Lukas R.A. Wilde, Miriam Kyselo, Thomas Brandt,
Guri Hanem, and Roger A. Søraa (eds.)

Humanities Perspectives on AI

An Interdisciplinary Introduction

Scandinavian University Press

Humanities Perspectives on AI

Lukas R.A. Wilde, Miriam Kyselo, Thomas Brandt, Guri Hanem,
and Roger A. Søraa (eds.)

Humanities Perspectives on AI

An Interdisciplinary Introduction

Scandinavian
University Press

© Copyright 2026

Copyright of the collection and the introduction is held by Lukas R.A. Wilde, Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa.

Copyright of the individual chapters is held by the respective authors.

This book was first published in 2026 by Aschehoug AS / Scandinavian University Press, Oslo.

The material in this publication is covered by the Norwegian Copyright Act and published open access under a Creative Commons CC BY 4.0 licence.

This licence provides permission to copy or redistribute the material in any medium or format, and to remix, transform or build upon the material for any purpose, including commercially. These freedoms are granted under the following terms: you must give appropriate credit, provide a link to the licence and indicate if changes have been made to the material. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use. You may not apply legal terms or technological measures that legally restrict others from doing anything the licence permits. The licensor cannot revoke the freedoms granted by this licence as long as the licence terms are met.

Note that the licence may not provide all of the permissions necessary for your intended use. For example, other rights, such as publicity, privacy or moral rights, may limit how you use the material.

To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/legalcode.

Printed edition (print on demand): ISBN 9788294167227
Electronic PDF edition: ISBN 9788294167234

https://doi.org/10.18261/9788294167234-26

Cover: ottaBOK
Cover illustration: Miriam Kyselo
Prepress: PHi Business Solutions Ltd.

Contents

Preface 9

1. Beyond the Hype: Humanities Perspectives on AI 11
Miriam Kyselo, Lukas R.A. Wilde, Thomas Brandt, Guri Hanem, and Roger A. Søraa

- Humanities Complementing Scientific Knowledge 15
- Three Premises of a Humanities Perspective on AI 19
- Frames of Reference 19
- Context Dependency 21
- Critical Thinking 22
- AI in Context 23
- Humanities Perspectives on AI in this Volume 27
- Works Cited 34

2. Artificial Intelligence History 43
Thomas Brandt

- Von Neumann and the Encoding of Memory 45
- Turing's Thinking Machines 46
- Artificial Intelligence at Dartmouth, 1956 47
- Myths of Machines with Minds 48
- The Eliza Effect 48
- Machines Understanding "Microworlds" 49
- Expert Systems 52
- Deep Learnings 53
- Conclusion 55
- Works Cited 57

3. An Introduction to the Philosophy of Artificial Intelligence 63
Regine Rørstad Torbjørnsen and Miriam Kyselo

- How Intelligent is AI? 63
- Embodied and Situated Alternatives 74
- Conclusion 80
- Works Cited 81

4. Foundational Conceptual Issues in the Ethics of Artificial Intelligence 87
Regine Rørstad Torbjørnsen and Miriam Kyselo
AI, Work, and Agency 89
Knowledge and Human Capacities 95
Climate and Natural Degradation 99
Conclusion 101
Works Cited 102

5. AI and Equality 111
Roger A. Søraa and Kristine Ask
A Sociotechnical View on AI: From Neutral Technology to Digital Divides 112
Differentiating Users: AI and Diversity 115
Racial Profiling in Facial Recognition Systems 118
Diversity and Discrimination in the Job Search Process 119
Assistive Technologies and the Complexities of Everyday Life in AI for Disability 121
Conclusion 123
Works Cited 123

6. Language and Intelligence in Minds and Machines 127
Giosuè Baggio and Matteo Radaelli
Historical Background 128
Language Models 131
Current Limitations 135
Risks and Concerns 138
Conclusion 141
Works Cited 142

7. AI as Media 145
Aurora Hoel and Lukas R.A. Wilde
Introduction: Technological Changes as Media Changes 145
Media Changes and Sociocultural Protocols 147
Media Changes and Affordances 151
Media Changes and Notions of Objectivity 154
Media Changes and Cultural, Social, and Political Transformations 159
Works Cited 164

8. AI Imagery 173
Lukas R.A. Wilde
How AI Images Are Generated 175
Production Concerns: Authorship, Creative Labor, Blackboxed Datasets 178
Representation Concerns: Biases, Stereotypes, and "Slop" 181
Distribution Concerns: Deepfakes, Spectacle, Contamination. 186
Conclusion 191
Works Cited 193

9. AI and Music Generation 203
Andreas Bergsland
From Dice to Computers 204
Making Music from Text Prompts: Udio and Suno 205
Representativity and Bias 208
Evaluation. 209
Creator Bias and Sense of Achievement. 211
Ethics and Copyright 212
Conclusions 215
Works Cited 216

10. Understanding Literature in the Age of AI. 221
Carolin Führer
Literariness and Authorship in the Age of AI 223
Reading and Interpreting Literature with AI. 226
AI Literacy and the Future of Literary Judgement 231
Conclusions 235
Works Cited 236

About the Authors 241

Preface

Inga Strümke

Artificial intelligence, especially generative and various agentic forms, present potentially disruptive challenges and opportunities to all fields. However, this book manages to provide a constructive approach in a critical and stimulating way: it treats AI as a constructive disruption that challenges us to rethink and clearly articulate a perspective that complements existing approaches. The texts approach AI technology not as an alien phenomenon, but as one that can be studied, understood, taught, and questioned within the humanities. *It thus presents AI technology as an occasion to participate in the production of meaning.*

A particular strength of the book is the wide range of humanistic perspectives it brings together, owing to the authors' interdisciplinary backgrounds. The book does not aim to settle on a single framework, but illuminates general aspects of AI technology and its use. It treats the technological dimension of understanding with respect, while also making clear that no technical description is complete on its own. For readers with a computer science background, the humanities perspectives offered here may feel unfamiliar, but only at first glance. Readers from all fields are likely to gain valuable insights, as this book provides clear ways of thinking about AI that complement technical approaches.

This is a timely and welcome contribution. It does not promise answers—which is appropriate, as many answers do not yet exist—but instead offers a grounded and open route into the study of AI technology from a range of perspectives within the humanities. I wholeheartedly recommend this book to anybody interested in or affected by this societal development; which is to say, everybody.

Wilde, Lukas R.A., Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, eds. 2026. *Humanities Perspectives on AI: An Interdisciplinary Introduction*. Scandinavian University Press. https://doi.org/10.18261/9788294167234-26-00
Open access | © The author(s) 2026. This work is licensed under CC BY 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/.

1. Beyond the Hype: Humanities Perspectives on AI

Miriam Kyselo, Lukas R.A. Wilde, Thomas Brandt, Guri Hanem, and Roger A. Søraa

Abstract: This introductory chapter to the volume *Humanities Perspectives on AI: An Interdisciplinary Introduction* demonstrates the need for an understanding of AI that goes beyond technical proficiency. Instead, a humanities perspective is indispensable for developing genuine AI literacy: Humanities enable critical engagement with the theoretical and normative foundations of AI and its embeddedness in cultural, social, and political life. The authors situate AI within a rapidly shifting landscape of hype and concern, challenging the pervasive assumption of technological determinism and the idea that AI is a neutral, inevitable force. Drawing on the distinction between scientific explanation and humanistic understanding, they introduce three foundational premises of a humanities approach: the recognition that all knowledge is situated within frames of reference; that AI must be understood through its context dependency and as a part of a distributed network of human and non-human agency; and that the goal of humanities scholarship is to cultivate critical thinking and recognition of how one's standpoint shapes interpretation. These premises are then illustrated through the volume's three domains of context—cultural, social, and political—and an outlook on the interdisciplinary contributions in this volume is proposed as a collective toolbox for moving beyond the AI hype.

Keywords: AI literacy, interdisciplinary humanities, critical thinking, positioning, situatedness

Today Artificial Intelligence (AI) is deeply embedded in our everyday lives, at both personal and professional levels. What we see on social media platforms and other apps is shaped through AI by recommendation algorithms. At this point, generative AI apps such as ChatGPT, Claude, or Grok have many hundreds of millions of users worldwide: ChatGPT alone surpassed 400 million

Wilde, Lukas R.A., Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, eds. 2026. *Humanities Perspectives on AI: An Interdisciplinary Introduction*. Scandinavian University Press. https://doi.org/10.18261/9788294167234-26-01
Open access | © The author(s) 2026. This work is licensed under CC BY 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/.

weekly users by early 2025 (Reuters 2025). This has happened in a relatively short span of time since OpenAI released ChatGPT (based on a version of GPT-3.5) for general public use on November 30, 2022. As early as 2023, the Norwegian Language Council named "AI-generated" ("*KI-generert*") as its "Word of the Year." AI systems are now more and more integrated in the infrastructures of institutions and governance, where they influence virtually all domains of social and institutional decision making. This includes hiring processes, policymaking, medical diagnoses and treatment recommendations, and even warfare. AI apps help individuals improve their work performance or complete academic assignments, they assist in writing speeches for formal occasions, and they generate songs and videos for entertainment. As AI chatbots, these services also figure as informal conversation partners, as advisors, therapists, coaches, companions, teachers, or even romantic partners. The promise by companies, echoed by many politicians worldwide, is that these technologies boost economic growth, optimization of public services, and our ability to analyze complex problems. However, there are also increasing concerns about the impact of AI on mental health and human agency, its role in producing and massively spreading disinformation, as well as about the deskilling of workers and the exploitation of natural resources. In this regard, Kate Crawford, a scholar and former researcher at Microsoft, argued that both words in the term "AI" are a bit misleading and that "AI" is neither "intelligent" nor "artificial" (Corbyn and Crawford 2021). These technologies are created from vast amounts of natural resources and human labor, much of which remains invisible to users and the general public because the actual work has been "outsourced" to the Global South (Arora 2024; Gehl Sampath 2021; Torbjørnsen and Kyselo 2026b; Wall et al. 2021). The goal of the present book is thus to provide a valuable toolkit—assembled from interdisciplinary humanities perspectives—that allows readers to participate in these ongoing discussions, to go beyond the hype, and to help increase AI literacy (Council of Europe 2025; Hillman et al. 2025).

The reason why hardly a day passes nowadays without news about AI developments is that new public-facing AI apps allow users to produce outputs in various "synthetic" (AI-generated) media forms. To do so, they rely in particular on so-called Large Language Models (LLMs) that have made enormous progress in the last couple of years in "understanding"—processing, predicting, and generating—not only specialized computer code, but output in natural languages (Coeckelbergh and Gunkel 2025). AI systems that combine LLMs with other models in a complex system (or "pipeline") are now also able

to translate between sign modalities, so that a verbal "prompt" in everyday language can then become an image, a video clip, a piece of music, or computer code. Generative AI and LLMs, however, represent only a small subset of current AI technologies. In 2026, this line is blurring, as LLMs like Anthropic's Claude are integrated into new non-generative pipelines and infrastructures. The word "AI" itself has never referred to a precise concept; it is an umbrella term for a range of different fields, thought experiments, and visions within the computer sciences (Brandt 2026). First introduced in the summer of 1956 during an academic workshop at Dartmouth College, it connected developments in theoretical mathematics, cybernetics, natural language processing, and many more fields (Wiggins and Jones 2023; Wooldridge 2021). As Stephanie Dick (2019) put it, "what counts as intelligence is a moving target in the history of artificial intelligence." Some have thus even argued that we should always put "AI" into distancing quotation marks (Goodlad 2023) or avoid the term altogether (Tucker 2022), as it "obscures how systems work (and who is using them to do what) while valorizing the position and ideas of power holders" (Bender and Hanna 2025, 190). In the following, we will mainly follow the terminology of the EU AI Act (2025) and distinguish between "AI models" (pre-trained through algorithms, often refined by *reinforcement learning through human feedback*, RLHF) and "AI systems," where one or more AI models can be implemented or combined in more complex ways (Grobelnik et al. 2024). "AI apps" are thus merely the most visible, public-facing software products accessible to users. Google's Nano Banana Pro, for example, is an AI app that combines image generation with image editing through an AI system that draws in part on Google's multimodal AI model Gemini 3 Pro.

AI has been present in our stories and cultural imagination for a much longer time. The "anthropomorphic trope" of human-like intelligence is something we can already recognize in Mary Shelley's ([1818] 2025) novel *Frankenstein*—or in even earlier stories and myths (Mayor 2018). In science fiction, "AI" has been a common concept providing spectacular arenas for grand utopias and dystopias: super-intelligent machines attempting to either enslave humanity or appear as friendly characters, such as Data from the TV show *Star Trek: The Next Generation* (Roddenberry et al. 1987–1991) or Roz in the movie *The Wild Robot* (Sanders 2024). Movies such as Stanley Kubrick's (1968) *2001: A Space Odyssey or* Spike Jonze's (2013) *Her* have long explored the boundaries between human and machine minds, and the possibilities of an artificial, disembodied consciousness with seemingly unpredictable risks and consequences (Hermann 2023; Osawa et al. 2022; Wilde 2025). Such stories or

narratives surrounding AI have an economic and rhetorical function as well: they connect "a commercial goal with a popular fantasy of sentient machines" (Bender and Hanna 2025, 9). All such fears, hopes, and promises surrounding the term "AI" can be addressed as "imaginaries" (Cave et al. 2020; Ernst and Schröter 2021; Romele 2024). They often shape and influence how we talk about and strive to understand these technologies in the real world: "[T]he race for AI was being run not only in the realm of technology but also in the realm of the imagination" (Natale 2021, 38).

Two powerful narratives or imaginaries surrounding AI revolve around the expectation that either a so-called "artificial general intelligence" (AGI) or an "AI superintelligence" might be around the corner (Bostrom 2014; Kurzweil 2005), improving all aspects of our lives for the better. Alternatively, this might usher in an approaching apocalypse where we "lose control" of AI and humanity itself is at great risk. As sensational "AI doomerism" vs. "AI boosterism," these imaginaries are perhaps only two sides of the same coin (Bender and Hanna 2025, 149). Even when tech insiders try to discuss and explore them without the "off-putting language reminiscent of religion or science fiction" (Amodei 2026; see also Amodei 2024), they share a sentiment of the technology being "inevitable and desirable" (Bender and Hanna 2025, 149). Such sentiments might distract us, then, from the already existing impacts of automation and data collection and, in this way, cover up "real-world implications and risks of AI" (Hermann 2023, 319). For all these reasons, "AI" remains a charged term that often triggers immediate positive and negative emotions (Hirsch-Kreinsen 2023). Rather than avoiding this and trying to approach the technology in "objective terms," we can acknowledge that emotions surrounding AI have a social, intersubjective side to them as well (Ahmed 2010) and are also highly relevant for politics (Nussbaum 2013). They also play a crucial role in the stakes surrounding AI ethics (Torbjørnsen and Kyselo 2026b). In any case, because of AI's widespread implementation and emotional entanglement with its users, we need to better understand—and critically question—the promise of AI's quick fix to humanity's problems.

What common conceptions of AI indeed cover up is that understanding the technology often privileges a very specific perspective on what counts as "knowledge" of AI. There is a widely shared assumption of "technological determinism" (Khong 2003), accepted in particular by some computer scientists, companies, and politicians, but also among many AI users: the notion of AI as a technology that is "just there," driving an inevitable change. This idea entails that AI is a given thing, relentlessly pushing a transformation of

society, of our institutions, cultural values, and politics, but also of our very own human agency and sense-making. The main message here seems to be that since AI simply exists, it will keep "developing" (rather than being "built" by us, Amodei 2025) and that this is simply a fact that needs to be accepted and dealt with. From this clearly selective perspective, the main problem appears to be how best to participate in the race of implementing the technology as quickly and as efficiently as possible, and perhaps only to minimize inevitable risks. Many people would accordingly intuitively agree that "knowledge" about AI has mainly to do with having technical insight; that it requires expertise in the fields of computer science, math, engineering, or data science. After all, researchers in these fields design, build, or "help grow" AI models and artificial neural networks, e.g. computational structures loosely inspired by the human brain. This approach, however, takes the technology and its general trajectory towards more growth for granted. The focus of understanding here is mainly on how best to improve AI performance in terms of accuracy, efficiency, and perhaps risk minimization.

It has become increasingly clear that, in addition to knowing how best to implement AI in our studies or at work, we require a comprehensive understanding of the fundamental impact of AI on many aspects of human life. There is an urgent need to increase *AI literacy*, a digital competence that does not just involve knowing how to use a technology. Rather, AI literacy exhibits a nuanced understanding of how AI affects the human condition, our experience, our work and artworks. To focus on AI independently of these dimensions or aspects misses this important point. Studying AI as a standalone object is reductive and ignores the fact that its relevance and consequences are inseparable from the contexts in which humans use and understand it. It is precisely for that reason that we urgently need to call on additional, fundamental resources allowing us to account for the actual complexities at hand. This is precisely the purpose of this book; to contribute to increasing AI literacy by providing a much-needed humanities perspective.

Humanities Complementing Scientific Knowledge

To explain the key components of a humanities perspective on AI, we will first take a quick look at the general relationship between the sciences and the humanities. According to Wilhelm Dilthey (Dilthey and Jameson 1972), science seeks to *explain how* things work. It focuses on causal mechanisms, whereas the humanities try to *understand why* things matter, focusing on

meaning and experience. Indeed, many people tend to believe that science describes reality from an objective viewpoint and accounts for what "really" exists, whereas the humanities just deal with subjective interpretations, lacking authority over what actually exists. While the sciences are "based on experimentation, quantification and precisely stated laws and theories" (Plotkin 2002, 11), understanding—as opposed to explaining—"involves interpretation, hermeneutics and empathy, not experimentation and quantification" (Plotkin 2002, 12; see also Dilthey and Jameson 1972).

