Germany, Liquid Democracy was introduced in the district of Friesland in 2012 and discontinued just 4 years later due to lack of use.

And yet, objecting to the idea of a digital republic, demonstrating that it is not feasible for systematic reasons and that moving toward it is not desirable, is not incompatible with advocating massive use of the new technological possibilities. There is no contradiction. The goal of digital humanism is to strengthen power of judgment and decision-making and thus individual and collective autonomy. To achieve this, digital information and decision-making technologies are to be used as a supplement to parliamentary, representative democracies based on the rule of law—but they are merely a support, not a substitute.

In this sense, the enrichment of public space through the involvement of as many citizens as possible would not replace representative liberal democracy but strengthen it. The opportunities for this are more favorable today than ever before.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits any noncommercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if you modified the licensed material. You do not have permission under this license to share adapted material derived from this chapter or parts of it.

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

18

Socio-economic Aspects of Digitalization

When Detective Spooner first confronts the CEO of U.S. Robotics, the largest robot manufacturer in the world of *I, Robot* (Alex Proyas. USA, 2004), he can't help but make a sarcastic comment: "I got an idea for one of your commercials. You can see a carpenter making a beautiful chair, and then one of the robots comes in and makes a better chair, twice as fast. And then you superimpose on the screen: 'USR—shittin' on the little guy.' That would be the fade-out."

Currently, there is much speculation about the changes an industry 4.0 would bring, in particular, the production of industrial goods where they are needed, on a data basis provided by internet communication. The basis of industry 4.0 should enable largely self-organizing production. Through the use of highly complex search engines[1] and the interconnectedness of people, machines, logistics and production facilities complex value chains are to be made more efficient and thus—after mechanization through steam engines, mass production through assembly lines and the introduction of computers—a "fourth industrial revolution" is to be initiated. Whether this will actually happen, however, is written in the stars. Disruptive technological innovations of the past have not yet been initiated by political programs and goals, but digitalization may soon enter a new phase that will expand the previous, rather ephemeral character of this technology (its focus on communication and interpretation, but above all data usage for advertising purposes) to

[1] The most complex search engine is the software system Watson produced by IBM, which—at least that is what IBM hopes—will one day be used on a large scale as a word and image recognition program.

© The Author(s) 2022

J. Nida-Rümelin, N. Weidenfeld, *Digital Humanism*,
https://doi.org/10.1007/978-3-031-12482-2_18

include the tangible dimension of digital, network-controlled manufacturing.

Before this concept has even taken on a concrete shape, its protagonists are already talking about the massive job losses that would accompany it. The World Economic Forum in Davos warned in 2016 that 7 million jobs would soon be lost in Europe, and some forecasts speak of the loss of 50% or more of all work processes, not only in manufacturing but also in the service sector. Some politicians, economists, and managers argue for a robot tax[2] to replace the loss of wages, while others advocate the introduction of an unconditional basic income in order to stem the social disruptions in time.

Such a fundamental restructuring of the welfare state—away from being tied to gainful employment and wage tax, toward basic security models financed by sales tax—is gaining more and more supporters given the expected further increase in the digitalization of the entire economy. Politically, this ranges from strong forces from employers and the associated economic institutes and business associations to left-anarchist supporters. The line of conflict "pro and contra unconditional basic income" does not follow a left-right pattern. Common to all proponents is the goal of a radical change of system: the various social security systems are to be replaced by a simple and (supposedly) fair model. For many, one of the attractive elements of the introduction of an unconditional basic income is that it would make welfare state institutions, bureaucracy, and administration largely superfluous.

At first glance basic income seems to be an appealing idea: while all the tedious work is taken over by robots that clean, cook, and make products for us, we can devote ourselves entirely to our interests and indulge in our pleasures. The sweet life would no longer be possible only for the rich, the digital land of milk and honey would be achieved.

In his 1567 painting, Pieter Bruegel the Elder presents us with his vision of the land of milk and honey as it might have looked before the first industrial revolution. People lie there relaxed on the ground, there is enough to eat and drink, class distinctions are abolished. A knight lies next to a peasant, who in turn lies next to a scholar. And yet this vision is appealing only at first sight. Taking a closer look, one realizes that the land of milk and honey is in fact a dystopian paradise that makes people fat and lazy. The message is quite clear: there is more to a good life than being fed.

[2] In 2016, Luxembourg MEP Mady Delveau-Stehres first brought the idea of a robot tax to the EU Parliament in the form of a draft report. While Bill Gates, for example, supported the idea, it was mostly negatively received at the time.

A modern version of the land of milk and honey is provided by the film *WALL*-E. After the earth has become uninhabitable due to environmental catastrophes, people have evacuated earth and live on a spaceship that has been flying through space for decades. The people on board lack nothing, robots do all the work: they wake them up, feed them, drive them around, and steer the ship. *WALL*-E is a disturbing vision of a software operated paradise. On this fully automated ship of eternal holidays, people have become dull and apathic. Fat and unable to move, they either sit in automated cars or lie under sunshades on artificial beaches with an artificial sun. Their bloated faces stare at screens where they are being entertained by the on-board program.

The roots of the idea of an unconditional basic income go back to the nineteenth century. The most important ideological context is that of anarchism and utopian socialism. In recent decades, the ideology of the end-of-the-labor society has been added as a further ideological, sociological context. This ideology states that in view of rationalization processes, gainful employment is generally becoming a scarce commodity. According to this view, sources of income other than gainful employment are increasingly necessary. This idea, which has had numerous proponents from André Gorz to Jeremy Rifkin to Ulrich Beck and is based on visions by Herbert Marcuse (Marcuse 1964; Gorz 1999; Rifkin 1995), has until now proven to be false. So far, technological innovations have triggered massive shifts in the labor markets but have not led to a comprehensive dwindling of gainful employment.

Conceptions of unconditional basic income go hand in hand with the vision of an era of freedom: everyone can decide at any time whether they want to pursue gainful employment, take up other work (civic engagement, family work, etc.), or devote themselves to leisure. The unconditional basic income would for the first time remove the compulsion to work not only for a few but for everyone.

The first question that arises is whether digitalization has brought about a new situation that will give a late triumph to the old and hitherto false thesis of the disappearance of gainful employment in industrial society. The long-term analyses of productivity increases in the US economy through technological innovations speak against this expectation. Studies that forecast a high loss of gainful employment are therefore either just as wrong as those made decades ago for the car industry (in fact, more people work in the car industry today than in the past, despite largely deserted production halls). Also, it is to be expected that these losses will be compensated for by new labor capacities that will be created by digitalization.

We know from economics that a massive loss of jobs with moderately growing GDP could only occur on balance if the growth in productivity were

significantly higher than the growth in GDP over many years. In fact, however, productivity growth has weakened rather than increased in the past years of increased digitalization; indeed, disappointingly, the third and fourth technological revolutions have not yet been reflected in productivity growth. As long as this remains the case, there can be no net job losses due to digitalization, at least not until there is a massive collapse in overall economic output due to digitalization—which no one seriously expects.

If, on the other hand, there should ever be a huge productivity boost from Industry 4.0 which would not compensated for by high growth rates, the volume of labor would shrink accordingly. For the sake of simplicity, let's assume that productivity doubles within a decade, while the gross national product only increases by 50%. In this case, the volume of labor would decrease by a quarter with the same working time per capita, i.e., additional unemployment of 25% would result. A 25% decrease in working time per capita would even be compatible (without rising unemployment) with a 25% increase in real income over 10 years.

According to empirical data so far, the productivity effects from digitalization are modest and not comparable with earlier technological revolutions. They were evident in the early phase of the introduction of PCs and browsers (1994–2004) but fell to a conspicuously low level after the turn of the century, which continues to this day. It is by no means impossible that this will change in the so-called third phase of digitalization.[3] Certainly, the focus of digitalization led by US internet giants on communication, entertainment, and consumption in the second phase of digitalization is partly responsible for this disappointing development. This could change however if there was to be new phase of industrially oriented digitalization. If this were to happen and the growth rates, at least in the mature economies, were not sufficient to prevent a significant shrinkage of the volume of work, then the new leeway should be used for an expansion of lifelong learning, a more flexible change between family and gainful employment, for sabbaticals and more time sovereignty, and not lead to a division of society into highly paid and productive gainfully employed workers and unproductive workers who are supported by an unconditional basic income.

Even a modest unconditional basic income in the amount of the minimum wage of a full job would lead to very high taxes. The amount of taxes depends,

[3] The first wave of digitalization (1985–1999) is understood as the development of the Internet and the creation of a new infrastructure. The second wave (2000–2015) refers to the time when Facebook, Google, various apps, etc. become an integral part of many people's lives. The third wave of digitalization, i.e., the time we are now at the beginning of, refers to a time of the "Internet of Things," i.e., a ubiquitous and comprehensive digital interconnectedness of devices, sensors, and products.

of course, on the extent to which use is made of the unconditional basic income, i.e., how large the proportion is of those who then voluntarily leave the labor force, temporarily or permanently, partially or completely. Since an unconditional basic income in the amount of the average earned income is unrealistic for tax reasons alone, an existence on the basis of the unconditional basic income would only be attractive for parts of the population, including in particular younger people (in the phase after completing their education and before entering gainful employment) and those who can secure their standard of living above the unconditional basic income, for example, through intra-family transfer payments from earned income.

The empirical evidence that prolonged absence from employment drastically reduces earning capacity is overwhelming. This means that graduates must be integrated into the labor force quickly after completing their training or studies, otherwise their qualifications lose value. The long-term unemployed are difficult to integrate into working life even if they have good qualifications. Incentives to stay out of the labor force for many years are therefore irresponsible. They lead to a division of society into the permanently employed and the permanently unemployed, even if they are provided with an unconditional basic income.