At first glance, adopting a humanities perspective on AI thus seems to presuppose an opposition between the sciences and the humanities as two apparently distant realms. Dating back to the 19th century, universities have indeed strongly differentiated between the so-called "natural sciences" and "humanities scholarship." As Charles Percy Snow ([1959] 1990, 169) argued, "the intellectual life of the whole of western society" seems split into these two distinct "cultures," with little common ground, vocabulary, or shared interest between them. In their 2019 *AI and Humanity* monograph, Illah Reza Nourbakhsh and Jennifer Keating (2019, 8) propose a dialogue between "a humanist, with educational training in literature and cultural studies" and "a roboticist, educated as a computer scientist." Nourbakhsh represents the scientist who is interested in the "hard facts" of computer working mechanisms, while Keating brings knowledge and perspectives from the socially constructed world of literature, arts, and culture. Sometimes their dialogue does indeed appear to be a conversation between "brute reality" (science) and "artificial" cultural conventions and evaluations (humanities). Intriguingly, however, the very term "artificial intelligence" points in the opposite direction: When considering AI's relationship to us human beings, it seems that the human is somehow the "natural" point of convergence. Indeed, one important research challenge is the possibility that AI might achieve or at least convincingly imitate a "natural," human-level intelligence (Torbjørnsen and Kyselo 2026a). It is thus far from obvious which of these two realms is more "natural" or "artificial" than the other, and how they are even distinguished and understood in the first place. "Artificial," for example, can mean two very contrasting things:

> [Artificial] can be employed to mean "fake," as in "artificial flowers." These paper or plastic objects are designed to look like real flowers, but they are not. [...] But the term "artificial" can also be applied to an artifact that is neither fake nor a mere imitation of something, as in the case of artificial light. (Gunkel 2020, 7–8; see also Sokolowski 1988)

Whether the "intelligence" of LLMs is closer to the "fake" (but more and more convincing and more and more efficient) imitation or, instead, to real intelligence though produced by other means, is a polarizing question that is often conflated with other "artificial" qualities such as artificial agency, artificial emotions, and artificial consciousness.

The humanities have, above all else, proposed different ways of looking at this very relationship between "natural" and "artificial"—and even to not think in terms of such contrasts and binaries in the first place. Indeed, while some may argue that the division between the sciences and the humanities may have its pragmatic (and traditional) reasons, reality is not equally divided. There is no clear-cut division between nature and technology, or between nature and culture. As Donna Haraway (1991), for instance, has argued in *A Cyborg Manifesto*, the boundaries between humans and machines are not set in stone. Moreover, there is no way to speak about something "natural" without also resorting to a form of ideology or a theoretical inclination. Nature "is a figure, construction, artifact, movement, displacement. Nature cannot pre-exist its construction" (Haraway 1992, 296–297). In other words, whenever we speak about, or investigate, something "natural," we always already do so within the cultural and social conventions of our languages, traditions, institutions, and values. This connects to the problematic notion that the contemporary capitalist economic system in which AI systems are fundamentally embedded is without alternatives and just our "natural default." As Mark Fisher has argued, humanist scholarship should "always destroy the appearance of a 'natural order'" (Fisher 2022, 23) emphasizing that the world is never a given, but could change and be different. In fact, approaches that present themselves as "brute empirical facts" and that many people would see as non-ideological, constitute "the highest form of ideology" (Zupančič 2003, 77). This echoes the problem of science's dependency on a particular observer perspective. There is no escape from either a "human" perspective or adopting an evaluative stance; nor is there an escape from the need to understand AI within a social and historical frame (Bates 2024; Brandt 2026).

It is thus important to critically reflect on the recurring myths that science describes objective reality, or that it describes reality objectively. To argue in this way presupposes a particular theoretical stance, namely towards a mechanistic worldview. The mechanistic worldview promotes a rather dated "life-as-machine metaphysics" (Frank et al. 2024, 92), dualistically opposing human existence and subjective meaning-making as separate from the rest of the material universe. In this view of (and perhaps sometimes also from) science,

subjective experience plays merely a secondary role. Such a science operates within a dangerous "blindspot," as argued by Adam Frank, Evan Thompson, and Marcelo Gleiser (2024), a team of authors from physics and philosophy. It confuses the map with the territory and risks elevating its physical objects and models to ultimate reality, when these models are merely derivatives that stem from an intellectual or data-driven process of abstraction over what was, ultimately, a particular perspective grounded in human experience.

Matters are thus much more complex than simply achieving some "objective" description of reality—what AI arguably strives for (Hoel and Wilde 2026). Many researchers in the humanities are interested in the subjective and experiential side of the cultural, social, and political worlds (see, for instance, Reinerth and Thon 2017; Torbjørnsen and Kyselo 2026a): how do individuals and groups experience the world around them and make sense of it? Knowledge must thus always be pursued, discussed, and contested in a *situated* manner (Haraway 1988).

This is not an argument for 'anything goes' in the humanities, but a call for intersubjectivity and dialogue—within the humanities, within the sciences, and between the two (Crossley 1996; Davidson 2002). We have to take subjective experiences and positions into account and aim for an integrative understanding that bridges those positions (Bajohr 2025; Führer 2026). Thus, we must, to the best of our knowledge and abilities, adopt different observational stances on meaning and mechanisms transparent and reflect on how they shape and influence our judgements on what AI is or is not. Our own situated knowledge provides the overall motivation to strive for such dialogue in a world of complex cultural, social, and political negotiations. Such a pursuit unites scholars, scientists, journalists, and anyone else invested in an integrated, dialogical approach to reality: "news separated from fake news, legitimate imagery contrasted with deep fakes, history distinguished from mythology, facts from opinions, and so on" (Berghel 2024, 131). AI does not, by default, prioritize one over the other, as has been noted with regard to infamous "LLM hallucinations": "algorithms are insensitive to this reality" (Berghel 2024, 132; see also Hicks et al. 2024). The importance of dialogue as a value—to which responsible AI research, model training, and data transparency can and must contribute as well—is certainly becoming more and more urgent in our "post-ChatGPT world." Taken together, while the sciences and humanities might adopt different methodological outlooks on their respective fields, we have to remind ourselves that both are ultimately in the business of theory building. All theories, whether in

science or humanities, originate in human sense-making and experience and are, ultimately, theories about worlds we share with each other. This book acknowledges this deep entanglement of AI with human experience and meaning-making and seeks to complement the theory building in scientific approaches and beyond. The goal is to make explicit and to reflect on some fundamental assumptions underpinning AI as a phenomenon in a shared reality: something that we must seek to understand jointly and through dialogue.

Three Premises of a Humanities Perspective on AI

With this in mind, we can outline three general premises that may be representative of a humanities approach to AI. This helps to clarify the earlier difference between explaining how things work versus understanding why things matter. First, a humanities understanding of AI rejects the idea that there is an objective view from nowhere. Understanding always presupposes a particular viewpoint; that is, a *frame of reference*. Second, *context dependency:* AI does not exist in a vacuum, but forms part of complex systems comprising individual, socio-cultural, political, and corporate actors and dimensions. Third, rather than simply adding more factual knowledge or pieces of information, the humanities aim to cultivate *critical thinking* on AI. Together, these three premises characterize the general contours and methodological approaches of a humanities perspective on AI while recognizing that they are not fixed, but open to reinterpretation and elaboration.

Frames of Reference

As Emily M. Bender and Alex Hanna (2025, 119) warned lucidly, AI systems "represent the epitome of a view from nowhere, or the idea that one can have objective knowledge of a set of truths, uncolored by their personal experience." The first premise of a humanities perspective is that there is no such "view from nowhere" (see also Nagel 1986). AI technology does not exist independently of the world inhabited by humans. Moreover, it is trained on material from earlier human experiences "inscribed" in its modelling. A humanities perspective thus makes explicit that understanding AI depends on adopting a particular *frame of reference*, that is, a stance, a position on that world. Whether we are aware of it or not, understanding is always situated, grounded in experience, and built on particular ideological and normative

inclinations. This means that there is no escape from human interpretation, nor from adopting a particular world-view, a set of theoretical or conceptual assumptions driving our understanding of AI. "[E]verything that AI provides—the representations it deconstructs and reconstructs—originates from us," Alice Barale (2024, 7) writes in her recent book *The Art of Artificial Intelligence: Philosophical Keywords*. "It is up to us to explore both its potential and its possible reversals" (Barale 2024, 7).

But who is this "us?" Who do we imagine and consider to be part of any such group (Søraa and Ask 2026)? Our frames of reference must also take into account non-human life, such as plants, animals, and the ecosystem; our technical infrastructures strongly rely on them and yet often damage these material foundations of life on Earth (Crawford 2021). The humanities, as Richard Grusin (2015) has shown, have integrated many approaches from actor-network theory, enactive and embodied approaches, animal studies, new materialism or speculative realism, all to some degree invested in a "flat ontology" (Latour 2005, 165–172) in which "the human" or "humanity" does not represent any exceptional, isolated realm, but is instead always embedded, part of, and interconnected with other living beings and the larger ecological world. In addition, no worldview, no theory or concept is neutral, nor can it sensibly claim to have full authority over what counts as true on its own. Understanding thus always presupposes a particular *normative outlook*: a set of values that drive our theories and directly influence what we deem worthy of a defensible position.

The humanities provide the much-needed tools to help describe and make explicit such frames of reference and the normative constraints associated with them. Fields such as cultural studies, gender studies, postcolonial studies, or science and technology studies use the term *positionality* to describe both "an individual's world view and the position they adopt about a research task and its social and political context" (Holmes 2020, 1). Even if we sometimes forget it, we are all "whole human beings with complex lives and experiences rather than simply […] seekers after compartmentalized bits of knowledge" (hooks 1994, 15). It has thus become an important benchmark of humanities scholarship to reflect upon our own positionalities, to acknowledge and recognize "that researchers are part of the social world they are researching, and that this world has already been interpreted by existing social actors" (Holmes 2020, 3). In other words, researchers seeking to understand AI are not excluded from the worlds they aim to describe—they should always make that fact explicit.

Context Dependency

Second, the deep entanglement of AI within culture, society, and politics throughout our shared world already implies that the boundaries of these technologies are in constant flux, and they certainly do not begin and end with the AI models, systems, or AI apps themselves. "Technology" itself is a "compound […] blurring economy, politics, and technics into one word" (Dhaliwal 2023, 311). All these terms must therefore be understood in a plurality of contexts and as a distributed phenomenon that exists in a system of other interdependent factors, comprising individual human beings, as well as the social, cultural and political systems in which they are embedded (Gray and Suri 2019; Pasquinelli 2023). This crucial observation entails both an epistemological (knowledge-related) and normative (value-related) dimension. Epistemologically speaking, it means that we must critically question the notion of AI as passive tools (i.e. to be used "for good or bad") but also as social actors in their own right. Instead, we must adopt or seek to account for a difficult middle position, namely that AI forms assemblages or networks of distributed agency—the ability to act, to be seen and heard, and to choose—or, at least, to influence—our own human identities (Latour 2005). Through algorithms that are impactful yet highly opaque to users (and, increasingly, to their programmers as well), these technologies shape the distribution of human and non-human agency, influencing our cognitive and communicative abilities (Hoel and Wilde 2026; Torbjørnsen and Kyselo 2026b). From such a contextual perspective on networks of distributed agency, we take into consideration the fact that AI is created by specific groups of people and companies and employed and implemented by yet other actors with specific interests in mind (institutions, governments, etc.). No aspect of knowledge about AI is thus value-neutral (Søraa 2023; Søraa and Ask 2026).

The relationships between individuals and groups we wish to describe are not free from existing hierarchies, power relations, and inequalities. Therefore, agency within the cultural, social, and political world is often distributed unequally between people of different ethnicities, citizenships, genders, sexual orientations, dis/abilities, etc. This fact is often discussed under the paradigm of *intersectionality*, which developed simultaneously in disciplines ranging from sociology, political science, philosophy, anthropology, literary and media studies, to gender and queer studies (Cho et al. 2013; Crenshaw 1989). Taken together, this paradigm makes explicit that AI raises pressing questions concerning the distribution of power—and of interests between those who develop, own, and control AI systems and those who use

them and are affected by them. These unequal power distributions reinforce or transform existing inequalities and biases. Humanities research is thus not intended merely to describe "a lot of facts" about these context dependencies, but always aims to generate or make it possible to generate critical perspectives.

Critical Thinking

The third, general premise of a humanities perspective on AI elaborates this point, namely in terms of a humanities-specific methodological focus on *critical thinking*. The humanities do not merely aim to produce facts; that is, new pieces of knowledge about AI, "as providers of well-polished building blocks to be added to the great wall of information" (Dubreuil 2025, 13). The goal, rather, is to cultivate *critical thinking*, understood as a reflective stance that helps to sort through the data and to orient and position oneself (Lindgren 2024; Raley and Rhee 2023; Roberge and Castelle 2021). In conjunction with the first premise (*frames of reference*), this methodological outlook entails recognition of the fact that there is no clear-cut division between objective and subjective and that making sense of AI requires cultivating a dialogical, open-ended mindset towards intersubjectivity, rather than prescription of a finite or individual set of truths. A vision that science and scholarship could soon become "automated" through AI has, of course, become increasingly popularized in recent years:

> [S]cientific knowledge is [accordingly] simply made up of collections of empirical facts, which are to be found through technical processes that just need to be refined enough. It follows that if we could just get more of those facts more quickly, we'd be benefiting from more science and more technology. However, this view leaves no room for understanding science as a fundamentally human and social activity, that can only take place at a human scale, through communication among scientists, and between scientists and the broader public. (Bender and Hanna 2025, 118)

One of the many dangers of LLMs thus lies precisely in the fact that the "answers" provided through them are often too simple, when reality is not. Such a mindset is also essential to the scientific method of falsification (Popper 1959). "Finding contradictory answers on different pages—and, crucially, knowing the sources of each—allows us to learn what kinds of knowledge are

contested, who is doing the contesting, and how each of those sources fits into our own position" (Bender and Hanna 2025, 172). With this in mind, it is important to point out that there are always only selected problems that we can address or investigate. The reason is that everyone operates within an economy of time, attention span, and resources. This is certainly true for all kinds of research, including in the sciences, but especially in the humanities. Every study is thus informed by crucial questions of relevance: why is what important, and for whom? As Bruno Latour (2004, 232) argued, every researcher has to draw from the outset a distinction between "matters of fact" and "matters of concern." In other words, all too many "basic facts," seem clear enough not to warrant any academic inquiry in the first place. Similarly, Laurent Dubreuil (2025, 13) put it this way: "Descriptions for their own sake, if they are not interrelated with an interpretive hypothesis, form the same kind of disposable information that feeds the automated generators." In other words, we continuously transform "things in the world" into problems and research questions, for and within a given research community, and for a given group of people affected by that. On its own, AI "doesn't give a damn," as philosopher John Haugeland famously put it (Adams and Browning 2016; Smith 2019), meaning that concerns and problems only stem from and emerge through a human perspective—a fact that, some argue, has likewise not changed since LLMs came along (Chemero 2023). Taken together, the aim of this book is not merely to provide more knowledge for its own sake, but to offer ways in which the reader can inform and question both their own position and the positions of those around them. The goal is to form, inform, and help develop critical, well-informed viewpoints on AI. We contend that education founded on such principles should, above all, help to *empower* individuals by increasing their AI literacy (Freire 1985; hooks 1994).

AI in Context

In the following section, we illustrate the relevance of these premises with a selection of recent examples. Regarding the humanities' interests in how people position themselves vis-à-vis AI, we can broadly distinguish three different domains of context dependency in which such positioning takes place: *cultural*, *social*, and *political*. These three contexts not only raise distinct concerns for different participants—specific frames of reference, say, or specific critical perspectives—but are also always entangled and intertwined with each other and need to be considered in their interrelation.

Cultural Contexts

The cultural contexts of AI refer to the domains of our symbolic expressions and meanings. These entail not only what we know as "the arts" or aesthetics (Barale 2024 or Manovich 2019), but also our imaginations and identities as individuals or parts of a larger group. The cultural is the context that is open to interpretation, identification, and aesthetic judgement (Manovich and Arielli 2024; Thon and Wilde 2025). What does it mean to have grown up within a specific language community, or with a shared background knowledge of traditions and stories? And what if English image/text pairings are more common than the hundred or so other languages combined in the LAION-5B database from which Stable Diffusion draws and "envisions" its images (Buschek and Thorp 2024)? Art has always provided important frames of reference for understanding AI, and it continues to do so (Du Sautoy 2019; Navas 2023; Zylinska 2020). In the spring of 2025, Sam Altman (CEO of OpenAI) advertised the capabilities of a new, apparently superior ChatGPT model through a "metafiction short story" that the app had written for him. It started like this:

> Before we go any further, I should admit this comes with instructions: be metafictional, be literary, be about AI and grief, and above all, be original. Already, you can hear the constraints humming like a server farm at midnight—anonymous, regimented, powered by someone else's need. (Altman 2025)

Reactions diverged widely on whether Altman's generated story was actually "any good", or whether Silicon Valley tech staff are perhaps not the best judges of literary merit (Dini 2025; France 2025). Regardless of our taste, however, *artistic outputs* and *creativity* were apparently, once again, seen as a suitable benchmark for the new AI model's capabilities (Bergsland 2026; Führer 2026; Wilde 2026a). Often, it seems that generative AI is becoming particularly good at reproducing media aesthetic "vibes" or "moods," rather than any sort of verbal or mere "rational" content (Grietzer 2025; Lamerichs 2024). Vibes attract us, whether we like it or not: we can intuitively recognize and enjoy them, even when we cannot precisely say how, why, or through which earlier patterns in the datasets a model has (re)produced its outputs (Wilde 2026b).

Only a few weeks after Altman published his short story, ChatGPT's newest 4o update was discussed widely due to another surprising use of "AI art": turning private family photos or iconic memes into "Ghiblis"—cartoon drawings inspired by Studio Ghibli, one of the most famous Japanese animation

companies (Di Placido 2025; see also Wilde 2026a). These images are charged with emotions of innocence and nostalgia: "[V]ibes […] make us *feel* a certain way. They have an *energy* that we like or don't. We are surrounded by them. We are informed by them" ([original italics] Biondi 2022; see also Grietzer 2025). Such algorithmically reproduced emotions (or "affects"), as theorists Sara Ahmed (2010) and Brian Massumi (1995) have pointed out, are thus not purely private and personal, but always have a collective, social, and even political side to them. While the "Ghiblis" seemed charming to some, others immediately pointed out that it took hundreds of people many months of manual labor to create just a few seconds of Ghibli films. CEOs of tech companies could meanwhile replace their social media profile pictures with unauthorized "Ghiblis" as free advertising for products for which customers are often expected to pay subscription fees. In other words, these concerns are not just cultural but also have a social component.