The introduction of an unconditional basic income would deepen the already existing cultural division of society into the occupationally integrated and the occupationally non-integrated, whether through precarious and frequently changing employment or through unemployment. The introduction of an unconditional basic income would be tantamount to capitulation. Instead of a strategy of integration and inclusion in the working society, the final and then soon irreversible withdrawal from the working society would be rewarded. The bold justification of the thesis "Why Surfers Should Be Fed" (Van Parijs 1991) by the most important theorist of an unconditional basic income, the Belgian economist and philosopher Philippe Van Parijs, is to be taken quite seriously. Whereas up to now, such a form of existence has only been an option for most people during transitional phases of their biography, it would become a widespread phenomenon in the future, the quantitative dimension of which would depend on the amount of the unconditional basic income. The already observed prolongation of adolescence in the sense of dependent forms of existence in Western societies would receive a further boost, and entry into working life would not be accelerated as at present, but delayed, for many probably to the point where they voluntarily stay away from work.

It is likely that the apocalyptic vision that has accompanied all disruptive technological innovations is also unfounded in our times of digitalization. All

comparable upheavals have led to job losses in certain industries and professions; just think of the rural exodus in the nineteenth century, triggered by the use of machines in agricultural production, which at the same time enabled a huge increase in jobs elsewhere. In the case of digitalization, there is much to suggest that the economies of the future will use the resources freed up by rationalization to satisfy cultural interests, for example. Already today, for example, the production of digital games and virtual realities is one of the fastest growing industries. The cultural sector as a whole is a growth driver in all Western economies. Increasing digitalization need not and must not result in an exit from the working society. Such a development is economically undocumented and culturally undesirable.

At the end of WALL-E, the spaceship with the last remaining humans on board heads for Earth. As life has become possible again on Earth, people can now return to their old planet after centuries in barren space. When the hatches open, the fat humans, who have forgotten how to walk in the meantime, crawl outside. Wide-eyed, they stare at the huge skyscraper skeletons lying destroyed in front of them. The air is dusty and dry. They know there is a lot of work ahead of them to make the Earth beautiful and habitable again. And yet the captain smiles. The supposed land of milk and honey has come to an end and life has meaning again.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits any noncommercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if you modified the licensed material. You do not have permission under this license to share adapted material derived from this chapter or parts of it.

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

19

Transhumanist Temptations

When the policeman Alex wakes up from his artificial coma and finds himself in his new body, a kind of black steel Batman suit, he—understandably—panics.

"What the hell have you done with me?" he asks, horrified.

Alex, who fell victim to a car bomb and lost most parts of his body apart from his torso, was operated into a robotic suit by Dr Norton, an expert in cybernetic prosthetics. While this suit gives Alex mobility and strength, it also turns him into a kind of robot on a psychological level. For the technical parts, such as artificial arms and hands, to function well, the electrical impulses from the brain must be perfectly implemented, i.e., Alex must be emotionally stabilized.

Alex is unaware that the mega-corporation OmniCorp, which sponsored his expensive operation, has an ulterior motive: Since the idea of armed robots has been so far met with resistance from the population as well as from the government due to a common law enacted by Senator Dreyfus,[1] OmniCorp CEO Raymond Sellars hopes to present the public with Alex a cyborg so likeable and efficient that people will want more of these RoboCops in the future. Of course, in order to then be able to distribute them with the greatest possible profit.

When Alex sees himself in his robot suit for the first time, he breaks down. He doesn't want to live like this. Only when Dr Norton tells him that his wife

[1] Named after the real-life philosopher Hubert L. Dreyfus, who was one of the most prominent critics of AI in the USA. See Dreyfus (1972, 1992).

© The Author(s) 2022

J. Nida-Rümelin, N. Weidenfeld, *Digital Humanism*,
https://doi.org/10.1007/978-3-031-12482-2_19

has consented to his transformation he calms down and decides to get used to his new existence.

There is another problem, however: Alex is far less capable of combat than real robots. Robots shoot straight away; Alex however considers beforehand whether it is morally okay to do so. In doing so, however, he loses time. To overcome this—from OmniCorp's perspective—disadvantage, Dr Norton equips Alex virtual vizor. In combat situations, this vizor automatically comes down, making the robotic part of him take control of his body and his actions. The perfidious thing about this is that a chip implanted in Alex's brain by Dr Norton causes him to believe that these are still his own decisions. Alex becomes increasingly cold and emotionless as the film progresses. Neither his wife nor his son recognize him. Where the robot begins—*RoboCop* (José Padilha. USA, 2014) tells us—the human ends. Another problem raised by the film is the question of responsibility. In *RoboCop*, the power over the robots does not lie with the state but is in the hands of a company. And this company has only one goal, namely to increase its profits. Alex is defenseless against the machinations of OmniCorp, which decides how he or to be precise his robotic parts are to be programmed.

Cyborgs—hybrid beings between man and machine—have existed in a certain sense for a long time. Aren't glasses also an artificial aid that humans have been using for centuries to improve their abilities? Prosthetic legs or hearing aids also belong to this category. But what about implanted microchips that allow you to start a car, open your front door, or pass on your contact details? Or augmented reality lenses which allow one to see more of one's surroundings or to superimpose information from the internet directly into one's field of vision? Or implanted chips that allow one to sense the movements of others or to hear colors as sounds? Will we become as accustomed to these aids as we are to glasses for long-sightedness? On a website called *I am Robot*,[2] the company of the same name (under the slogan "Upgrade your body with new features") is already offering to send these chips by post.

The US Defense Advanced Research Projects Agency, an agency of the US Department of Defense, has been researching for years the extent to which electric shocks to certain brain regions can increase attention, suppress fatigue, and influence moral behavior. Neuroenhancement is the name of this technology, which is designed to improve the brain, partly with the help of digital chips, or to equip it with better combat capabilities.[3]

---

[2] Online at: https://chip-implants.com/. Accessed 26 February 2022.

[3] For example, the well-known entrepreneur and Tesla CEO Elon Musk is currently developing chips with his start-up Neuralink, which are supposed to improve the brain's performance and enable thought transmission.

The fantasies are particularly inflamed by what is referred to as *brain-computer interface*, i.e., technical connections between brain processes and software systems. For example, the European Union is also funding the VERE project (Virtual Embodiment and Robotic Re-Embodiment), whose explicit research goal is to permanently link the human sense of self to avatars or robots. There have already been successful trials in which a subject's ideas about movement are read via magnetic resonance imaging and passed on to robots, which then execute this movement.[4] The positive sides of such technology, including the accompanying Proteus effect, i.e., the successful identification of the ego with another artificial body, are undeniable: physically paralyzed people could walk and move in the world again thanks to a new body or new body parts. But what about the negative sides? What if, thanks to clever neurological manipulations, this technique is misused to make people do certain things? What if impulse control works less well with such a technique? And what if, within a military deployment, the identification with a strong robotic soldier body acting together with others in a group leads to what is known as the "Lucifer effect,"[5] i.e., to the user being tempted to act excessively aggressively and sadistically because he finds himself in a new role?

Advocates of these new technologies made to expand human capabilities call themselves transhumanists. They endorse using the most modern technological possibilities to arrive at a completely new dimension of human cognition (especially regarding the ability to grasp complex processes) and practice. Transhumanism is a global movement. There are several global transhumanist think tanks as well as transhumanist "parties" in the USA, Australia, Korea, India, Great Britain, Austria, and Germany.

As euphoric as transhumanists are about the expansion of human capabilities through new technologies, there are also counter-movements that fear a new division of society between those who, with the help of new technologies (not only digital but also medical and pharmaceutical or nanotechnological), place themselves on a higher level of human development and those who are denied this due to a lack of economic or technical means.

Contemporary transhumanism is the concretization of an age-old human dream, namely, to be able to defy the human condition, to go beyond all limitations of human nature, and to develop superhuman powers and abilities.

---

[4] See the Israeli-French pilot study by Cohen et al. (2012).

[5] The "Lucifer effect" comes from the "Stanford Prison Experiment" conducted by Philip Zimbardo and colleagues in 1971, in which students were asked to take on the roles of guards and prisoners. After a short time, students who had taken on the role of guards treated the prisoners sadistically and aggressively. The reason for this was both the fact that the students were allowed to slip into powerful roles and felt the social pressure of the group. Cf. Zimbardo (2007).

Some transhumanists also hope to be able to fulfill the childlike wish for immortality via digital technologies. For example, Alcor Life Extension Foundation, a US non-profit organization, offers to preserve the brain after death, so that the owner of this brain can be resurrected decades, perhaps even centuries later in a cloned, healthy, young, or even in an artificial machine body. The machine body that is to fuse with the brain thus becomes a kind of double or revenant that, unlike the previous human, is supposed to live forever.

From a psychoanalytical point of view, such a desire must be characterized as regressive and narcissistic, because it is one of the most elementary conditions of adulthood to accept one's limitation as a human being and also one's own mortality. Freud also described the idea of doppelgangers, i.e., the idea of a double, as pathological and narcissistic and mentions in this context the ancient Egyptian tradition of sarcophagi, which, as images of the dead, were supposed to guarantee their immortality. Robots that are to be implanted with our brains are, in this sense, nothing other than the modern manifestation of these regressive ideas.

What is not (yet) possible in reality is played out in science fiction films, probably most extremely in the film *Ghost in the Shell* (Rupert Sanders. USA, 2017), based on a manga by Masamune Shirow. Major Mira Killian, the film's protagonist, is the result of a fusion of a human brain and a purely synthetic body that gives her incredible elasticity and strength. Tony Stark the hero of *Iron Man* (Jon Favreau. USA, 2008) also has superhuman powers, he has become a cyborg with the help of a steel battle suit. Unlike Alex in *RoboCop* however, Tony, as a genius engineer, not only has the power over his own programming, but as the immensely wealthy owner of a gigantic company (Stark Industries), he can thus sponsor himself. The problem of a company exterting control over a person's identity does not arise here.