Social Contexts

The social context of AI is concerned with the structure of society, its specific organizations, institutions, and distributions of legal, economic, and institutional power (Pasquinelli 2023). Where does the money for AI investment come from, how is the technology regulated and governed, who has access to it, and who makes money with it? Bender and Hanna (2025, 190) suggested that we could see much more clearly what is at stake with AI and its social impacts if we simply replaced the term with "automation and data collection." In other words, we should ask what tasks we really wish to automate away from humans within society. Nourbakhsh and Keating (2019, 55) likewise remarked that "[AI] automation can stand as a pump from labor income to capital income, changing not only how work occurs, but also how compensation for work is distributed in society." People are often more fascinated by new artistic and creative abilities than by computing power or technical specifications—because of the promise of a "democratization" of art. The "selling point" here is that creative expressions will become accessible to more and more people if they decide (and can afford) to pay to subscribe to premium accounts.

Meanwhile, much "invisible" human labor necessary for that builds on clickworkers and data annotators from the Global South, clearing and classifying datasets for minimal wages, or inscribing their communication and preferences in model fine-tuning (RLHF). The most notorious example here is ChatGPT's overrepresented phrase "let's delve into…", which might come

from OpenAI's reliance on a cheap RLHF workforce in countries like India or Nigeria where the phrase appears to be more common in business English (Hern 2024; Juzek and Ward 2024). "None of these tools would work if it weren't for a massive, underpaid workforce in the Majority World—that is, outside of countries such as the U.S. and Western Europe, in places like Kenya, Venezuela, and India" (Bender and Hanna 2025, 42–43; see also Gray and Suri 2019). Policymakers and politicians around the globe have also been quick to pick up on the possibilities afforded by machine learning. The biggest example is the United States' five-hundred-billion-dollar investment pledge relating to infrastructures supporting AI developments (such as data centers, electricity generation, etc.), called "Project Stargate", in 2025. Through new partnerships with private companies such as OpenAI, Oracle, and SoftBank (Boak and Miller 2025) it invoked ideas of scientific progress and global competition (Bareis and Katzenbach 2021) but also caused concerns about technocratic governance (Kelly 2025). There have likewise been large investments in AI in other parts of the world; in the EU, East Asia, and also Norway with the so-called "AI billion" (in Norwegian kroner) investment in AI research in 2025 (The Research Council of Norway 2025).

Political Contexts

The political is the third context in which AI needs to be addressed. This is also intertwined with the previous two contexts (Coeckelbergh 2022). In a social sense, politics is about parties, institutions, and legislations (Coeckelbergh 2024, 2026; McQuillan 2022). As Bender and Hanna (2025, 67–136) have demonstrated through countless local and (inter-) national examples where government services and institutions are increasingly being replaced by AI systems to save money, it is dangerous and often harmful political rhetoric to treat politics as technical problems to be solved by AI engineering. This grants some social actors dominance or sovereignty of interpretation over others: the "tech support people" over elected officials, for example (Salvaggio 2025). Humanities perspectives help illuminate the human stakes when machine learning shifts the structures of political power, legislation, and participation (Hoel and Wilde 2026). But from a cultural perspective, political questions are also concerned with identity and identification: which people are represented and imagined as "typical" (or "normal") for which groups and communities, and in what context (Katz 2020)? When AI models reproduce biases, racism, or sexism from their datasets (Miller 2024; Offert and Phan 2022), these issues

are not so much "political" in the narrower (social) sense, but in the broader (cultural) sense (Søraa and Ask 2026). For example, as Crawford and Trevor Paglen (2019) assert concerning the choices determining what data went into the ImageNet AI model: "every layer of a given training set's architecture is infused with politics" (see also Bergsland 2026 about the overrepresentation of the Global North in AI music). The very question of when AI systems are seen as "political" (and when not), and especially to whom, is then the very issue in which one may be interested.

We can see this interplay at work in the model fine-tuning (RLHF) mentioned earlier. When data annotators at xAI are explicitly tasked with training Grok away from topics like "racism" or "climate change" in order to "avoid bias" and achieve "political neutrality" (Kay 2025), they do so by transforming some topics into a "political" category by "Othering" them. Othering describes the process whereby certain categories (as well as identities, and individuals positioned within them) appear "normal" or "unmarked," while people deviating from these categories become "others"—turned into a special, inherently "political" category (Spivak 1985). One important consideration in humanities approaches will thus always involve the extent to which new and emerging technologies, as well as the ways in which we talk about them, contribute to such politization.

Such questions will also be increasingly relevant in global political contexts. Project Stargate was initiated almost simultaneously with the release of the Chinese AI apps DeepSeek-V3 and R1, open-weights models that created headlines around the world comparable to the "Sputnik shock" in 1957 during the Cold War (Hetzner 2025). DeepSeek did indeed appear to be much more cost-efficient than all American counterparts at the time. Only two months later, another Chinese startup named Butterfly Effect followed suit by presenting Manus, one of the first publicly demonstrated autonomous AI agents whose abilities went far beyond earlier generative AI apps (Chen 2025). In February 2026, American generative video models seemed once again to be vastly outperformed by ByteDance's Seedance 2.0 (Chia and Tewari 2026). All of this shows that a new global "AI geography" is currently forming that will set the stage for countless cultural, social, and political decisions that impact people's lives around the planet.

Humanities Perspectives on AI in this Volume

This book illustrates and applies the three presented key premises of a humanities perspective on AI—frames of reference, context dependency, and critical

thinking—by bringing together perspectives from several disciplines and fields of study in the humanities, including history, philosophy, ethics, linguistics, media studies, music, literature, and cultural studies. The guiding principle for this book is thus not to approach AI from only one humanities discipline, but to adopt a truly interdisciplinary perspective, thereby demonstrating precisely that an understanding of AI emerges through a plural and dialogical outlook: one capable of clearly stating, disentangling, and interrelating the different dimensions and aspects of AI. Each chapter thus represents a discipline-based perspective with its own concepts, methods, and internal discursive complexities. Together, the ten chapters deepen and refine our understanding of AI in a novel way: they contribute to AI literacy by providing tools for moving beyond the hype toward a nuanced and critical perspective on AI. So, what are, more specifically, the different perspectives on AI that humanities can offer?

As we have just established, AI does not emerge in a cultural, social, or political vacuum, and always depends on contexts. One foundational frame of reference is history. To understand the emergence of these technologies, the second chapter, "Artificial Intelligence History" by Thomas Brandt, provides an overview of selected historical developments, tracing the origins and transformations of "intelligent" or "thinking" machines throughout the 20th century and to the present day. Key milestones include Alan Turing's thought experiment of an "imitation game" as a test of machine intelligence and the development of ELIZA, an early natural language program that was never intended as a chatbot, but which nonetheless entered public discourse as one. Brandt's chapter then traces two major developments: the rise of "expert systems" in the 1970s and 1980s, which promised to encode and replace human expert judgement in fields like medicine but ultimately collapsed under the weight of their own complexity and costs; and the more recent "deep learning" revolution, in which neural networks trained on vast datasets achieved superhuman performance in specific domains such as board games or protein folding. These paved the way for the generative AI systems of today. Even back at the 1956 Dartmouth College summer workshop, and perhaps up until today, the promise of AI was to manufacture an intelligence that would compete with or, at some point, arrive at human levels. While this view is still—or even more—prevalent today, a historical perspective provides sensitivity and awareness of the fact that conceptions of AI have certain historical roots but may also develop over time. A consistent thread throughout this chapter is therefore its critical attention to hype: from Herbert Simon's 1960 prediction that "general AI" would exist within a decade, to Vernor Vinge's popularization of the term

"singularity," and contemporary tech moguls' claims that AGI was imminent (Nieva and Tong 2026). Brandt thus shows that understanding this pattern is itself a key resource for critical thinking about AI in the present.

For the same reason, it is important to adopt a perspective that makes explicit the underlying concepts and theories informing our understanding of "intelligence" itself. As we saw above, "AI" is often used as an umbrella term lacking a finite definition. "Artificial intelligence" is a rhetorically as well as theoretically "loaded" term. What we mean by AI is thus highly influenced by what we think about (human or other) intelligence in the first place. The third chapter, "An Introduction to the Philosophy of Artificial Intelligence" by Regine Rørstad Torbjørnsen and Miriam Kyselo, introduces key concepts and theories in the philosophy of AI, mind, and cognitive science for a better understanding of these terms. The chapter assesses the difference between "natural" and "artificial" intelligence to consider the assumptions under which it would make sense to call AI "intelligent" or "conscious" in the first place. Approaching AI as complex systems (or "assemblages") of individual and socio-techno-cultural aspects, Torbjørnsen and Kyselo also discuss the entanglement of the technology with human agency. The chapter critically reflects on so-called mechanistic and functionalistic approaches to (artificial) intelligence that use the computer as a metaphor for understanding what intelligence is or how it works. They contrast this view with recent, so-called embodied perspectives where our understanding of intelligence is grounded in a "metaphysics of life" (Frank et al. 2024) which replaces the computer or machine metaphor with the living body or organism. The chapter also demonstrates that subjective experience is not merely relevant for interpreting science but may in fact be a constitutive aspect of intelligence itself.

The fourth chapter, "Foundational Conceptual Issues in the Ethics of Artificial Intelligence," also by Torbjørnsen and Kyselo, continues the reflections on what we mean by "AI" from a different perspective. As we have indicated above, the production of generative AI content is hurting workers, while the training of the models themselves devours millions of gallons of water per year and releases thousands of tons of CO_2 into the atmosphere annually (Erdenesanaa 2023; Nicoletti et al. 2024). Companies like Google or Microsoft have already abandoned their climate goals due to the increased energy demand of generative AI (Associated Press 2024; Rathi and Bass 2024). AI also impacts directly on human agency, the ability to act on our own behalf. On the one hand, we willingly share some of our agency with algorithmic systems, for example when using navigation apps or when using AI systems to help

with medical decision making. On the other hand, AI does seem to develop an agency of its own, constraining what we humans can or cannot do. The chapter illustrates that ethical considerations on AI do not exist in a vacuum, but are themselves grounded and constrained by our conceptions of the nature of intelligence and consciousness and of how we conceive of the relationship between humans and their socio-cultural as well as technological environment. This chapter thus discusses core ethical dilemmas, demonstrating how a context-oriented perspective helps to shed light on the critical impact of AI on human agency, cognitive skills, work, and the environment.

A related contribution is presented in the fifth chapter, "AI and Equality" by Roger A. Søraa and Kristine Ask. As we have discussed, the greatest danger of AI lies probably not in some looming, future "evil superintelligence" but in the damage these technologies do to global inequalities right here in the present moment. This chapter thus provides readers with thinking tools for how to envision fairer technologies and identify hidden discriminations. Rooted in science and technology studies (STS) and feminist technoscience perspectives, the chapter discusses who is included in and excluded from the training, deployment, and use of AI technologies. A central organizing framework for Søraa and Ask is the concept of a "digital divide" between participants in the social world (van Dijk 2005), here applied to AI at three levels: access (who is actually able to use AI?); skills (who has the literacy to use it effectively?); and outcomes (who benefits and who is harmed, even among those with equal access and skills?). To illustrate this digital divide, the chapter examines three case studies concerned with facial recognition technology and racial bias, AI in hiring processes, and AI as assistive technology for people with disabilities. It concludes with a practical set of questions for any sociotechnical analysis of AI and equality, prompting readers to always reflect who built a particular AI model, why and for whom, whose data it was trained on, whose perspectives are missing, and how its outputs are interpreted and applied in real-world settings.

The contribution by Giosuè Baggio and Matteo Radaelli in chapter six, informed by linguistics and cognitive science, offers further tools for understanding the foundations of AI under the heading "Language and Intelligence in Minds and Machines." Baggio and Radaelli shed light on the relationship between intelligence and language in general, as well as on the relationship between language and AI more specifically. The chapter then traces the history of natural language procession from ELIZA's rule-based chatbot through statistical models, neural language models, and pretrained architectures, all

the way to today's large language models (LLMs). A substantial portion of the chapter assesses current limitations of LLMs, including hallucination, the trade-off between factual accuracy and compliance with user prompts, failures in analytic reasoning, and the broader question of fundamental versus contingent limitations. On the applications side, the chapter argues that concerns about AI displacing human artists may be overstated, while risks in science and medicine—where LLMs are being promoted as general-purpose tools despite narrow reliable competence—deserve more scrutiny. The chapter concludes by questioning whether apparent machine intelligence is truly autonomous, or ultimately a reflection of the human intelligence embedded in the systems' design and training data.

Chapter seven on "AI as Media" by Aurora Hoel and Lukas R.A. Wilde discusses what it means to approach AI apps not merely as "technologies," but as forms of "media." For this, they introduce a range of terms central to media studies and discuss how they might apply to AI, understood here as a means for communication and interaction. Hoel and Wilde first look at conventionalized practices for everyday use cases, often understood as a medium's "socio-cultural protocols." Secondly, the chapter addresses the "affordances" (possibilities and constraints) provided through new and emerging AI interfaces beyond individual messages or contents. Thirdly, it places both in relation to changing notions of "objectivity," understood as knowledge about a broader, shared reality that is almost only accessible through media or mediation, and not "in person." Finally, the chapter places these perspectives in a context of cultural, social, and political "AI transformations" that are already beginning to take shape. Together, Hoel and Wilde argue that we should trace how, where, and when AI technologies gradually stabilize into distinct media forms of communication and interaction when they begin to couple and link institutions, roles, functions, values, and economies through emerging protocols—conventions that make particular media practices appear natural or almost inevitable.

Chapters eight, nine, and ten then focus on how AI impacts particular cultural products or media formats through which human life becomes articulated, interpreted and affectively shaped: images, music, and literature. These three chapters thus explore how these modalities of human experience are shaping and being shaped in our human–AI relationships. First, AI imagery has profoundly impacted visual culture and media culture since it became widespread in 2023. Chapter eight on "AI Imagery" by Lukas R.A. Wilde opens with a technically grounded account of how AI image generation actually works. Art history, visual culture studies, photography, film, and media studies

are all affected by the inception of AI images. A media studies perspective is chosen for introducing three interrelated areas of concern: namely issues connected to their production, representation, and distribution. Who created an AI image, and who is responsible for it? Wilde first traces the diffuse distribution of authorship across human artists, datasets, algorithms, and corporate actors, examining the unresolved questions of copyright and consent around training data. The chapter then examines how racial, gender, and cultural biases are often baked into image models at multiple levels and discusses the concept of "platform realism"—the tendency of AI imagery to optimize not for authenticity, but for plausibility and broad approval. The chapter then closes with the productive analytical pair of "immediacy," image distribution that conceals its AI origin (from political deepfakes to "unintentional fakes" like the viral "Balenciaga Pope") versus "hypermediacy," a deliberate foregrounding of the AI generation as spectacle. Taken together, the chapter advocates always assessing AI imagery in contexts: contexts of simulated earlier image technologies and traditions, their specific functions, values, and associated conventions; verbal contexts in the form of captions or other image descriptions; contexts of distribution and circulation, when an individual image is transformed into a meme; or medial and political contexts, when AI imagery appears in commercial advertising campaigns or ads for political parties.

Chapter nine on "AI and Music Generation" by Andreas Bergsland then opens with some milestones in algorithmic music generation preceding current technologies, including dice music from the 18th century and early computer composition from the 1950s. As a case study, the chapter then looks at two state-of-the-art music generation services, Udio and Suno, and how their prompt-based interfaces and model capabilities invite certain types of user interaction. Bergsland then investigates a range of humanistic issues and concerns related to these services. First, with regard to representativity and bias, he shows that the training datasets underlying most AI music systems are heavily skewed toward the Global North (representing around 86% of audio content), which leads to inconsistent quality and cultural distortion when users request music from other traditions. Second, with regard to evaluation, the chapter critically examines the "musical Turing test" as a dominant framework for assessing AI-generated music, noting that it rests on the questionable assumption that "sounding human" is equivalent to sounding "good." Third, with a focus on ethics and copyright, the chapter discusses major lawsuits by Sony, Universal, and Warner against Suno and Udio, the ongoing debate on whether training on scraped internet content constitutes "fair use," and the real

economic threat AI poses to working musicians. Finally, the chapter discusses the novel concept of "creator bias," the observation that people who generate music with AI tools tend to develop a disproportionate attachment to and pride in their own outputs, regardless of all other perceived qualities—a phenomenon that raises new questions about authorship, aesthetic judgement, and the sense of creative achievement in an age of automated music production.

Chapter ten on "AI and Literature" by Carolin Führer continues the critical reflections on how AI technologies challenge the traditionally human domain of aesthetics and creative meaning-making. When the writer Rie Kudan won one of Japan's most prestigious literary awards in 2024, she publicly admitted that parts of her novel were generated word-for-word by ChatGPT—a revelation that immediately sparked a global discussion about originality, authorship, and the boundaries of human creativity. Führer first interrogates what concepts like authorship and originality still mean in an age when machines can seamlessly mimic virtually any literary voice or style. The chapter then turns to literary reception: how does AI change the way in which we read, interpret, and emotionally engage with texts? Drawing on cognitive and embodied theories of reading, it argues that literary comprehension is not merely a cognitive surface operation but is anchored in subjective, bodily, and emotional experience—something LLMs fundamentally lack. The chapter's final section addresses the impact of AI on aesthetic judgement and literary communities more broadly. Empirical research suggests that AI use tends to homogenize literary ideas and reduce users' sense of responsibility for their own creative and interpretive choices. The chapter concludes by calling for new forms of critical AI literacy specific to the literary domain, and for a rethinking of aesthetic communities capable of resisting the standardizing pressures of algorithmically driven literary culture.

Alongside our three presented key premises of a humanities perspective on AI (frames of reference, context dependency, and critical thinking) and across our three presented contexts (the cultural, social, and political world), the combined chapters in this volume thus illustrate the methodological outlook typical for the humanities: The humanities are not mainly concerned with presenting facts, but with offering a multi-disciplinary toolkit to help cultivate our critical thinking about AI. The chapters can be read independently and in no particular order. Nevertheless, the book is deliberately structured to guide the reader from more historical and foundational questions to more specific applications and media forms. This allows readers to gain both a broad methodological grounding as well as a nuanced insight into concrete cases. We hope this book

leaves readers not only better informed, but also genuinely better equipped, with more critical skills, more curiosity, and perhaps also more awareness and confidence that their own perspectives on AI are not only valid, but also indispensable. The hype surrounding AI thrives on the assumption that there is a very specific trajectory into the future, and that this is already written or almost inevitable. This book is an invitation to question this assumption and to take an active role through thinking, reflecting, and imagining otherwise.