Freud would certainly have seen this film as a typical expression of regressive, unconscious fantasies of omnipotence and invulnerability.[6] And yet these fantasies have no doubt often fueled innovations, such as the desire to fly, for example, the realization of this age-old dream of humankind, in the end became a technical and economic reality. Automobile-based individual transport, the movement of dozens of horsepower with a light hand and gentle foot pressure, the movement of a vehicle weighing tons with minimal use of one's own physical strength, is also the realization of an old dream of unlimited

[6] For the cultural scientist Klaus Theweleit, *Iron Man* would probably be an example of the fantasy of a hard, "soldierly body" of a man who is incapable of building relationships with other people because of his ego disorder. Theweleit (1987).

mobility. Natural science and technology have repeatedly received important impulses from radical ideas in our lifeworld. The development of nuclear-power technology and its peaceful use in the form of light-water reactors or the far more advanced technology of fast-breeder reactors were intended to realize the vision of an unlimited, clean, sustainable, and carefree use of energy. As we know today, things turned out quite differently in the end, and nuclear power as an energy source is still regarded in many countries at best as a transitional technology to a decentralized economy based on renewable energies, which, interestingly enough, falls back on quite old models of energy production (wind turbines, hydropower, manure, geothermal energy, etc.).

New technologies have almost always been accompanied by utopian visions, may it be the invention of aircraft, electricity, or the first automobile which was hailed by Henry Ford in his book *Machinery: A New Messiah* as a "new messiah" which would bring eternal peace and prosperity on earth. However, even if these utopian expectations have never been met, they have often improved the conditions for human life. This is why humanists welcome new technologies. Humanists are guided by the idea of responsible agency and self-determination. They aim at improving conditions of human authorship within the limits of the *conditio humana*.

Giovanni Pico della Mirandola, in his small but very influential writing *De hominis dignitate*, praised almost hymn-like the special human abilities in the arts and sciences. For him, the special human dignity is shown above all in the fact that man—as the image of God, so to speak—harbors within himself divine abilities of creativity and freedom, which come to full development under favorable conditions. However, this humanist idea of human self-development remains limited by natural conditions. The common thread of humanist thought since antiquity, to keep moderation and to defend the middle against the extremes (the *mesotes* doctrine of Aristotle in the *Nicomachean Ethics*), gains new relevance through the new technological possibilities and the transhumanist movement.

Only in the modern age and in times of advanced information technologies, which in conjunction with contemporary neuroscience give rise to expectations of overcoming some of these natural barriers, can humanism turn into transhumanism. Transhumanism differs from humanism in that it questions the natural conditions, even if they belong to the traditional human self-image, in a sense extending the humanist idea of self-determination and self-design beyond all boundaries. This swift from humanism to transhumanism has a famous precursor in Friedrich Nietzsche and his transhumanist ideal of the *Übermensch*, who does not have to take anything or anyone into consideration and can disregard the herd of supposed mediocrity and the values of humanity.

In the course of the film *RoboCop*, the protagonist Alex loses more and more of his moderation and center. He becomes arrogant, aggressive, and increasingly loses his humanity. In the end, when the US government speaks out against the use of cyborg police officers, Dr Norton confesses in an interview that he regrets having worked on the creation of an armed cyborg: "I admit that we know less than we thought. I do believe my research program is sound but what I did with it was wrong."

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits any noncommercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if you modified the licensed material. You do not have permission under this license to share adapted material derived from this chapter or parts of it.

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

20

On the Metaphysics of Digitalization

Los Angeles in 2019. A gloomy city, where it is constantly raining and furnaces throw fire into the air at regular intervals. The city has become so multicultural that people can barely communicate with each other. The dystopian world that Ridley Scott imagines in *Blade Runner* in 1982 accompanied by the melancholic synthesizer sounds by Vangelis is anything but inviting. The city's only bright building, a huge shiny gold pyramid-like structure, belongs to a man named Tyrell. The small man with huge glasses is the inventor and designer of robots that are sent to Mars to make the planet habitable under the harshest conditions. These robots called "replicants" are indistinguishable from humans in their behavior and appearance. Only when one studies their emotional reactions up close, they can be distinguished from humans.

After four of the replicants have illegally left Mars and made their way to Earth, police officer Rick Deckard is asked to find and eliminate them. Deckard is a good replicant hunter, so it doesn't take him long to eliminate almost all of them. Only one remains until the end—Roy. Of all the replicants, Roy is not only the most intelligent and strongest but also the one who undergoes the greatest development in the course of the film. From a kind of primitive robot who speaks in a choppy voice and feels no empathy for humans he occasionally kills, he goes through stages of development much like the developmental stages of a human being. In the beginning, he is impulsive and taciturn like a child, then aggressive and searching for meaning like an adolescent, who then turns into a Nietzschean *Übermensch* until at the moment of his death, he becomes compassionate and spiritual.