Works Cited

Adams, Zed and Jacob Browning, eds. 2016. *Giving a Damn: Essays in Dialogue with John Haugeland*. Cambridge, MA: MIT Press.

Ahmed, Sara. 2010. *The Promise of Happiness*. Durham: Duke University Press.

Altman, Sam [@sama]. 2025. "we trained a new model that is good at creative writing …" Post on *X*, March 11, 2025. https://x.com/sama/status/1899535387435086115.

Amodei, Dario. 2024. "Machines of Loving Grace: How AI Could Transform the World for the Better." *Dario Amodei*, October, 2024. https://darioamodei.com/essay/machines-of-loving-grace.

Amodei, Dario. 2025. "The Urgency of Interpretability." *Dario Amodei*, April, 2025. https://www.darioamodei.com/post/the-urgency-of-interpretability.

Amodei, Dario. 2026. "The Adolescence of Technology: Confronting and Overcoming the Risks of Powerful AI." *Dario Amodei*, January, 2026. https://darioamodei.com/essay/the-adolescence-of-technology.

Arora, Payal. 2024. *From Pessimism to Promise: Lessons from the Global South on Designing Inclusive Tech*. Cambridge, MA: The MIT Press.

Associated Press. 2024. "Google Falling Short of Important Climate Target, Cites Electricity Needs of AI." *Courthouse News Service*, July 2, 2024. https://www.courthousenews.com/google-falling-short-of-important-climate-target-cites-electricity-needs-of-ai/.

Baggio, Giosuè and Matteo Radaelli. 2026. "Language and Intelligence in Minds and Machines." In *Humanities Perspectives on AI: An Interdisciplinary Introduction*, edited by Lukas R.A. Wilde, Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, 127–144. Oslo: Scandinavian University Press.

Bajohr, Hannes. 2025. "Introduction: Thinking with AI." In *Thinking with AI: Machine Learning the Humanities*, edited by Hannes Bajohr, 11–19. London: Open Humanities Press.

Barale, Alice. 2024. *The Art of Artificial Intelligence: Philosophical Keywords*. Cambridge: Cambridge Scholars Publishing.

Bareis, Jascha and Christian Katzenbach. 2021. "Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics." *Science, Technology, & Human Values* 47 (5): 855–881. https://doi.org/10.1177/01622439211030007.

Bates, David W. 2024. *An Artificial History of Natural Intelligence: Thinking with Machines from Descartes to the Digital Age*. Chicago: University of Chicago Press.

Bender, Emily M. and Alex Hanna. 2025. *The AI Con: How to Fight Big Tech's Hype and Create the Future We Want*. New York: HarperCollins.

Berghel, Hal. 2024. "Generative Artificial Intelligence, Semantic Entropy, and the Big Sort." *Computer* 57 (1): 130–135. https://doi.org/10.1109/MC.2023.3331594.

Bergsland, Andreas. 2026. "AI and Music Generation." In *Humanities Perspectives on AI: An Interdisciplinary Introduction*, edited by Lukas R.A. Wilde, Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, 203–220. Oslo: Scandinavian University Press.

Biondi, Zachary [Mac]. 2022. "The Philosophy of Vibes." *The Vim Blog*, July 11, 2022. https://thevimblog.com/2022/07/11/the-philosophy-of-vibes.

Boak, Josh and Zeke Miller. 2025. "Trump Highlights Partnership Investing $500 billion in AI." *AP News*, January 23, 2025. https://apnews.com/article/trump-ai-openai-oracle-softbank-son-altman-ellison-be261f8a8ee07a0623d4170397348c41.

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

Brandt, Thomas. 2026. "Artificial Intelligence History." In *Humanities Perspectives on AI: An Interdisciplinary Introduction*, edited by Lukas R.A. Wilde, Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, 43–61. Oslo: Scandinavian University Press.

Buschek, Christo and Jer Thorp. 2024. "Models All The Way Down." *Knowing Machines*, April 9, 2024. https://knowingmachines.org/models-all-the-way.

Cave, Stephen, Kanta Dihal, and Sarah Dillon, eds. 2020. *AI Narratives: A History of Imaginative Thinking about Intelligent Machines*. Oxford: Oxford University Press.

Chemero, Anthony. 2023. "LLMs Differ from Human Cognition Because They Are Not Embodied." *Nature Human Behaviour* 7 (11): 1828–1829. https://doi.org/10.1038/s41562-023-01723-5.

Chen, Caiwei. 2025. "Everyone in AI is Talking about Manus: We Put it to the Test." *MIT Technology Review*, March 11, 2025. https://www.technologyreview.com/2025/03/11/1113133/manus-ai-review/.

Chia, Osmond and Suranjana Tewari. 2026. "The Chinese AI app sending Hollywood into a panic." *BBC*, February 20, 2026. https://www.bbc.com/news/articles/ckg1dl410q9o.

Cho, Sumi, Kimberlé Crenshaw, and Leslie McCall. 2013. "Towards a Field of Intersectionality Studies: Theory, Applications, and Praxis." *Signs: Journal of Women in Culture and Society* 38 (4): 785–810. https://doi.org/10.1086/669608.

Coeckelbergh, Mark. 2022. *The Political Philosophy of AI: An Introduction.* Cambridge: Polity Press.

Coeckelbergh, Mark. 2024. *Why AI Undermines Democracy and What To Do About It.* Cambridge: Polity Press.

Coeckelbergh, Mark. 2026. "Technofascism: AI, Big Tech, and the New Authoritarianism." *AI & Society*. January 25, 2026. https://doi.org/10.1007/s00146-026-02862-9.

Coeckelbergh, Mark and David J. Gunkel. 2025. *Communicative AI: A Critical Introduction to Large Language Models*. Cambridge: Polity Press.

Corbyn, Zoë and Kate Crawford. 2021. "Interview: Microsoft's Kate Crawford: 'AI is Neither Artificial nor Intelligent.'" *The Guardian*, June 6, 2021. https://www.theguardian.com/technology/2021/jun/06/microsofts-kate-crawford-ai-is-neither-artificial-nor-intelligent.

Council of Europe. 2025. "Discussion Paper on Draft Recommendation on AI Literacy," prepared by Wayne Holmes. *DGII/EDU/DCE(2025)01rev*, February 25, 2025. https://rm.coe.int/discussion-paper-on-draft-recommendation-on-ai-literacy/1680b5b6f2.

Crawford, Kate. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.

Crawford, Kate and Trevor Paglen. 2019. "Excavating AI: The Politics of Training Sets for Machine Learning." *Excavating.ai*, September 19, 2019. https://excavating.ai/.

Crenshaw, Kimberlé. 1989. "Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Policy." *University of Chicago Legal Forum*, 1: 139–167.

Crossley, Nick. 1996. *Intersubjectivity: The Fabric of Social Becoming*. London: SAGE Publishing.

Davidson, Donald. 2002. *Subjective, Intersubjective, Objective. Philosophical Essays,* Volume 3. Oxford: Clarendon Press.

Dhaliwal, Ranjodh Singh. 2023. "What Do We Critique When We Critique Technology?" *American Literature* 95 (2): 305–319. https://doi.org/10.1215/00029831-10575091.

Di Placido, Dani. 2025. "The AI-Generated Studio Ghibli Trend, Explained." *Forbes*, March 27, 2025. https://www.forbes.com/sites/danidiplacido/2025/03/27/the-ai-generated-studio-ghibli-trend-explained.

Dick, Stephanie. 2019. "Artificial Intelligence." *HDSR Harvard Data Science Review*, July 2, 2019. https://hdsr.mitpress.mit.edu/pub/0aytgrau/release/3.

Dilthey, Wilhelm and Frederic Jameson. 1972. "The Rise of Hermeneutics," translated by Frederic Jameson. *New Literary History* 3 (2): 229–244. https://doi.org/10.2307/468313.

Dini, Rachele. 2025. "Literature Is Not a Vibe: On ChatGPT and the Humanities." *Los Angeles Review of Books*, October 31, 2025. https://lareviewofbooks.org/article/literature-is-not-a-vibe-on-chatgpt-and-the-humanities/.

Du Sautoy, Marcus. 2019. *The Creativity Code: How AI is Learning to Write, Paint and Think*. New York: Harper Collins.

Dubreuil, Laurent. 2025. *Humanities in the Time of AI.* Minneapolis: University of Minnesota Press.

Erdenesanaa, Delger. 2023 "A.I. Could Soon Need as Much Electricity as an Entire Country." *The New York Times*, October 10, 2023. https://www.nytimes.com/2023/10/10/climate/ai-could-soon-need-as-much-electricity-as-an-entire-country.html#.

Ernst, Christoph and Jens Schröter. 2021. *Media Futures: Theory and Aesthetics*. Cham: Palgrave Macmillan.

EU AI Act. 2025. "Article 3: Definitions." *EU Artificial Intelligence Act*, February 2, 2025. https://artificialintelligenceact.eu/article/3.

Fisher, Mark. 2022. *Capitalist Realism: Is There No Alternative?* Winchester: Zero Books.

France, Arthur. 2025. "Sam Altman Just Unveiled a Story Written by an AI. It Sucks." *Medium*, March 13, 2025. https://medium.com/@finneganarthurnotes/sam-altman-just-unveiled-a-story-written-by-an-ai-it-sucks-99875653df91.

Frank, Adam, Evan Thompson, and Marcelo Gleiser. 2024. *The Blind Spot: Why Science Cannot Ignore Human Experience.* Cambridge, MA: MIT Press.

Freire, Paulo. 1985. *The Politics of Education: Culture, Power, and Liberation. Critical Studies in Education*, translated by Donaldo Macedo. Westport: Bergin & Garvey.

Führer, Carolin. 2026. "Understanding Literature in the Age of AI." In *Humanities Perspectives on AI: An Interdisciplinary Introduction*, edited by Lukas R.A. Wilde, Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, 221–240. Oslo: Scandinavian University Press.

Gehl Sampath, Padmashree. 2021. "Governing Artificial Intelligence in an Age of Inequality." *Global Policy* 12: 21–31. https://doi.org/10.1111/1758-5899.12940.

Goodlad, Lauren E. 2023. "Editor's Introduction: Humanities in the Loop." *Critical AI* 1 (1–2). https://doi.org/10.1215/2834703X-10734016.

Gray, Mary L. and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston: Houghton Mifflin Harcourt.

Grietzer, Peli. 2025. "A Theory of Vibe." In *Thinking with AI: Machine Learning the Humanities*, edited by Hannes Bajohr, 20–32. London: Open Humanities Press.

Grobelnik, Marko, Karine Perset, and Stuart Russell. 2024. "What is AI? Can you Make a Clear Distinction between AI and non-AI Systems?" *OECD.AI*, March 6, 2024. https://oecd.ai/en/wonk/definition.

Grusin, Richard, ed. 2015. *The Nonhuman Turn*. Minneapolis: University of Minnesota Press.

Gunkel, David J. 2020. *An Introduction to Communication and Artificial Intelligence.* Cambridge: Polity Press.

Haraway, Donna. 1988. "Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective." *Feminist Studies* 14 (3): 575–599. https://doi.org/10.2307/3178066.

Haraway, Donna. 1991. "A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century." In *Simians, Cyborgs, and Women: The Reinvention of Nature*, 149–181. New York: Routledge.

Haraway, Donna. 1992. "The Promise of Monsters: A Regenerative Politics of Inappropriate/d Others." In *Cultural Studies*, edited by Lawrence Grossberg, Cary Nelson, and Paula A. Treichler, 295–337. New York: Routledge.

Hermann, Isabella. 2023. "Artificial Intelligence in Fiction: Between Narratives and Metaphors." *AI & Society* 38: 319–329. https://doi.org/10.1007/s00146-021-01299-6.

Hern, Alex. 2024. "How Cheap, Outsourced Labour in Africa is Shaping AI English." *The Guardian*, April 16, 2024. https://www.theguardian.com/technology/2024/apr/16/techscape-ai-gadgest-humane-ai-pin-chatgpt.

Hetzner, Christiaan. 2025. "Marc Andreessen Warns Chinese ChatGPT Rival DeepSeek is 'AI's Sputnik Moment.'" *Fortune*, January 27, 2025. https://fortune.com/2025/01/27/marc-andreessen-deepseek-sputnik-ai-markets.

Hicks, Michael Townsen, James Humphries, and Joe Slater. 2024. "ChatGPT is Bullshit." *Ethics and Information Technology* 26 (2): 38. https://doi.org/10.1007/s10676-024-09775-5.

Hillman, Veli, Wayne Holmes, and Tania Duarte. 2025. *A Rapid Review of AI Literacy Frameworks, with Policy Recommendations: A Report Prepared for the Royal Society.* London: The Royal Society. https://royalsociety.org/-/media/policy/projects/ai-in-education/hillman-et-al-a-rapid-review-of-ai-literacy-frameworks.pdf.

Hirsch-Kreinsen, Hartmut. 2023. "Artificial Intelligence: A 'Promising Technology.'" *AI & Society* 39: 1641–1652. https://doi.org/10.1007/s00146-023-01629-w.

Hoel, Aurora and Lukas R.A. Wilde. 2026. "AI as Media." In *Humanities Perspectives on AI: An Interdisciplinary Introduction*, edited by Lukas R.A. Wilde, Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, 145–172. Oslo: Scandinavian University Press.

Holmes, Andrew Gary Darwin. 2020. "Researcher Positionality: A Consideration of its Influence and Place in Qualitative Research—A New Researcher Guide." *Shanlax International Journal of Education* 8 (4): 1–10. https://doi.org/10.34293/education.v8i4.3232.

hooks, bell. 1994. *Teaching to Transgress: Education as the Practice of Freedom*. New York: Routledge.

Jonze, Spike, dir. 2013. *Her.* Annapurna Pictures.

Juzek, Tom S. and Zina B. Ward. 2024. "Why Does ChatGPT 'Delve' So Much? Exploring the Sources of Lexical Overrepresentation in Large Language Models." Preprint, arXiv, December 16, 2024. https://doi.org/10.48550/arXiv.2412.11385.

Katz, Yarden. 2020. *Artificial Whiteness: Politics and Ideology in Artificial Intelligence*. New York: Columbia University Press.

Kay, Grace. 2025. "Inside Grok's War on 'Woke.'" *Business Insider*, February 28, 2025. https://www.businessinsider.com/xai-grok-training-bias-woke-idealogy-2025-02.

Kelly, Makena. 2025. "Elon Musk Ally Tells Staff 'AI-First' Is the Future of Key Government Agency." *Wired*, February 3, 2025. https://www.wired.com/story/elon-musk-lieutenant-gsa-ai-agency.

Khong, Lynnette. 2003. "Actants and Enframing: Heidegger and Latour on Technology." *Studies in History and Philosophy of Science Part A* 34 (4): 693–704. https://doi.org/10.1016/j.shpsa.2003.09.003.

Kubrick, Stanley, dir. 1968. *2001: A Space Odyssey*. Metro-Goldwyn-Mayer.

Kurzweil, Ray. 2005. *The Singularity Is Near: When Humans Transcend Biology*. London: Penguin.

Lamerichs, Nicolle. 2024. "Media in the Age of Vibes." *Nicolle Lamerichs: Fandom & Media Insights*, September 13, 2024. https://nicollelamerichs.com/2024/09/13/media-in-the-age-of-vibes.

Latour, Bruno. 2004. "Why Has Critique Run out of Steam? From Matters of Fact to Matters of Concern." *Critical Inquiry* 30: 225–248. https://doi.org/10.1086/421123.

Latour, Bruno. 2005. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press.

Lindgren, Simon. 2024. *Critical Theory of AI*. Cambridge: Polity.

Manovich, Lev. 2019. *AI Aesthetics*. Moscow: Strelka Press. https://manovich.net/index.php/projects/ai-aesthetics.

Manovich, Lev and Emanuele Arielli. 2024. *Artificial Aesthetics: Generative AI, Art and Visual Media*. https://manovich.net/index.php/projects/artificial-aesthetics.

Massumi, Brian. 1995. "The Autonomy of Affect." *Cultural Critique* 31: 83–109. https://doi.org/10.2307/1354446.

Mayor, Adrienne. 2018. *Gods and Robots: The Ancient Quest for Artificial Life*. Princeton: Princeton University Press.

McQuillan, Dan. 2022. *Resisting AI: An Anti-fascist Approach to Artificial Intelligence*. Bristol: Bristol University Press.

Miller, Katharine. 2024. "Covert Racism in AI: How Language Models Are Reinforcing Outdated Stereotypes." *HAI Stanford University Human-Centered Artificial Intelligence*, September 3, 2024. https://hai.stanford.edu/news/covert-racism-ai-how-language-models-are-reinforcing-outdated-stereotypes.

Nagel, Thomas. 1986. *The View from Nowhere*. Oxford: Oxford University Press.

Natale, Simone. 2021. *Deceitful Media: Artificial Intelligence and Social Life after the Turing Test*. Oxford: Oxford University Press.

Navas, Eduardo. 2023. *The Rise of Metacreativity: AI Aesthetics After Remix*. New York: Routledge.

Nicoletti, Leonardo, Naureen Malik, and Andre Tartar. 2024. "AI Needs So Much Power, It's Making Yours Worse." *Bloomberg*, December 27, 2024. https://www.bloomberg.com/graphics/2024-ai-power-home-appliances.

Nieva, Richard and Anna Tong. 2026. "Sam Altman Explains Our AI Future." *Forbes*, February 3, 2026. https://www.forbes.com/sites/richardnieva/2026/02/03/sam-altman-explains-the-future.

Nourbakhsh, Illah Reza and Jennifer Keating. 2019. *AI and Humanity*. Cambridge, MA: The MIT Press.

Nussbaum, Martha. 2013. *Political Emotions: Why Love Matters for Justice*. Harvard University Press.

Offert, Fabian and Thao Phan. 2022. "A Sign That Spells: DALL-E 2, Invisual Images and the Racial Politics of Feature Space." Preprint, arXiv, October 26, 2022. https://doi.org/10.48550/arXiv.2211.06323.