© The Author(s) 2022
J. Nida-Rümelin, N. Weidenfeld, *Digital Humanism*,
https://doi.org/10.1007/978-3-031-12482-2_20

If one adopts the emergentist perspective,[1] according to which the next higher level is not determined by the lower level, one cannot rule out the possibility that software-controlled systems will one day have mental states, indeed the capacity for insight. There is no principled argument that only biological and non-physical materiality enables feelings, beliefs, intentions, decisions, etc. Experience with biology teaches us that such transitions are usually fluid, gradual: The newborn child probably has no beliefs and pursues no intentions, but a few months later there can be no doubt about either. The danger in dealing with Artificial Intelligence is to confuse simulation with realization. This often enough leads us to use an inappropriate vocabulary, according to which software systems have "perceptions" and "make decisions." For example, when the navigation system commands "Turn right," we attribute an intention to the system, namely, to make us turn right. But as things stand, that is, at the current state of digital technologies, this would be a mystification—even if the film *Blade Runner* (Ridley Scott, USA, 1982) seems to suggest that there is such an emergentist development from an AI that merely simulates human faculties to an AI that realizes these faculties.

Software systems do not feel, think and decide, humans on the contrary do, as they are not determined by mechanical processes. Thanks to their capacity for insight as well as their ability to have feelings, they can determine their actions themselves, and they do this by deciding to act in this way and not in another. Humans have reasons for what they do. Humans as rational beings are able to recognize mathematical and logical truths, they can distinguish good from bad reasons. By engaging participate in theoretical and practical reasoning we influence our mental states, our thinking, feeling, and acting thereby exerting a causal effect on the biological and physical world. If the world were to be understood reductionistically, all higher phenomena from biology to psychology to logic and ethics

[1] Emergentism is understood in different ways but has two basic principles in common: (1) The properties and laws of the higher (here: biological) level are not reducible to those of the lower (here: physical) level, i.e., it is not already physically determined which biological organisms arise and which properties they have. (2) The properties and laws of the higher (here: biological) level are compatible with the properties and laws of the lower (here: physical) level. Biological organisms consist of physical parts, and all physically describable events and processes of an organism can be explained by physical laws. The biological laws do not violate the physical ones. Analogously, one can imagine the relationship between neuroscience and psychology or between psychology and logic. Often a third property of emergentism is added, according to which something that is the same with regard to the properties of the lower level also does not differ with regard to the properties of the higher level.

would be determined by physical laws: Human decisions and beliefs would be causally irrelevant in such a world.[2]

In one of the most beautiful moments in film history, the last dramatic scene of *Blade Runner*, there is a showdown between Deckard and Roy. Roy, who is seriously hurt and close to dying, is still obsessed by killing Deckard. Like a wild animal, he chases Deckard through a run-down skyscraper, his bare upper body is covered with white paint. Eventually, both reach the roof of the building. When Deckard tries to jump onto another roof, he slips. With the last of his strength, he is able to hold on to an iron bar. He knows that if he lets go, he will fall and die. At that moment, Roy appears above him, half-naked, bleeding, and confident of victory, a Nietzschean *Übermensch*. Roy looks at Deckard, sees him struggling for his life. The spectator expects him to kick Deckard off the roof but instead of this something else happens. Roy looks in Deckard's eyes and reaches out his hand to Deckard to help him back onto the roof.

Roy sits down opposite his former enemy. He knows he is about to die. The rain drips down his face. "I've seen things," Roy says, "you people wouldn't believe. Attack ships on fire off the shoulder of Orion. I watched C-beams glitter in the dark near the Tannhäuser Gate. All these moments will be lost in time. Like tears in rain."

And after a short pause: "Time to die." Then he lowers his head. He is dead.

At this moment, we see a dove fly from the roof into the sky. The bird—easily read as a symbol for Roy's soul—makes it clear what Ridley Scott wants to tell the viewer here: replicants can—if they have enough time and enough experiences and memories—become sentient, empathetic, spiritual beings. We should however be careful not to understand this as a realistic prophecy of how Artificial Intelligence will evolve but as a metaphor for the power of transformation of humans who are able to expand their capacities in order to gain moral sentiments like forgiveness and empathy.

[2] A theory T2 can be reduced to a theory T1 if T2 can be completely derived from T1, which presupposes that the terms of T2 can also be defined with the help of terms of T1. A weaker form of reducibility exists if all empirical predictions of T2 can already be derived from T1 (empirical reduction). Physicalism is the most prominent form of reductionism, according to which all science can be traced back to physics. So far, this has only been successful for parts of inorganic chemistry and has otherwise remained science fiction. Even the reducibility of biology to physics is highly implausible; the reducibility of the social sciences or even literary studies to physics is completely out of the question. This is due, among other things, to the fact that even in the social sciences, but especially in cultural studies and the humanities, terms such as "meaning," "intention," "belief," or "emotion" occur that cannot be translated into physical terms: Intentions or even reasons are not a possible object of physics.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits any noncommercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if you modified the licensed material. You do not have permission under this license to share adapted material derived from this chapter or parts of it.

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

21

Afterword

There is a red thread running through this book. It is the critique of what we call "Silicon Valley ideology," or put positively, its replacement by the idea of a digital humanism which demarcates itself from a misinterpretation of Artificial Intelligence.

In the broadest sense, Artificial Intelligence refers to everything that can be realized with digital techniques, such as computing, from pocket calculators to autonomously acting, self-learning software systems. The greatest misinterpretation is referred to in philosophy as "strong AI" (*strong AI*), according to which there is no categorical difference between humans and computers; software systems that imitate human behavior, judgment, and decision-making also have human characteristics. There are two possible readings of strong AI: a materialistic and an animistic one.