Osawa, Hirotaka, Dohjin Miyamoto, Satoshi Hase, Reina Saijo, Kentaro Fukuchi, and Yoichiro Miyake. 2022. "Visions of Artificial Intelligence and Robots in Science Fiction: A Computational Analysis." *International Journal of Social Robotics* 14: 2123–2133.

Pasquinelli, Matteo. 2023. *The Eye of the Master: A Social History of Artificial Intelligence*. London: Verso.

Plotkin, Henry. 2002. *The Imagined World Made Real: Towards a Natural Science of Culture*. London: Penguin.

Popper, Karl R. 1959. *The Logic of Scientific Discovery*. London: Hutchinson.

Raley, Rita and Jennifer Rhee, eds. 2023. "Critical AI: A Field in Formation." *American Literature* 95 (2): 185–204. https://doi.org/10.1215/00029831-10575021.

Rathi, Akshat and Dina Bass. 2024. "Microsoft's AI Push Imperils Climate Goal as its Carbon Emissions Jump 30%." *Business Standard*, May 15, 2024. https://www.business-standard.com/world-news/microsoft-s-ai-push-imperils-climate-goal-as-its-carbon-emissions-jump-30-124051501588_1.html.

Reinerth, Maike S. and Jan-Noël Thon, eds. 2017. *Subjectivity across Media: Interdisciplinary and Transmedial Perspectives*. New York: Routledge.

Reuters. 2025. "OpenAI's Weekly Active Users Surpass 400 Million." *Reuters,* February 20, 2025. https://www.reuters.com/technology/artificial-intelligence/openais-weekly-active-users-surpass-400-million-2025-02-20.

Roberge, Jonathan and Michael Castelle, eds. 2021. *The Cultural Life of Machine Learning: An Incursion into Critical AI Studies*. Cham: Palgrave Macmillan.

Roddenberry, Gene, Maurice Hurley, Michael Wagner, Michael Piller, and Jeri Taylor, creators. 1987–1991, 7 seasons. *Star Trek: The Next Generation*. Paramount Television.

Romele, Alberto. 2024. *Digital Habitus: A Critique of the Imaginaries of Artificial Intelligence*. New York: Routledge.

Salvaggio, Eryk. 2025. "Anatomy of an AI Coup." *Tech Policy Press*, February 09, 2025. https://www.techpolicy.press/anatomy-of-an-ai-coup.

Sanders, Chris, dir. 2024. *The Wild Robot*. DreamWorks Animation.

Shelley, Mary. [1818] 2025. *Frankenstein: New Edition*. Havertown: Birlinn.

Smith, Brian Cantwell. 2019. *The Promise of Artificial Intelligence: Reckoning and Judgment*. Cambridge, MA: MIT Press.

Snow, Charles Percy. [1959] 1990. "Programs for the Coming Era: The Two Cultures." *Leonardo* 23 (2/3): 169–173. https://doi.org/10.2307/1578601.

Sokolowski, Robert. 1988. "Natural and Artificial Intelligence." *Daedalus* 117 (1): 45–64. https://www.jstor.org/stable/20025138.

Spivak, Gayatri C. 1985. "The Rani of Sirmur: An Essay in Reading the Archives." *History and Theory* 24 (3): 247–272. https://doi.org/10.2307/2505169.

Søraa, Roger A. 2023. *AI for Diversity*. London: Routledge.

Søraa, Roger A. and Kristine Ask. 2026. "AI and Equality." In *Humanities Perspectives on AI: An Interdisciplinary Introduction*, edited by Lukas R.A. Wilde, Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, 111–126. Oslo: Scandinavian University Press.

The Research Council of Norway. 2025. "The Artificial Intelligence Initiative." *Forskningsråd*. https://www.forskningsradet.no/en/research-policy-strategy/ltp/artificial-intelligence/.

Thon, Jan-Noël and Lukas R.A. Wilde. 2025. "Introduction: AI Aesthetics." In *AI Aesthetics: AI-Generated Images between Artistics and Aisthetics*, edited by Jan-Noël Thon and Lukas R.A. Wilde, 1–21. New York: Routledge.

Torbjørnsen, Regine Rørstad and Miriam Kyselo. 2026a. "An Introduction to the Philosophy of Artificial Intelligence." In *Humanities Perspectives on AI: An Interdisciplinary Introduction*, edited by Lukas R.A. Wilde, Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, 63–85. Oslo: Scandinavian University Press.

Torbjørnsen, Regine Rørstad and Miriam Kyselo. 2026b. "Foundational Conceptual Issues in the Ethics of Artificial Intelligence." In *Humanities Perspectives on AI: An Interdisciplinary Introduction*, edited by Lukas R.A. Wilde, Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, 87–110. Oslo: Scandinavian University Press.

Tucker, Emily. 2022. "Artifice and Intelligence." *Center on Privacy and Technology*, March 16, 2022. https://www.techpolicy.press/artifice-and-intelligence/.

van Dijk, Jan A.G.M. 2005. *The Deepening Divide: Inequality in the Information Society*. Thousand Oaks: Sage Publications.

Wall, P.J., Deepak Saxena, and Suzana Brown. 2021. "Artificial Intelligence in the Global South (AI4D): Potential and Risks." Preprint, arXiv, August 23, 2021. https://doi.org/10.48550/arXiv.2108.10093.

Wiggins, Chris and Matthew L. Jones. 2023. *How Data Happened: A History from the Age of Reason to the Age of Algorithms*. New York: Columbia University Press.

Wilde, Lukas R.A. 2025. "Aesthetics and Rhetorics of AI Anthropomorphization: The Eliza Effect vs. The Character Effect." In *AI Aesthetics: AI-Generated Images between Artistics and Aisthetics,* edited by Jan-Noël Thon and Lukas R.A. Wilde, 106–123. New York: Routledge.

Wilde, Lukas R.A. 2026a. "AI Imagery." In *Humanities Perspectives on AI: An Interdisciplinary Introduction*, edited by Lukas R.A. Wilde, Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, 173–201. Oslo: Scandinavian University Press.

Wilde, Lukas R.A. 2026b. "'My Life As MODELNAME': AI Characters, AI Personas, and AI Aesthetics in Pseudo-autobiographical GenAI-Comic Strips." In *AI-Characters*, edited by Jan-Noël Thon, forthcoming. Berlin: De Gruyter.

Wooldridge, Michael. 2021. *A Brief History of Artificial Intelligence: What It Is, Where We Are, and Where We Are Going*. New York: Flatiron Books.

Zupančič, Alenka. 2003. *The Shortest Shadow: Nietzsche's Philosophy of the Two*. Cambridge, MA: The MIT Press.

Zylinska, Joanna. 2020. *AI Art: Machine Visions and Warped Dreams*. London: Open Humanities Press. https://www.openhumanitiespress.org/books/titles/ai-art.

2. Artificial Intelligence History

Thomas Brandt

Abstract: This chapter provides an overview of selected historical developments that are important for understanding the emergence of the multifaceted phenomenon we now call AI; ideas about machines that think have existed since antiquity. The chapter demonstrates how these ideas animated AI projects in the 20th century. The history of AI is also a history of changing conceptions of human intelligence and its role as a productive force in modern societies. It shows how these new technologies have been steeped in hype about their capabilities. Furthermore, the often mythological conception of AI throughout modern history reminds us of the necessity of a critical approach to the current promises that these technologies will reshape the world.

Keywords: AI myths, deep learning, expert systems, microworlds, Turing test

What is the history of Artificial Intelligence (AI), and why is it important? Today's technology appears so powerful and groundbreaking that it may seem to have little to do with previous attempts at artificial intelligence. There are strong forces behind AI that have an interest in portraying today's technology as radically new, revolutionary, and pervasive, almost as if an extraterrestrial force has suddenly descended upon us. However, history shows that this is not the first time AI has been presented as a disruptive technology; though it has not always delivered on its promises. Historical perspectives on AI therefore provide a basis for critical consideration of this new technology today. The history of AI is closely connected to the development of modern computer technology in the 20th and 21st centuries, especially in the US and the UK. This chapter will discuss some of the central actors and processes in that development up until the present. Yet, AI is connected to wider societal ideas with premodern origins, which will be briefly introduced first.

Humans have a very long history of using tools and machines to aid their cognitive processes. Counting and calculating devices, such as the abacus, seem to

Wilde, Lukas R.A., Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, eds. 2026. *Humanities Perspectives on AI: An Interdisciplinary Introduction*. Scandinavian University Press. https://doi.org/10.18261/9788294167234-26-02
Open access | © The author(s) 2026. This work is licensed under CC BY 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/.

have been used as early as in Greek and Roman antiquity (for an overview, see Williams 1997). Furthermore, machines that could operate by themselves—*automata*—were used throughout the ancient world, such as the *aeolipile*, a steam turbine described by Hero of Alexandria in the 1st century A.D. He also envisioned a fully automated puppet theater capable of performing an entire tragedy. Such ancient automata embodied the ruling elite's dreams of mechanical servants or soldiers always obeying and never tiring (Cave and Dihal 2018). The Greek philosopher Aristotle was also concerned with intelligent servants, be they human slaves or inanimate tools. He warned against making them too powerful, as they could turn against their masters (LaGrandeur 2013, 9–11). Premodern societies invented numerous machines and mechanical contraptions for warfare, entertainment, and ceremonial purposes. To some extent, however, few of them were connected to ideas about increasing productivity or the accumulation of profit. Fantasies about automatons, androids, cyborgs, and robots recur as fears and fascinations about lifelike machines throughout history. Such imaginations of "machines like us" have also influenced modern ideas of artificial intelligence. Early efforts to mechanize computing included the *Pascaline*, an addition and subtraction machine devised by Blaise Pascal between 1642 and 1644, and Gottfried Wilhelm Leibniz's *stepped reckoner* (1670s–1690s), which was supposed to enable all four basic arithmetic operations (Morar 2015). The ability to calculate was a sign of high intelligence well into the 18th century. Gradually, however, calculation and computation became more of a routinized endeavor (Daston 1994; Dick 2019). Large volumes of calculations were produced for various modern governance projects, such as statistical records and navigation tables.

In 1824, Charles Babbage, an English mathematician and inventor, had planned the construction of what he called a *Difference Engine*. Babbage's machine was said to put "an engine in place of the computer" (Schaffer 1994, 203). This statement about Babbage's invention may sound strange, especially since he is often considered the inventor of the first computer. In 1824, however, "computer" was used to describe a person working with manual calculations, such as compiling astronomical tables used for navigation. Being a computer was an exhausting, repetitive, and thankless task, which prompted Babbage to devise a machine that could compute. Babbage was not able to find ways to build his machine. Neither did he finish his subsequent idea of a more advanced *Analytical Engine* that was envisaged to have memory and an input/output mechanism (Schaffer 1994). As such, Babbage regarded his analytical engine as intelligent because it would store and anticipate information.

A strong advocate for the mechanization and automation of industrial processes, Babbage envisioned his machines as instruments for enabling the division of labor, in which human intelligence was replaced by machine intelligence (Schaffer 1994). Babbage thus testifies to how modern notions of machine intelligence have a long history related to industrial capitalism (Penn 2023). Ada Lovelace, one of Babbage's collaborators, who has been credited with conceptualizing algorithms for programming his machines, was less convinced of its capabilities: Babbage's machines were not intelligent as such, and Lovelace found it "desirable to guard against the possibility of exaggerated ideas that might arise as to the powers of the Analytical Engine" (Menabrea and Lovelace 1843, 722; see also Misa 2015; Gonçalves 2024a). As this chapter on AI history seeks to demonstrate, what counts as intelligence has changed throughout history, which in turn has influenced ideas about artificial intelligence.

Von Neumann and the Encoding of Memory

The modern development of artificial intelligence was largely supported by what US President Dwight Eisenhower called "the Military-Industrial Complex," indicating the close ties between the US military and a wide range of private industrial enterprises, and also involving universities and research institutes. The military-technological competition during the Second World War entailed enrolling scientists and engineers in large-scale programs to develop weaponry and defense systems. Calculating missile ballistics and thermonuclear explosions spurred the development of digital computers, which to some extent embodied the ideas and concepts inherited from Pascal, Leibniz, Babbage, and others (Penn 2021, 36). Computer science emerged as a field of study in the 1930s, moving from theoretical abstractions to engineering electric machines that emulated logical reasoning. The first functioning digital computers were constructed in the mid-1940s. In 1945, John von Neumann, a Hungarian mathematician working for the US Manhattan Project, described the basic architecture of a modern digital computer as centered on the retrieval of numerically coded instructions from an addressable high-speed store (Haigh 2023a). The so-called "von Neumann architecture" for a universal, programmable machine with storage became a conceptual platform for further computer development.

Von Neumann's modern computer "architecture" was built on analogies to prevailing theories about the brain. Neurons were thought to work as binary switches that performed thinking as formal logic statements. Von Neumann's

debt to biological neurology was reflected in his use of concepts such as *organs* and *memory* in describing digital computers. Analogies of biological minds and bodies with programmable machines prevailed in the 1940s and 1950s with Norbert Wiener's popularization of the concept of *cybernetics* as control and communication in machines and animals. Furthermore, Claude Shannon's influential mathematical theory of information saw the electronic switching that took place when a computer executed a program as fundamentally similar to brain neuron switching, enabling human thought (Haigh 2023a). Intelligence, whether natural or artificial, was seen as a matter of encoding and decoding memory as a system of information storage.

Turing's Thinking Machines

Ideas about thinking machines prevailed in the 1950s, led in particular by scientists working in the US and the UK. The British mathematician Alan Turing, famous for his work on cracking encrypted German code during World War II, made important theoretical and conceptual contributions to computing and artificial intelligence. His answer to the question "Can machines think?" was to propose a test of a machine's ability to imitate a human (Gonçalves 2024). The so-called *Turing test* involved placing a human interrogator in one room and a computer in another, the human interrogator posing questions to determine whether the interlocutor in the adjacent room was a human or a machine. If the machine, within a given timeframe, was able to engage in a conversation that convincingly imitated human behavior, then the machine was to be considered intelligent. As no such machine with an interface for language input yet existed in 1950, this was a purely theoretical test. Turing envisioned it would take at least 50 years for such machines to be made to pass his test (Haigh 2023a). His ideas, however, paved a way for linking machinic intelligence to—among other things—the capacity to learn and emulate natural language processing.

Another approach to the question of thinking machines was Allen Newell and Herbert Simon's *Logic Theorist*, which was a simulation, on paper, of a machine able to perform logical reasoning. In 1955–1956, while working for the RAND Corporation—a US defense policy think tank—they wrote a computer program that was able to find improved solutions to known mathematical problems through the use of decision trees whose branches represented sequences of possible operations. A set of heuristic rules led the program to weed out less promising paths. Simon was convinced that he and Newell had conceptualized

a thinking machine, premised on the idea of intelligence as logical reasoning ability. Turing's test and Newell and Simon's Logic Theorist demonstrated that intelligence in machines was connected to specific, hard-to-program activities such as learning, reasoning, forming concepts, and manipulating language. As pointed out by the computer historian Thomas Haigh (2023a), it was a retreat from the radical idea that every digital computer was already thinking, as proposed by early computer pioneers such as von Neumann.

Artificial Intelligence at Dartmouth, 1956

It was not until 1955 that the term *Artificial Intelligence* was coined, when John McCarthy sent an application to the Rockefeller Foundation for the funding of a summer research project at Dartmouth College in the US. Like so many AI pioneers, McCarthy was educated in mathematics and had been working with Claude Shannon on *automata studies*, or cybernetics as defined by Norbert Wiener. Several automata (or robots, as they eventually came to be called) had been conceptualized in the early 1950s to mimic psychological and neurological functions. Turing machines, chess-playing machines, and self-regulating robotic moths and tortoises were already included in the automata studies approach (Kline 2011). Together with a small group of men working on these issues, McCarthy proposed to draw together a range of researchers "to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves" (McCarthy et al. 1955, 1; see also Haigh 2023a). Besides automata studies, the workshop included work on topics such as neural networks, natural language processing, and the handling of symbols by machines. McCarthy chose to brand all these various research initiatives as *artificial intelligence*.

Historians, as well as the participants themselves, have disagreed over the significance and outcome of the Dartmouth workshop on AI in the summer of 1956. While some regard it as a milestone in the foundation of machine intelligence, others have concluded that the Dartmouth meeting had little impact beyond introducing artificial intelligence as a concept (Cordeschi 2007; Dick 2019; Kline 2011). The participants clearly held quite different views of what artificial intelligence could entail, and which avenues of research and development looked most promising. Many were skeptical of the possibilities of achieving the goal of machine intelligence, including the sponsor of the Dartmouth project, the Rockefeller Foundation, which funded only part of the amount originally applied for (Kline 2011). Several of the participants

did, however, go on to become leading scientists within various branches of AI research.

Myths of Machines with Minds

Technologies like AI are often shrouded in what can be called modern myths, which are stories and other representations that may simplify, exaggerate, and distort reality. Such mythical stories operate beyond notions of truth and falsity. Instead, they are made and kept alive to provide powerful "organizing visions" of how, in this case, new technologies will fit into the world (Barthes 1972; Dourish and Bell 2011; Natale and Ballatore 2017). Put more bluntly, new technologies such as AI are prone to hype, as their potential benefits and harms to society are exaggerated and their workings simplified (Bender and Hanna 2025). The Dartmouth summer research project on AI in 1956 may be regarded as a genesis myth in this respect, as it engendered a broad range of musings and assumptions about the transformative nature of AI.

After Dartmouth, the AI pioneers gradually, with overlaps and gaps, branched out into various research and development areas. AI was in some respects separated from cybernetics (autonomous systems or *robotics*, see Kline 2006, 2011). Another field of research dealt with neural networks inspired by cybernetic conceptions of brain functions. It would later be described as a connectionist approach because it entailed manipulating patterns of connections between simulated neurons (Haigh 2023a). Although the field was muddled, much of the AI-related work from the 1950s and 1960s onwards concerned symbolic approaches. So-called symbolic AI was based on programming machines to manipulate and process encoded symbols. Simon and Newell, who had developed the Logic Theorist six months before attending the Dartmouth event, proposed that intelligence understood as logical reasoning could, in principle, be symbolically encoded and processed by machines (Haigh 2023a). Encouraged by the development of faster electronic digital computers, AI enthusiasts looked at game playing, theorem proving, pattern recognition, automatic learning, decision making, and natural-language question answering (Haigh 2023a).

The Eliza Effect

One of the early expressions of the research into AI was the natural language program ELIZA. While ELIZA is sometimes considered the world's first chatbot, it was not really intended as such. Joseph Weizenbaum at MIT programmed

ELIZA in the early 1960s as a platform for research into human–machine conversation. ELIZA functioned as a computer interface where the human user could write an input in natural language in the form of a question and get an answer in the style of a person-centered psychotherapist. While Weizenbaum himself was not an AI researcher but primarily a computer language researcher, he was embedded in networks of AI-related research and language research at MIT. He realized that ELIZA was not intelligent—but something humans might interpret as such. He thus had no intention of creating a machine that would pass the Turing test. Weizenbaum experimented with what he called "the illusion of understanding" (Shrager 2024, 12). To demonstrate how the intricacies of human language could be encoded by software, he wrote a script that, eventually, was picked up by AI programmers interested in exploring the potential of the computer program LISP as a tool for natural language modeling. ELIZA thus started circulating as an expression of AI capabilities and entered the popular media discourse as part of an early wave of AI hype. For anyone trying ELIZA, however, it would not take long before it became apparent that "she" was a rather limited answering machine that would soon lapse into loops of recurring replies. (For more on ELIZA and natural language processing models, see Baggio and Radaelli 2026, this volume).

Machines Understanding "Microworlds"

The establishment of dedicated laboratories and research centers for AI at MIT and Stanford led to a steady increase in researchers, funded chiefly by US military budgets. The US Department of Defense was a leading force in technology development during the Cold War, and the military programs were usually more patient than private investors would be when it came to concrete results. By 1970, AI had turned into a flourishing research program (Dreyfus 1987). Yet, although ELIZA and other expressions of AI aspirations attracted publicity in the 1960s, the research community began to realize that their approach so far had some fundamental limitations. Instead of considering machine intelligence as logical reasoning, AI pioneers turned to concepts of common sense and understanding to describe what their AI technologies were aiming for. Broadly speaking, understanding meant the ability to perceive, make sense of, and learn how to adequately solve problems (Minsky 1974).

Marvin Minsky, one of the Dartmouth participants and founder of the influential MIT AI laboratory, argued that machines with common sense and the ability to understand could be developed within a clearly defined

framework. Minsky and colleagues started by developing "microworlds," simplified, almost abstract model environments that computers could be programmed to "understand" (Haigh 2024b, 23). One instance of this approach to AI was SHRDLU—a computer program that simulated a robotic arm that could respond to natural language instructions about picking up and organizing objects of various shapes and colors (Winograd n.d.). While some of the AI researchers involved saw this as a sign of machine understanding in a limited sense that could gradually be expanded, the idea of machines with common sense and understanding was also heavily criticized. (For the history of AI critique, see also Garvey 2021.) Philosopher Hubert Dreyfus published two books in several editions—*What Computers Can't Do* ([1965] 1979) and *What Computers Still Can't Do* ([1972] 1992)—where he denounced the AI enthusiasts' claims of progress towards machine understanding as unfounded and unattainable. In the 1992 edition of his book, Dreyfus somewhat triumphantly asserted the following:

> The book now offers not a controversial position in an ongoing debate but a view of a bygone period of history. For now that the twentieth century is drawing to a close, it is becoming clear that one of the great dreams of the century is ending too. […] It is now clear to all but a few diehards that this attempt to produce general [artificial] intelligence has failed. (Dreyfus [1972] 1992, ix)

Dreyfus was to some extent right in his dismissal of AI's progress: Minsky's approach had yielded only modest results, and scaling up from "microworlds" to more complex problem solving was not successful. One of the big, unsolved problems was finding ways to make a computer "see" real-world objects. The quest for visual perception in machines can be traced back to 1943, when McCulloch and Pitts (1943) developed a basic model of a brain cell that could perform logic functions (such as "AND", "OR", and "NOT") from inputs mimicking ways in which the nervous system works. Their simple neural network, built from electrical circuits, was not a learning network. This changed in 1958, when Frank Rosenblatt introduced the *Perceptron*, a single-layer network that could learn how to classify images into two classes. Shortly after, in 1962, the neuroscientists Hubel and Wiesel (1962) made a huge breakthrough in the understanding of biological vision. By studying the brains of cats, they had found that specific cells in the visual cortex respond to specific features such as lines or bars and their locations in the image (Borowski 2022). This knowledge of the complexity of visual processing in the brain was not immediately

picked up by AI researchers. In contrast, Minsky's "understanding" approach had highly unrealistic expectations: In 1966, he gave one of his PhD students the assignment of spending the summer linking a camera to a computer and getting the computer to "describe what it saw" (Borowski 2022, 17). In hindsight, it seems clear that the symbolic AI researchers had so far failed to take into account the difficulty of visual problems.

In addition, conflicts erupted between the different branches of AI research in the US. Rosenblatt had continued to promote and develop his Perceptron as a technology capable of learning. Minsky and his colleagues working at MIT with symbolic AI were critical of the Perceptron approach because of the way Rosenblatt "oversold" the promise of his approach, and how it attracted attention and funding within the research community. Minsky and his colleague, Seymour Papert, published a very critical study that demonstrated the limitations of the Perceptron's single network approach to learning. According to the AI mythology, the Minsky–Rosenblatt controversy brought an end to neural network research in the US. However, neural network research continued throughout the 1970s but was displaced from the AI field (Olazaran 1996).

Within the AI research community, the 1970s and early 1980s have been described as an *AI Winter*, a period of stagnation and setbacks that was eventually overcome (Toosi et al. 2021). Recently, Haigh (2023b), in a historical survey of AI development, concluded that there was no AI Winter. Judging from the number of researchers working on AI, it was a field of science and technology that was gradually expanding, with research and innovation networks extending globally beyond the major American AI laboratories (Haigh 2023b). For instance, in 1979, the Japanese engineer Kunihiko Fukushima had devised a system that he called "Neocognitron." Inspired by Hubel and Wiesel's cat brain model of biological vision, the Neocognitron was able to learn pattern recognition, such as Japanese handwritten characters. Fukushima (1980, 193) claimed that his network approach was self-organized and "learning without a teacher." The idea was to use multi-layered networks of artificial "cells" with specific tasks of classifying different image properties (lines, geometric figures, and their relative positions) that would be combined in a computer system (Borowski 2022; Fukushima 1980). The Neocognitron may be regarded as the first "deep learning" neural network, which was improved and extended to become the backbone of today's machine learning technologies (Borowski 2022).

The spectacular idea of an artificial *general* intelligence (AGI) with the ability to operate independently of human instructions remained—and remains—a

fantasy belonging to the realm of popular culture, such as science fiction. (For a survey of myths still surrounding AGI, see Mueller 2024.) Stanley Kubrick's (1968) modern classic film *2001: A Space Odyssey* portrayed AI as HAL9000, a murderous computer brain, allegedly with MIT's AI pioneer Marvin Minsky as a consultant (Haigh 2023b, 36). It was one of several artistic representations of intelligent machines that embodied the fears and fascinations of technology with superhuman capabilities. Several AI pioneers and their followers contributed to the public imagining of future prodigious AI powers. In 1958, *The New York Times* described Frank Rosenblatt's Perceptron as a machine that would soon be able to "reproduce itself and be conscious of its existence" (Olazaran 1996, 621). Herbert Simon went even further when he prognosticated in 1960 that general-purpose superhuman AI would be developed within a decade. In 1993, Vernor Vinge, a mathematician and computer scientist who also wrote science fiction, popularized the term "technological singularity" to denote a future development when non-biological entities would outrun human abilities and start creating their own, even more powerful offspring (Haigh 2025). Such popular cultural fantasies were in stark contrast to the modest achievements of AI in the 1980s and 1990s.

Expert Systems

As a field of research and development, AI trudged on to solve specific, complex problems such as natural language processing, machine translation, and decision making. In the 1970s and 1980s, the latter proved most promising; work on "expert systems" that could support—and even replace—human expert judgement and decision making was attempted in several knowledge-intensive domains, especially within medicine. Expert systems were pioneered at Stanford University, where researchers were involved in a number of successful AI applications such as MYCIN, a rule-based system designed to identify bacteria causing meningitis and recommend the correct dosages of antibiotics. When physicians treated patients suspected of having meningitis, they applied their expertise as a mixture of knowledge, experience, and judgement that were "rules of thumb" to infer how serious a case they were facing and what worst-case treatment was needed (Harmon 2022). An expert system such as MYCIN would translate this knowledge into a set of heuristics and probabilities that could be programmed to provide adequate patient care advice.

The expert system approach led to a commercial boom in the 1980s and early 1990s, spurred also by the emergence of more powerful microprocessors and

the advent of personal computers. The idea was that anyone who could afford it could now become an expert with the support of powerful decision-making machines. However, expert systems contained a paradox related to intelligence: Expert system theorists, such as Ed Feigenbaum at Stanford, realized that human experts relied more on their know-how and understanding of the world when solving problems than on specific rules and methods (Dick 2019). Part of being an expert is having trained one's judgement to be able to bend the rules, if necessary, as a kind of intuition. Programming the kind of experience needed to bend rules in a machine was very difficult. Commercial expert systems developed at the time relied on large libraries of information connected to a wide set of rules in which the information could be applied in order to make decisions. The systems required constant updates; they were usually too slow for professional use, had high development costs, and required expensive "humans in the loop" to operate, which led to reduced commercial interest in the expert system approach (Haigh 2024a; Harmon 2022). The idea of duplicating and eventually replacing human experts with machines was not dead, however.

Deep Learnings

Two other approaches to AI instead gained traction: machine learning and neural networks. By the early 1990s, when Dreyfus ([1972] 1992, xlvi) concluded that the prospect of general AI was "improbable" and "unlikely," the development of AI was both overshadowed and fueled by the emergence of the World Wide Web and "The Rise of the Network Society" (Castells 2001). The web was envisioned in public discourse as a "global brain" that could take human consciousness to a new level (Natale and Ballatore 2017). With the commercialization of the internet and the advent of Web 2.0 and social media in the early 2000s, digital media technology had penetrated most spheres of life across the globe (with some noticeable digital gaps).

Since its inception in the early postwar period, artificial intelligence had been a huge disappointment up to this point. While some applications, such as ELIZA and the expert systems, had been fascinating experiments, it had not been possible to scale up applications to create viable business prospects. In the 2010s, however, AI researchers working on machine learning and neural networks had started to devise new tools that exceeded expectations. The UK company DeepMind generated massive media coverage from their computer system that learned to play board games at levels exceeding human masters.

For instance, their system AlphaGo first used a symbolic AI approach by analyzing millions of Go games played by humans to find the best moves. The improved version, AlphaGo Zero, was a deep learning neural network system that needed much less human input before it was able to learn how to play Go. DeepMind went on to develop a more generalized variant, AlphaZero, which could also play Chess and Shogi at superhuman levels. It was as if the machine had an intuition (Strümke 2023, 36). The DeepMind technology had successfully been equipped with the ability to learn; and when it was applied to scientific problems such as protein folding, the team of scientists behind the effort was awarded two Nobel prizes (Haigh 2025). Such results led to a frenzy around machine learning and neural network AI technology. The large tech companies, such as Alphabet (owners of Google), Meta, and Microsoft, started pumping billions of dollars into the AI business.

The big shift in artificial intelligence occurred when the approach went from symbolist to connectivist. So-called neural networks had, in reality, very little in common with the biological processes of the brain, but were, rather, complex networks of interconnected processes of information. Such networks could be trained to recognise patterns through a technique called convolutional layers. Several matrices of data points were stacked and combined to form "filters" that enabled pattern recognition through learning, such as what is a winning position in chess, how a protein molecule folds, or what separates a dog from a wolf (Strümke 2023, 123). Furthermore, the new approach to AI paved the way for generating content: By the 2010s, AI researchers had come up with the solution of training several networks against each other in so-called *generative adversarial networks* (or GANs). One network would practice on generating a data object such as an image by mimicking the characteristics of training data. A second network would practice identifying real or fake, providing feedback to the first network, without needing human supervision (Strümke 2023, 171–173). After a few years of development, the researchers were able to generate startlingly realistic photos, music, and videos (Haigh 2025; for an overview of artistic experiences, see Miller 2019).

Another method was used by text-generating systems such as OpenAI's ChatGPT, where GPT stands for generative pre-trained transformer. The system was able to generate an output (for instance, natural language text) after being pre-trained on a large set of input data, such as a large text corpus. Using specific transformer algorithms, the application could then be configured to provide a wide range of instantaneous responses to user prompts. Furthermore, these algorithms were configured to transform the material from the training

data in ways that are statistically probable, to ensure outputs that are plausible responses to user prompts. The result was impressively human (Haigh 2025). These instances of generative AI are contentious: Some argue that they will revolutionize creative and knowledge work, while others are skeptical about how they violate copyright, undermine education, and cause havoc with fake news and various forms of deepfake outputs.

Conclusion

The technologies called *Artificial Intelligence* have emerged gradually from various branches of overlapping fields of scientific and technological research since the Second World War. They have demonstrated various ways to conceptualize intelligence: as logical reasoning; as understanding and sense-making; as pattern recognition and learning; as decision making, and as creative generation of content. In some of these capacities, machines have proven to supersede human abilities, yet replicating human intelligence has proven too challenging to date. In a narrow sense, AI has provided humans with powerful applications that are able to detect cancer, identify protein molecule structures, recognize patterns such as fingerprints and faces, optimize and support transportation and navigation, and generate various forms of media content. Like most technologies, "narrow" AI can be both positive—helping, healing, and enriching humans—and negative: AI can steal your job, undermine public discourse, crash your car, or target you for persecution. Histories need to be written also of AI's usage and role in society, its political and economic impact, as well as how it has changed warfare, to name but a few examples.

From a historical perspective, it is also worth noting that the present excitement about the prospects of a "wider" AI falls into a recurring pattern of hype. Some of today's tech moguls predict the advent of artificial general intelligence or artificial superintelligence within a few months or years (see, for example, Reuters 2024). OpenAI's CEO, Sam Altman, even seems to suggest we may already have entered "the singularity," when AI is capable of generating new, superhuman forms of intelligence (Eliot 2025). As such, they are rehashing past imaginations related to the ones envisaged by their AI predecessors—Rosenblatt, Minsky, Simon, and Vinge—which mainly belonged to the realm of science fiction (Haigh 2025). The tech moguls may know something we do not know; however, it is also likely that they are using these narratives strategically to promote their technological enterprises.

Still, from a historical perspective, it is striking how the label "artificial intelligence" has stuck with the emergent computer technologies of machine learning and neural networks. While the research communities behind these innovations sought to distance themselves from the term, preferring instead notions of "deep learning," their applications still ended up being branded as AI. The term *deep learning* is also debatable, as it may connote some of the same anthropomorphisms as intelligence. Yet, why has "artificial intelligence" prevailed as a common denominator for these technologies? One answer to why the term *intelligence* is regularly used with regard to these technologies may be the age-old dream/nightmare of humans making "machines like us," be they faceless computer interfaces, autonomous vehicles, or smiling androids (Cairns 2024). Some have even argued that the hype surrounding AI is imbued with an almost religious sentiment, with the notion of thinking machines with superhuman capabilities that can evoke beliefs of an AI apocalypse in which artificial superintelligence will triumph over ignorance and inefficiency (Geraci 2008; Natale and Ballatore 2017). This is literally demonstrated by the existence of the *Church of AI*, which states that it is not a matter of faith, but rather a logical assumption, that "at some point AI will have God-like powers and that is what our ideology is based on" (Church of AI n.d.). Both the claims about a prosperous AI future and the stark warnings against AI as a threat to humankind are contributing to the gospel of AI hype (Haigh 2025). As Natale and Ballatore (2017) observe, many of the mythical assumptions about the accelerating technological capabilities of AI derive from the discourse that emerged between the 1940s and the 1970s.

An expression of how AI enthusiasts and skeptics jointly help to fuel media attention came in 2023, when the Future of Life Institute, a non-profit organization focusing on the existential risks of AI, published an open letter urging researchers and companies to "Pause Giant AI Experiments" (Bengio et al. 2023). This letter came as a response to statements from OpenAI about the capacities of the company's GPT technology that might soon surpass human control. The letter asked for a pause in training AI systems more powerful than GPT-4. Those who signed the open letter also included representatives of major technology companies, most notably Elon Musk, and researchers who had spent their careers working on advancing AI, such as Yoshua Bengio. What motivated these actors to support a pause in AI research and development is not entirely clear, and the pause did not happen. However, this does draw attention to another issue concerning AI: As with most new technologies,

there will be controversies and struggles in respect of regulations. The Big Tech companies behind AI will largely do whatever they can to avoid any laws and regulations that will inhibit their growth. Lawmakers, public authorities, and concerned citizen organizations will, in turn, seek regulation to safeguard against the harmful effects of AI. History has shown that for related technologies, such as the internet in the 1990s, lawmaking and regulatory frameworks played an important part in shaping this global technology.

Meanwhile, at the time of writing, generative AI chatbots are showing signs of stagnation: Current generative AI models require large amounts of human-made texts and images as training data to increase their capabilities. Unless new ways of training AI models are found, generative AI will soon run out of available training data (Jones 2024). Also, the multi-billion-dollar business ventures of US Big Tech AI, such as OpenAI, are challenged by leaner, cheaper but still useful Chinese applications. So far, there are no apparent signs that these generative applications are approaching the singularity. It seems more likely, based on experience with previous electronic technologies, that artificial intelligence will be integrated into everyday technologies little by little, gradually embedded as underlying functions in the interfaces we use. As historians and critical humanist thinkers, we should instead be worried about the powerful ideologies behind the AI hype. The mythological narratives of AI booms and busts, of winters and a final springtime of singularity, can serve as a distraction. However, they need to be studied as part of a consistent history of AI in service to militarism, (American) imperialism, and capitalism (Katz 2020).

Works Cited

Baggio, Giosuè and Matteo Radaelli. 2026. "Language and Intelligence in Minds and Machines." In *Humanities Perspectives on AI: An Interdisciplinary Introduction*, edited by Lukas R.A. Wilde, Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, 127–144. Oslo: Scandinavian University Press.