In the materialist reading, human brains are nothing more than complex computers. Therefore, the language of mental properties is fundamentally redundant, and with the progress of natural science, *mentalese* (the language of mental properties) will die out.

If one were to take digital materialism seriously as a worldview, this would mean the end of the human way of life as it is.

In the animistic version one can find in many Hollywood films, software systems are regarded as animate beings that are endowed with mental properties just like humans.

The weak variant of Artificial Intelligence (*weak AI*) does not claim that there is no categorical difference between humans and computers, but merely that all human, cognitive capacities can in principle also be performed by computers.

© The Author(s) 2022
J. Nida-Rümelin, N. Weidenfeld, *Digital Humanism*,
https://doi.org/10.1007/978-3-031-12482-2_21

The optimistic expectation that the cognitive abilities of software systems are capable of unlimited development is often accompanied by a kind of hope of salvation, according to which digital technologies will free people from the toils and limitations of their existence, create new ways of interaction and communication, give us perfect (digital) partners, and enhance human perceptual and cognitive abilities. The message that comprehensive digital transformation will herald a brave and radically new world, as preached by many entrepreneurs from Silicon Valley, is not coincidentally reminiscent of the millenarian rhetoric of salvation that has always had great significance in US-American culture. As a kind of technological millenarianism, Silicon Valley perverts Christian eschatology and presents the digital revolution as the answer to all our economic, social, and even spiritual problems.

Digital humanism counters this ideologization of digital technologies with an attitude of sobriety. Like all technologies of the past, digital technologies are ambivalent. Digital transformation will not automatically humanize our living conditions—it depends on how we use and develop this technology. Digital humanism argues for an instrumental attitude towards digitlization: what can be economically, socially, and culturally beneficial, and where do potential dangers lurk?

Furthermore, digital humanism opposes strong AI. There is nothing to suggest that software systems have perceptions or even emotions, that they can recognize and make decisions. What can be observed here is a more or less successful simulation of cognitive and emotional processes. We should beware of a specific kind of self-deception consisting of us first developing digital machines that simulate emotions, cognitions, and decisions, and then explaining this behaviour by attributing them mental states.

At this point, an argument comes into play that reaches deep into logic, mathematics, and epistemology. The meta-mathematical results of incompleteness and undecidability of Kurt Gödel and other logicians of the early twentieth century are crucial for this. We have interpreted this as a conclusive refutation of the weak AI thesis. With these results, it has been shown that there can be no complete simulation of human judgment and decision-making. There is much to suggest that the categorical difference between humans and machines is also related to this. This is at least the assumption developed by the mathematician and theoretical physicist Roger Penrose in two extensive monographs (*The Emperor's New Mind*, 1989 and *Shadows of the Mind: A Search for the Missing Science of Consciousness*, 1994). Furthermore, one should be aware that attributing mental properties to software systems, would make our everyday use of computers problematic and block further

technical progress in digitalization. If computers recognize, decide, and feel, we would need to treat them with consideration and also grant them human rights, depending on how similar they are to humans. Quite contrary to its intention, strong AI is becoming an obstacle to technological progress.

Digital humanism does not get carried away. It emphasizes the far-reaching immutability of human nature and the conditions of human development. It defends cultural achievements such as the separation of private and public spheres and informational self-determination. It pleads for the strengthening of democracy, also using the new digital possibilities; it warns against a decline of interpersonal connection in times of increasing anonymization and manipulation of Internet communication. It pleads for strengthening the power of judgment in order to enable reliable orientation in the face of an overabundance of data.

Digital humanism is not defensive; it does not want to put the brakes on technological progress in the age of Artificial Intelligence but advocates an acceleration of human progress using digital possibilities to make our lives richer, more efficient, and more sustainable. It does not dream of a whole new human form of existence like the transhumanists; it remains skeptical about utopian expectations but is optimistic about human creativity in dealing with digital potential.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits any noncommercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if you modified the licensed material. You do not have permission under this license to share adapted material derived from this chapter or parts of it.

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

References

Arrow, Kenneth. 1951. *Social Choice and Individual Values*. New York: Wiley.

Barthes, Roland. 1981. *Camera Lucida: Reflections on Photography*. New York: Hill and Wang.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Cellan-Jones, Rory. 2014. Stephan Hawking warns Artificial Intelligence could end mankind. BBC News, December 2. https://www.bbc.com/news/technology-30290540. Accessed 3 March 2022.

Chalmers, David. 2010. *The Character of Consciousness*. Oxford: Oxford University Press.

Cohen, Ori et al. 2012. fMRI Robotic Embodiment: A Pilot Study. *4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*: 314–319.

Condorcet, Marie Jean, and Antoine Nicolas Caritat marquis de. 1785. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des Voix*. Paris: L'imprimerie royale.

de Lagasnerie, Geoffroy. 2017. *The Art of Revolt. Snowden, Assange, Manning*. Stanford: Stanford University Press.

Dreyfus, Hubert L. 1972. *What Computers Can't Do*. New York: MIT Press.

———. 1992. *What Computers Still Can't Do. A Critique of Artificial Reason*. New York: MIT Press.

Freud, Sigmund. 1913. *Totem and Taboo*. Boston: Beacon Press.

Frey, Carl Benedict and Michael A. Osborne. 2013. The Future of Employment: How Susceptible Are Jobs to Computerisation? https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf. Accessed 6 March 2022.

© The Author(s) 2022
J. Nida-Rümelin, N. Weidenfeld, *Digital Humanism*,
https://doi.org/10.1007/978-3-031-12482-2

Gibbard, Allen. 1973. Manipulation of Voting Schemes: A General Result. *Econometrica* 41 (4): 587–601.

Gorz, André. 1999. *Reclaiming Work: Beyond the Wage-Based Society*. Cambridge: Polity Press.

Grice, Paul. 1991. *Studies in the Ways of Words*. Cambridge, MA: Harvard University Press.

Hevelke, Alexander, and Julian Nida-Rümelin. 2015. Responsibilty for Crashes of Autonomous Vehicles: An Ethical Analysis. *Science and Engineering Ethics* 21 (3): 619–630.

Kleene, Stephen C. 1952. *Introduction to Metamathematics*. New York: Ishi Press.

Lem, Stanisław. 1976. *The Star Diaries*. New York: Seabury Press.

———. 1985. *Golem XIV*. New York: Harcourt.

———. 2013. *Summa Technologiae*. Minneapolis: University of Minnesota Press.

Marcuse, Herbert. 1964. *One-Dimensional Man. Studies in the Ideology of Advanced Industrial Society*. Boston: Beacon Press.

Margalit, Avishai. 1996. *The Decent Society*. Cambridge, MA: Harvard University Press.

Maturana, Humberto. 1988. Reality: The Search for Objectivity or the Quest for a Compelling Argument. *The Irish Journal of Psychology* 9 (1): 25–82.

Milkowski, Marcin. 2013. *Explaining the Computational Mind*. Cambridge, MA: MIT Press.

Moravec, Hans. 1988. *Mind Children: The Future of Robot and Human Intelligence*. Cambridge, MA: Harvard University Press.

———. 1999. *Robot: Mere Machine to Transcendent Mind*. Oxford: Oxford University Press.

Nagel, Thomas. 1974. What Is It Like to Be a Bat? *The Philosophical Review* 83 (4): 435–450.

Nida-Rümelin, Julian. 2010. Reasons Against Naturalizing Epistemic Reasons: Normativity, Objectivity, Non-computability. In *Causality, Meaningful Complexity and Embodied Cognition*, ed. Arturo Carsetti, 203–210. New York: Springer.

———. 2018. Responsibility in Philosophy and Law. *Proceedings World Congress Philosophy*.

———. 2023. *A Theory of Practical Reason*. New York: Palgrave Macmillan.

Penrose, Roger. 1989. *The Emporer's New Mind. Concerning Computers, Minds, and the Laws of Physics*. New York: Oxford University Press.

———. 1994. *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford: Oxford University Press.

Popper, Karl. 1996. *Of Clouds and Clocks: An Approach to the Problem of Rationality and Freedom of Man*. St. Louis: Washington University.

Putnam, Hilary. 1981. *Reason, Truth and History*. Cambridge: Cambridge University Press.

Rifkin, Jeremy. 1995. *The End of Work: The Decline of the Global Labor Force and the Dawn of the Post-Market Era*. New York: Putnam.

Searle, John R. 1980. Mind, Brains, and Programs. *Behavioral and Brain Sciences* 3 (3): 417–424.

———. 1992. *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.

———. 1993. The Problem of Consciousness. *Consciousness and Cognition* 2 (4): 310–319.

Simanowski, Roberto. 2018. *Facebook Society. Losing Ourselves in Sharing Ourselves*. New York: Columbia University Press.

Singer, Wolf. 2001. Consciousness and the Binding Problem. *Annals of the New York Academy of Science* 929: 123–146.

Slotkin, Richard. 1973. *Regeneration through Violence*. Middletown, CT: Wesleyan University Press.

Smart, J.J.C., and Bernard Williams. 1973. *Utilitarianism: For and Against*. Cambridge: Cambridge University Press.

Theweleit, Klaus. 1987. *Male Fantasies*. Cambridge: Polity Press.

Tomasello, Michael. 2008. *Origins of Human Communication*. Cambridge, MA: MIT Press.

Turing, Alan. 1938. *Systems of Logic based on Ordinals* (PhD Thesis). Princeton University.

———. 2004. Intelligent Machinery. In *The Essential Turing*, ed. B. Jack Copeland, 431. Oxford: Oxford University Press.

Van Parijs, Philippe. 1991. Why Surfers Should Be Fed: The Liberal Case for an Unconditional Basic Income. *Philosophy and Public Affairs* 20 (2): 101–131.

Zimbardo, Philip. 2007. *The Lucifer Effect. Understanding How Good People Turn Evil*. New York: Random House.