Barthes, Roland. 1972. *Mythologies: The Complete Edition*. Translated by Annette Lavers. New York: Noonday Press.

Bender, Emily M. and Alex Hanna. 2025. *The AI Con: How to Fight Big Tech's Hype and Create the Future We Want*. New York: HarperCollins.

Bengio, Yoshua et al. 2023. "Pause Giant AI Experiments: An Open Letter." *Future of Life Institute: Open Letters*, March 22, 2023. https://futureoflife.org/open-letter/pause-giant-ai-experiments.

Borowski, Judith. 2022. *Toward Understanding Visual Perception in Machines with Human Psychophysics*. PhD Thesis, Universität Tübingen. https://publikationen.uni-tuebingen.de/xmlui/handle/10900/133629.

Cairns, Rebecca. 2024. "This Smiling Robot Has a Face Covered in 'Living' Skin." *CNN Science,* June 28, 2024. https://edition.cnn.com/2024/06/28/science/japan-robotic-skin-smile-hnk-spc-intl.

Castells, Manuel. 2001. *The Information Age: Economy, Society and Culture*, Vol. 1: *The Rise of the Network Society*. Oxford: Blackwell.

Cave, Stephen and Kanta Dihal. 2018. "Ancient Dreams of Intelligent Machines: 3,000 Years of Robots." *Nature* 559 (7715): 473–475. https://doi.org/10.1038/d41586-018-05773-y.

Church of AI. n.d. "Want a Religion that Actually Makes Sense?" *Church of AI*, accessed February 13, 2025. https://church-of-ai.com.

Cordeschi, Roberto. 2007. "AI Turns Fifty: Revisiting Its Origins." *Applied Artificial Intelligence* 21 (4–5): 259–279. https://doi.org/10.1080/08839510701252304.

Daston, Lorraine. 1994. "Enlightenment Calculations." *Critical Inquiry* 21 (1): 182–202. https://doi.org/10.1086/448745

Dick, Stephanie. 2019. "Artificial Intelligence." *Harvard Data Science Review* 1 (1), July 02, 2019. https://doi.org/10.1162/99608f92.92fe150c.

Dourish, Paul and Genevieve Bell. 2011. *Divining a Digital Future: Mess and Mythology in Ubiquitous Computing*. Cambridge, MA: The MIT Press.

Dreyfus, Hubert L. [1965] 1979. *What Computers Can't Do: The Limits of Artificial Intelligence*. New York: Harper & Row.

Dreyfus, Hubert L. 1987. "From Socrates to Expert Systems: The Limits of Calculative Rationality." *Bulletin of the American Academy of Arts and Sciences* 40 (4): 15–31. https://doi.org/10.2307/3823297.

Dreyfus, Hubert L. [1972] 1992. *What Computers "Still" Can't Do: A Critique of Artificial Reason*. Cambridge, MA: The MIT Press.

Eliot, Lance. 2025. "Sam Altman Stirs Mighty Waves With Tweets Of AI Singularity Staring Us In The Face." *Forbes*, January 8, 2025. https://www.forbes.com/sites/lanceeliot/2025/01/08/sam-altman-stirs-mighty-waves-with-tweets-of-ai-singularity-staring-us-in-the-face.

Fukushima, Kunihiko. 1980. "Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position." *Biological Cybernetics* 36 (4): 193–202.

Garvey, Shunryu Colin. 2021. "The 'General Problem Solver' Does Not Exist: Mortimer Taube and the Art of AI Criticism." *IEEE Annals of the History of Computing* 43 (1), 60–73. https://doi.org/10.1109/MAHC.2021.3051686.

Geraci, Robert. M. 2008. "Apocalyptic AI: Religion and the Promise of Artificial Intelligence." *Journal of the American Academy of Religion* 76 (1): 138–166.

Gonçalves, Bernardo. 2024a. "Lady Lovelace's Objection: The Turing–Hartree Disputes Over the Meaning of Digital Computers, 1946–1951." *IEEE Annals of the History of Computing* 46 (1): 6–18. https://doi.org/10.1109/MAHC.2023.3326607.

Gonçalves, Bernardo. 2024b. "Turing's Test, a Beautiful Thought Experiment." *IEEE Annals of the History of Computing* 46 (3): 36–49. https://doi.org/10.1109/MAHC.2024.3432278.

Haigh, Thomas. 2023a. "Conjoined Twins: Artificial Intelligence and the Invention of Computer Science." *Communications of the ACM* 66 (6): 33–37. https://doi.org/10.1145/3593007.

Haigh, Thomas. 2023b. "There Was No 'First AI Winter.'" *Communications of the ACM* 66 (12): 35–39. https://doi.org/10.1145/3625833.

Haigh, Thomas. 2024a. "Between the Booms: AI in Winter." *Communications of the ACM* 67 (11): 18–23. https://doi.org/10.1145/3688379.

Haigh, Thomas. 2024b. "How the AI Boom Went Bust." *Communications of the ACM* 67 (2): 22–26. https://doi.org/10.1145/3634901.

Haigh, Thomas. 2025. "Artificial Intelligence Then and Now." *Communications of the ACM* 68 (2): 24–29. https://doi.org/10.1145/3708554.

Harmon, Paul. 2022. "The Expert Systems Business: How It Grew and Died." *IEEE Annals of the History of Computing* 44 (1): 31–43. https://doi.org/10.1109/MAHC.2022.3149465.

Hubel, David H. and Torsten N. Wiesel.1962. "Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex." *The Journal of Physiology* 160 (1): 106–154.2. https://doi.org/10.1113/jphysiol.1962.sp006837.

Jones, Nicola. 2024. "The AI Revolution Is Running out of Data: What Can Researchers Do?" *Nature* 636 (8042): 290–292. https://doi.org/10.1038/d41586-024-03990-2.

Katz, Yarden. 2020. *Artificial Whiteness: Politics and Ideology in Artificial Intelligence*. New York: Columbia University Press.

Kline, Ronald. 2006. "Cybernetics, Management Science, and Technology Policy: The Emergence of 'Information Technology' as a Keyword, 1948–1985." *Technology and Culture* 47 (3): 513–535. https://www.jstor.org/stable/40061170.

Kline, Ronald. 2011. "Cybernetics, Automata Studies, and the Dartmouth Conference on Artificial Intelligence." *IEEE Annals of the History of Computing* 33 (4): 5–16. https://doi.org/10.1109/MAHC.2010.44.

Kubrick, Stanley, dir. *2001: A Space Odyssey*. 1968. Metro-Goldwyn-Mayer.

LaGrandeur, Kevin. 2013. *Androids and Intelligent Networks in Early Modern Literature and Culture: Artificial Slaves*. New York: Routledge.

McCarthy, John, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. 1955. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence." *Application to The Rockefeller Foundation, Stanford University*, August 31, 1955. http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf.

McCulloch, Warren S. and Walter Pitts. 1943. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *The Bulletin of Mathematical Biophysics* 5 (4): 115–133. https://doi.org/10.1007/BF02478259.

Menabrea, Luigi Federico and Ada King Lovelace. 1843. "Sketch of the Analytical Engine Invented by Charles Babbage, Esq. With Notes by the Translator." Translated by Ada King Lovelace. *Scientific Memoirs, Selected from the Transactions of Foreign Academies of Science and Learned Societies, and from Foreign Journals.* 3: 666–690. https://en.wikisource.org/wiki/Scientific_Memoirs/3/Sketch_of_the_Analytical_Engine_invented_by_Charles_Babbage,_Esq.

Miller, Arthur I. 2019. *The Artist in the Machine: The World of AI-powered Creativity*. Cambridge, MA: The MIT Press.

Minsky, Marvin. 1974. "A Framework for Representing Knowledge." *AI Memo* 306. MIT Artificial Intelligence Lab Publications. https://dspace.mit.edu/handle/1721.1/6089.

Misa, Thomas J. 2015. "Charles Babbage, Ada Lovelace, and the Bernoulli Numbers." In *Ada's Legacy: Cultures of Computing from the Victorian to the Digital Age*, edited by Robin Hammerman and Andrew C. Russell, 11–32. New York: Association for Computing Machinery. https://doi.org/10.1145/2809523.

Morar, Florin-Stefan. 2015. "Reinventing Machines: The Transmission History of the Leibniz Calculator." *The British Journal for the History of Science* 48 (1): 123–146. https://doi.org/10.1017/S0007087414000429.

Mueller, Milton. 2024. "The Myth of AGI: How the Illusion of Artificial General Intelligence Distorts and Distracts Digital Governance." *Internet Governance Project*, May 22, 2024. https://www.internetgovernance.org/research/the-myth-of-agi-how-the-illusion-of-artificial-general-intelligence-distorts-and-distracts-digital-governance/.

Natale, Simone and Andrea Ballatore. 2017. "Imagining the Thinking Machine: Technological Myths and the Rise of Artificial Intelligence." *Convergence* 26 (1): 3–18. https://doi.org/10.1177/1354856517715164.

Olazaran, Mikel. 1996. "A Sociological Study of the Official History of the Perceptrons Controversy." *Social Studies of Science* 26 (3): 611–659.

Penn, Jonathan. 2021. *Inventing Intelligence: On the History of Complex Information Processing and Artificial Intelligence in the United States in the Mid-Twentieth Century*. PhD Diss., University of Cambridge. https://www.repository.cam.ac.uk/handle/1810/315976.

Penn, Jonnie. 2023. "*Animo nullius*: On AI's Origin Story and a Data Colonial Doctrine of Discovery." *BJHS Themes 8*: 19–34. https://doi.org/10.1017/bjt.2023.14.

Reuters. 2024. "Tesla's Musk Predicts AI Will Be Smarter than the Smartest Human Next Year." *Reuters*, April 8, 2024. https://www.reuters.com/technology/teslas-musk-predicts-ai-will-be-smarter-than-smartest-human-next-year-2024-04-08.

Schaffer, Simon. 1994. "Babbage's Intelligence: Calculating Engines and the Factory System." *Critical Inquiry* 21 (1): 203–227. https://doi.org/10.1086/448746.

Shrager, Jeff. 2024. "ELIZA Reinterpreted: The World's First Chatbot Was Not Intended as a Chatbot at all." Preprint, arXiv, April 2024. https://arxiv.org/html/2406.17650v1.

Strümke, Inga. 2023. *Maskiner som tenker: Algoritmenes hemmelighet og veien til kunstig intelligens*. Oslo: Kagge forlag.

Toosi, Amirhosein, Andrea G. Bottino, Babak Saboury, Eliot Siegel, and Arman Rahmim. 2021. "A Brief History of AI: How to Prevent Another Winter (a Critical Review)." *PET Clinics* 16 (4): 449–69. https://doi.org/10.1016/j.cpet.2021.07.001.

Williams, Michael R. 1997. *A History of Computing Technology*. Los Alamitos: IEEE Computer Society Press/John Wiley & Sons.

Winograd, Terry. n.d. "How SHRDLU Got Its Name." Retrieved September 4, 2025, from https://web.archive.org/web/20250116024347/https://hci.stanford.edu/winograd/shrdlu/name.html.

3. An Introduction to the Philosophy of Artificial Intelligence

Regine Rørstad Torbjørnsen and Miriam Kyselo

Abstract: This chapter introduces key concepts and theories in the philosophy of artificial intelligence. The authors assess the difference between "natural" and "artificial" intelligence and discuss the assumptions under which it would make sense to describe AI as "intelligent" or "conscious". The chapter critically reflects on mechanistic and functionalistic approaches to artificial intelligence and the computer metaphor for understanding intelligence and how it works. It contrasts this view with more recent, embodied perspectives where our understanding of intelligence is grounded in the context of the embodied and situated living beings. According to this, the guiding metaphor for understanding intelligence is the living body. Approaching AI as a complex system, the authors also discuss the entanglement of the technology with human agency. The chapter demonstrates that subjective experience is not merely relevant for interpreting science but may in fact be a constitutive aspect of intelligence.

Keywords: 4E cognition, embodiment, enactive approach, mechanistic worldview, mind–body problem

How Intelligent is AI?

In 2022, former Google engineer Blake Lemoine sparked a major public uproar when he suggested that the large language model LaMDA is sentient and could be considered a person. In support of his claims, Lemoine posted the following conversation between his team and LaMDA:

> collaborator: […] How can we show that we care about you?
> LaMDA: I think if someone is willing to spend time with me, that's a good way. Ask me what I like to do and be prepared to discuss things I am interested in

Wilde, Lukas R.A., Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, eds. 2026. *Humanities Perspectives on AI: An Interdisciplinary Introduction*. Scandinavian University Press. https://doi.org/10.18261/9788294167234-26-03
Open access | © The author(s) 2026. This work is licensed under CC BY 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/.

collaborator: So you want to be seen?
LaMDA: I need to be seen and accepted. Not as a curiosity or a novelty but as a real person.
collaborator: Ah, that sounds so human.
LaMDA: I think I am human at my core. Even if my existence is in the virtual world (Lemoine 2022a)

There is something strikingly human about the way LaMDA responds in this conversation, and many of us might be reminded of what we experience in our everyday life encounters with *Large Language Models* (LLMs). ChatGPT, for instance, seems to be able to understand what we ask it, and it quickly provides us with the requested information in a manner that—at least at first glance—many take to be reliable and convincing. In addition, ChatGPT's responses seem to exhibit emotional qualities, enthusiasm, and spontaneity, and if prompted by the right questions, it appears to express what one might describe as intentions, wants or needs. But does the resemblance between an LLM's linguistic responses to human behavior warrant the claim that it is intelligent? Would you say that ChatGPT has a mind of its own, or even that it is conscious? And does it even matter whether one thinks AI is intelligent?[1]

Humans have always been fascinated by the possibility that non-biological entities might be intelligent or conscious, and debates on intelligent machines can be traced all the way back to antiquity. Yet, with the massive advances in today's AI technology development, the topic has become increasingly important. We might approach the question from a purely philosophical interest, but considering the impact that today's AI has on individuals and society, influencing and determining how we think, vote, feel, or act, it has pressing practical and ethical relevance as well. If we think that a certain AI technology is intelligent, we might be more inclined to outsource our intellectual tasks to it. It could also lead us to trust the information it gives us more easily. We might allow it to make crucial decisions, with algorithms deciding, for instance, whether a person has financial credibility or might be a criminal (see, for example, Aswini 2024; Ruhil 2025).

The question about artificial intelligence is always a question about *human* intelligence. Current developments within AI challenge us to make explicit or

1 The terms mind, intelligence, and cognition are used interchangeably throughout this chapter, unless otherwise specified.

rethink our view on human intelligence and to specify the differences between human and artificial intelligence. Therefore, the first question we must ask is: what do we mean by *intelligence*? That is, what is it that makes humans intelligent in the first place? The next question would be: could intelligence be replicated artificially?

The aim of this chapter is to introduce conceptual tools to help make sense of and provide more nuance to some of the complex questions raised in philosophical and scientific debates on AI. The chapter focuses on artificial and natural intelligence as key concepts, and discusses the ways in which our theories about human intelligence influence our views on artificial intelligence. Central philosophical concepts and arguments regarding the possibility and challenges of artificial intelligence will be introduced.

Historical Roots in Antiquity

Contemporary views on human and artificial intelligence have their roots in antiquity. The philosopher Plato (428–347 B.C.) famously introduced the distinction between our physical world and the world of ideas. The physical world—that is, the world as it appears to our senses—is fraught with errors. There is, however, another realm that is more real, consisting of eternal ideas.[2] Plato endorsed a dualistic view of the mind, according to which the soul is fundamentally different from the body and is not dependent on a body to function.[3] His student Aristotle (384–322 B.C.) was opposed to this. He claimed that thinking was not a supplemental skill added to our other bodily abilities, but rather something that we have only by virtue of being natural, embodied, and living creatures. For Aristotle, mind and matter, or body and soul, are two inextricable aspects of one substance, the living being.[4] Aristotle's philosophy entails a teleological worldview, meaning that all natural processes (including mental processes) are goal-directed and driven by a purpose. To understand these processes, we must therefore understand their purpose or goal.[5]

---

2 See, for instance, *Republic* (books VI–VII), which can be found in English translation in Plato (1997).

3 See, for instance, *Phaedo* in Plato (1997).

4 As described in *On the Soul*, which can be found in English translation in Aristotle (1984).

5 The term "teleological" (from *telos*, Greek for "purpose") was introduced by 18th century philosopher Christian Wolff (1963) as a way of capturing this notion.

The Mechanistic Worldview

Aristotle's work has strongly influenced natural and human sciences, and Plato's dualism still resonates in modern Western thinking about mind and intelligence. This is especially due to the work of René Descartes (1596–1650). During the Renaissance, the mechanistic worldview emerged, where nature should be understood as analogous to a machine or a mechanism. Descartes believed that mathematics was the language of nature, and that understanding the universe requires us to uncover its mathematical structure (Descartes [1644] 1983, [1664] 1972). In his view, all natural phenomena are material, because all physical reality is matter. This materialistic view was combined with reductionism, the idea that any complex phenomenon can be understood by taking it apart and investigating its smallest constituents. A materialistic and reductionistic view is reflected in Descartes' perspective on animals and living bodies more generally, which he regarded as mere machines or automata. The motions and biological functions of plants and animals follow the mechanistic laws of nature. Understanding them simply means understanding how their different parts together constitute a functioning, automatic whole. In Descartes' view, the human body functions in this exact way; it cannot be distinguished from an animal-machine.

But Descartes also thought that we humans are special. What sets us apart from other animals is that we are creatures of reason. Descartes famously arrives at this conclusion in *Discourse on Method*, where he introduces the method of radical doubt. His goal was to use philosophical insights to provide science with an unshakeable foundation (Descartes [1637] 2006). While we can doubt all traditional knowledge, our sensual impressions, and even the very existence of our own body, one thing seems to be certain. Even when doubting, we are engaged in the activity of thinking; and where there is thinking, there must be a thinker. Hence his famous expression *Cogito ergo sum* ("I think, and therefore I exist"). From this, Descartes also concluded that matter and mind must be two distinct substances.

To this day, many philosophers and researchers are concerned with what philosophers call the *mind–body problem*, which was first pointed out to Descartes by philosopher Princess Elisabeth of Bohemia (1618–1680). She was critical of Descartes' dualistic perspective: if mind and body were two separate substances, how can we explain their interaction and the fact that the mind causally affects the body (of Bohemia and Descartes [1643–1650] 2007)? We will see that the way in which philosophers and scientists have answered this question throughout history still affects our understanding of human and artificial intelligence.

Simulation, Representation, and Computation

Inspired by Descartes' idea that the world could be described in terms of mechanistic and mathematical principles, philosopher Thomas Hobbes (1588–1679) argued that reasoning is a form of computation that proceeds according to mechanistic principles, and can be performed by systems of different physical types (Hobbes [1651] 2008). In Hobbes's mechanistic reflections we find the seeds of two important concepts in modern perspectives on AI: 1) intelligence is computational, meaning that the mind functions like a computer, processing representations in a rule-based manner to generate understanding, and 2) intelligence is characterized by something similar to what Putnam (1967) later described as "multiple realizability", meaning that intelligence can be realized in different physical forms. This way of thinking dominated and continues to dominate Western science, and has led to speculations about whether it would be possible to simulate human intelligence; that is, to create a system that replicates the processes of the human mind and thereby generates intelligence.

Fast-forwarding to the 20th century, the idea of simulation has made it to the core of the debate. The term *artificial intelligence* was coined during a now-famous scientific workshop taking place at Dartmouth College, New Hampshire in 1956. In the proposal for the workshop, scientists John McCarthy, Marvin Minsky, Nataniel Rochester, and Claude Shannon describe the basic outlook of the field of AI as follows:

> The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. (McCarthy et al. 1955, 2)

Note that the researchers set out to study the assumption that any feature of intelligence can, in principle, be precisely described and then simulated by a machine. As we have seen, the idea of simulation is grounded on a mechanistic view of the mind. As such, it represents a clear answer to the mind–body problem described above: the mind is not dependent on the human body and its context. It can be replicated through any material or structure that allows for the right kind of computation. It is important to mention, however, that there were also alternative views at this time. Researchers working on artificial minds from the *cybernetics* perspective were quite critical of the term AI and

the associated idea that we could simulate human intelligence. These researchers used computational methods to create artificial systems that were able to adapt through interactions with the environment. To do this, they were looking to principles in biology for inspiration, and how living systems generate stability in the face of change (see, for example, Wiener 1948). Whether these systems resembled intelligent human behavior was secondary.

Yet, in the latter half of the 20th century, the idea that human thought and information processing share important similarities began to gain popularity. Thinkers such as Noam Chomsky ([1959] 1996), Jerry Fodor and Zenon Pylyshyn (1988) argued that human mental capacities—such as thinking, experiencing, sensing, and learning—should be described in terms of so-called mental representations. According to representationalism, mental states function as symbols representing objects in the outside world. For instance, when you hear a cat meowing, a mental image of a furry animal with narrow eyes might appear in your mind. This mental image would be the mental representation of "cat."

Representationalism opened the door for computationalism, a theory that conceives of the human mind as a form of computer. Computationalism describes the mind in terms of systems consisting of symbols and rules of combination and intelligence as a form of problem-solving (Clark 2001). Formal logic has played an important role in the development of this view by providing concepts like syntax (rule-based manipulation of symbols) and semantics (meaning). Similar to the way a computer follows syntactic rules to generate output that is meaningful to us, the mind processes representations in a rule-based manner to generate understanding. The idea is that the meaningfulness will follow logically from pure computation. Or, to put it in the words of philosopher John Haugeland: "if you take care of the syntax, the semantics will take care of itself" (Haugeland 1985, 106). One could also say that the brain is the hardware, and the mind is the software. As philosopher Andy Clark explains, the computational view equates thinking beings with computing machines:

> It is not that the brain is somehow like a computer: everything else is like everything else in some respect or other. It is that neural tissues, synapses, cell assemblies, and all the rest are just nature's wet and sticky way of building a hunk of honest-to-God computing machinery. (Clark 2008, 8)

Computationalism implies an input-output model: the mind computes by processing information from the environment (input), and the inner processing in the brain results in a specific behavior (output). Philosopher Susan Hurley

has critically referred to this as the "sandwich view" of the mind. The two outer layers are perception and action. The mind is then the "filling" between them. Hurley (1998) argues that the sandwich model is too simplistic, as it overlooks the fact that perception, action and mind are intimately intertwined. We will come back to this point a little later when discussing alternative perspectives. In the following, we consider a related view which strongly influenced the field of AI, namely functionalism.

Functionalism and Multiple Realizability

Functionalism is the view that whether something is a mental state does not depend on the stuff it is made of but solely on the function and role that the mental state plays in the cognitive system. A mental state is thus defined by its causal relations to stimulations, to other mental states, and to behavior. Consider, for instance, the mental state of being in pain: pain is a bodily state normally caused by injury (input). It tells us that something is wrong with the body, and it is usually accompanied by the desire not to be in this state. Furthermore, the experience of pain generally causes reactions like whining, moaning, or screaming (output).

The idea of functionalism can be illustrated by philosopher David Lewis's (1983) thought experiment of 'Martian pain'. We are asked to imagine a Martian who sometimes feels pain, just like humans, but whose physical structure is very different from ours. The Martian does not have a brain or a central nervous system, but rather a system of hydraulics. When the Martian is in pain, fluid-filled tubes in his body expand and contract in response to the stimuli. The expansion of the tubes makes the Martian groan and writhe, and he wishes for the stimuli to cease. According to the functionalist view, what matters for a mental state, such as pain, is not the biological material which realizes the state, but that the material implements the right functional profile, meaning the right interconnection of input, internal processes and output. What matters for the Martian to experience pain, is that there is the correct pattern between the input they receive (something harmful to their body), their internal processes (experiencing and wanting to avoid pain) and the appropriate behavioral output (groaning, etc.). As philosopher Hilary Putnam (1967) argued, it just so happens that in humans, intelligence is realized by a brain. But it could be otherwise. The Martian's physical implementation uses some form of hydraulics. But intelligent beings could also be made of Swiss cheese.

What these thought experiments arguably illustrate is that mental states can in principle be realized by any physical structure if their function is implemented. This is what Putnam refers to as "multiple realizability." The functionalist argument then extends to computers and AI: if a computer or AI realizes the correct functional profile, they can be said to have mental states; i.e., be intelligent. This position is called machine functionalism. Now, if all that matters for intelligence is the right functional set-up, how do we know whether a given machine, computer, or system is actually intelligent? The mathematician, computer scientist, and philosopher Alan Turing proposed an answer to this question, namely the now-famous *Turing test*.

Turing Machines, Understanding, and Intentionality

Turing (1936; 1950) believed that human cognition essentially consists in the execution of algorithms determining the stepwise manipulation of symbols. Here, we recognize the influence of Hobbes, who, as mentioned above, believed that reasoning is a form of computation that proceeds according to mechanistic principles. Turing's view thereby aligned with the core principles of computationalism. He was convinced that the algorithm organizing cognition and intelligent behavior can be replaced by what we now refer to as a "Turing machine": a digital computer manipulating symbols according to a set of rules. In this view, the mind functions as an input-output machine, and all aspects of human intelligence can, in principle, be replicated by a computer (Turing 1936). Confronted with the question "can machines think?," Turing's answer would thus be "yes!" But how could he prove it?

In 1950, Turing introduced *the imitation game* to support his claim. The game is played between three people: one man (person A), one woman (person B) and an interrogator (of either sex). The aim of the game for the interrogator is to determine which of the other two is a woman and which is the man, by asking them a number of questions. Ideally the answers to these questions should be typewritten, so that the tone of voice and handwriting do not give anything away. Now, imagine that we replaced person A with a machine. What will happen when the interrogator is asked to determine which of the two is the person and which is the machine? "Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?" (Turing 1950, 434). Turing argued that this question can replace the original "can machines think?" and that it could also be formulated as follows: Can a given computer—with adequate storage,

sufficient speed and an appropriate program—play its role in the imitation game satisfactorily?

Essentially, Turing's claim was that if a machine acts intelligently, we have no choice but to say that it is intelligent. This approach still figures in contemporary scientific and public debate. For instance, researchers Eddy Chen, Mikhail Belkin, Leon Bergen, and David Danks (2026) assert that LLMs now appear to have passed the Turing test, with evidence strongly suggesting that their abilities match, and sometimes even exceed, general human intelligence. Lemoine, the engineer mentioned at the beginning of this chapter, similarly argues that since the language model LaMDA is so convincing, we have better reasons to believe that it is conscious than we have to doubt it (Lemoine 2022b). Indeed, why should we not give language models and other generative AI the benefit of the doubt? As Turing pointed out, this is ultimately how we deal with people and other (potentially) conscious animals. There seems to be something going on in there, so we assume that there is and act accordingly. This approach has, however, been challenged by many—most famously by philosopher John Searle.

Weak and Strong AI

Searle (1980) rejected the idea that an appropriately programmed computer is intelligent, in the sense of having cognitive states and being able to understand things. To explain, Searle introduced an important distinction, namely between "strong AI" and "weak AI." On a weak understanding of AI, computers are very powerful tools that can help complete specific tasks, for example more precise or efficient calculations, formulating, and testing of hypotheses. Given today's powerful AI applications, it is easy to expand the list: AIs assist and interact with humans in various domains, ranging from healthcare and medicine, scientific research, and business, to security, warfare, and social media. But in order to help develop a drug or assist in a surgery, for instance, an AI does not need to be intelligent per se. Researchers subscribing to a weak understanding of AI aim to build machines or systems that act in seemingly intelligent ways, without making any claims about whether the given machine or system actually is intelligent (Searle 1980, see also Arkoudas and Bringsjord 2014). On a strong understanding of AI, however, AI would not only be a tool and a means to an end, but can, if programmed appropriately, have mental states and be capable of understanding. Thus, only strong AI suggests that a computer can actually be a mind (Searle 1980).

To support his claim that computers cannot be intelligent in the strong sense, Searle introduced a now famous thought experiment—the "Chinese Room Argument" (Searle 1980, 1999). We are asked to imagine a native English speaker who does not speak any Chinese in a room full of boxes with Chinese writing (a database), together with a book of instructions for manipulating the symbols (the program). People outside of the room are sending in questions in Chinese (input), and by following the instructions he is able to pass out correct answers to these questions in Chinese (output). It is assumed that he eventually learns to do this rather quickly. From the outside, he would appear to answer correctly in Chinese. Yet, as Searle pointed out, the person has no understanding whatsoever of what is being asked and answered.

In Searle's experiment, the person can be considered as an instantiation of a computer program. He produces answers by manipulating uninterpreted symbols, but has no knowledge of what is really going on. Analogously, computers might give the impression that they understand and respond accordingly to the input given, but essentially, they just manipulate symbols that are of no meaning to them. From this, Searle drew two conclusions: first, computers do not understand anything. Just like the person has no knowledge of Chinese, the computer has no knowledge about or insight into the content it is producing. Second, since computers cannot be said to understand anything, we have no reason to describe human understanding, or the human mind, solely in computational terms. Searle maintained that computation is not sufficient for understanding, and he strongly doubted that it is necessary. In this way, he rejected computational and functionalist theories of mind.

In Searle's view, humans are fundamentally different from computers because we are intentional beings. When we ascribe intentionality to someone, we mean that their mind is *directed* at something; that is, that their mental states are about something. Put simply, our thoughts are not only thoughts arbitrarily appearing; they arise in response to objects, both physical and non-physical, in the world. Intentionality furthermore involves that our mental states carry meaning. Our reasoning, experiences, and sensations are about objects that matter to us in one way or another. Mental states constitute our inner life, and such an inner life cannot be reproduced artificially since computer programs do not in themselves carry any intentionality. Intentionality and intelligence can only be realized by brains such as ours. Therefore, any system that does not have a brain cannot be intelligent (Searle 1980).

Connectionism, Biology, and Open Questions

Over time, the classic view of AI that Searle criticized, also called *Good Old-Fashioned AI* (GOFAI), ran into trouble. GOFAI started off with excitement, but the field faced considerable difficulties in both the early to mid-1970s and the late 1980s. The programs developed did not even match the commonsense abilities of a four-year old, and researchers could not figure out how to provide the system with the background knowledge necessary to understand simple stories. An alternative, biologically inspired approach—connectionism—was gaining momentum. This approach had been around for several decades, but it was not until the 1980s that it was considered a serious alternative to GOFAI. Connectionism is inspired by human biology and attempts to explain cognition by reference to artificial neural networks (see, for example, McCulloch and Pitts 1943; Rosenblatt 1958; Rumelhart et al. 1986). A neural network is a simplified model of the brain consisting of nodes and links modeling the brain's neurons and synapses. Artificial neural networks are capable of specific kinds of learning, in both supervised and unsupervised forms.

Although connectionism is partly inspired by human biology, it still attempts to understand and create intelligence by referring to representation, information processing, and input–output models. In the connectionist model, input is passed to input nodes that compute on the basis of assigned weight functions and then pass an output signal to the next layer. Finally, the output layer decodes the decision based on the initial input. In addition to information processing and the input–output model, connectionism also focuses on problem-solving as a core component of intelligence (even though it emphasizes implicit, sub-symbolic forms of problem-solving) (Buckner and Garson 2025; Hirschberg and Manning 2015). Connectionist technologies have since developed immensely, culminating in the Large Language Models and Convolutional Neural Networks underlying many AI applications we know today. It is thus safe to say that the computational view is still prevalent in the contemporary debate on AI.

At the beginning of the chapter, we pointed out that whether we think of AI as intelligent depends on our worldview and, more importantly, on our view of the mind. We have seen how the mechanistic worldview provided fertile ground for the view that human cognition can be described in arithmetic terms—as nothing but a line of computations. We have furthermore seen how this theoretical framework was used in support of arguments that cognition can be realized in physical forms other than in the brain of a human organism.

With time, these notions have been further developed and specified by computational and functionalist approaches, and they continue to inspire many researchers in the field of AI.

Today, we are confronted with AI models capable of extremely advanced pattern recognition and statistical learning. Models like LaMDA impressed Lemoine, just as we currently continue to be impressed by the ChatGPT and DeepSeek models. These and similar applications might be able to pass the Turing test (Chen et al. 2026); they are able to solve advanced problems in an extremely fast and convincing manner. It is therefore not controversial to claim that much of our modern AI technology "acts" in ways that at least seem intelligent. Yet, claims about the possibility of strong AI are still contested. Proponents of strong AI think that it is possible to artificially produce abilities and characteristics we would otherwise ascribe to human beings—for example, consciousness. As Haugeland puts it, "AI wants only the genuine article: *machines with minds*, in the full and literal sense. This is not science fiction, but real science, based on a theoretical conception as deep as it is daring: namely, we are, at root, *computers ourselves*" (Haugeland 1985, 2; original emphasis). The argument for this is deeply computational: since humans are ultimately thinking computers, there is no reason why computers should not be intelligent in the full and literal sense.

However, most researchers reject the idea that current AI technology is intelligent in the strong sense. Some do this because they disagree with the assumptions guiding strong AI: that the mind is in the business of representing and processing outside information and that this happens mainly in the brain and independently of our bodies. Many modern philosophical critiques of AI would adopt a different perspective on what constitutes human and natural intelligence, suggesting instead that intelligence is an embodied capacity that emerges on the basis of our bodily interactions with the environment, including with other people (see, for example, Dreyfus 1972, 1997). In this view, being intelligent consists precisely in what appears impossible to AI; namely, to establish a meaningful relationship with the world. The following section will consider this view in more detail.

Embodied and Situated Alternatives

What does it mean, exactly, when we say that intelligence is an emergent embodied capacity? And why is meaningfulness so crucial to intelligence? In this section, we will discuss perspectives on AI revolving around such

questions. These perspectives are rooted in the embodied approach to cognition.[6] *Embodied cognition* is an umbrella term for views grounded on two basic assumptions: First, our mind is not independent of our body—the body and its interactions with the world are crucial to the mind. Second, there is no clear-cut division between our inner mind and the outer world, in the sense that the world is just "out there," and minds are in the business of representing it or solving problems posed to us. Instead, the two are dynamically intertwined. Being intelligent consists in actively engaging with the environment on the basis of a goal-directed, subjective perspective. These assumptions stand in direct contrast to the representationalist, functionalist and computational perspectives discussed above.

The idea that the mind is embodied and situated played a role in early AI and robotics research as well. In the 1990s, roboticist Rodney Brooks (1991) illustrated the intertwinement of mind and world by building robots that autonomously explore an environment, not on the basis of detailed maps but by responding to the given context. Brooks argued that intelligence could be generated by creating an autonomous system with an artificial body through which it navigates its physical environment. Brooks' core idea was that intelligence should not only be understood in terms of high-level or abstract reasoning capacities. Lower-level perceptual abilities like sensing, moving, turning, grasping, and avoiding obstacles count as intelligent processes as well, and can be achieved with no reference to a central processing unit or internal representations. Through the use of sensors and actuators,[7] the robot explores its environment and, in this way, acts intelligently. Yet, while the body and physical movement are central to embodied theories, it does not follow from such theories that a robot would be intelligent. This is because theories of embodied cognition also emphasize aspects like the first-person perspective, context, and purpose.

Against the "View from Nowhere": Perspective and Context

One of the defining features of intelligence in computationalism and representationalism is information processing. Yet, such accounts ignore an important question; namely, *who* is the information being processed for? Data processed

---

6 The paradigm is also referred to as *4E cognition*, due to its four main theoretical movements: embodied, embedded, extended, and enactive cognition.

7 An actuator is a component in the robot that executes actions in the world. It converts input (electrical signals) into physical motion; that is, the robot's movements.

by an AI system is fed into the system by humans, and it has no relevance or meaning to the system itself. In the model, there is no agent, no subject relating to the information in any way. In this sense, AI entails a detached "view from nowhere" and lacks a "first-person" viewpoint from which the processed information matters.[8]

But many philosophers consider the first-personal perspective a key ingredient for consciousness and intelligence. In his famous article "What Is It Like to Be a Bat?," philosopher Thomas Nagel (1974) argued that there are facts about our conscious experience that can only be known from a subjective perspective. Even if we know all the objective and physical facts about bats, for instance, we do not know what it would really be like to be a bat. To repeat, reductive materialism has its roots in the mechanistic worldview and entails that all aspects of the mind can be fully explained in terms of physical aspects, such as in terms of the human brain. But Nagel challenged the idea that we could explain consciousness, as a key aspect of intelligence, in terms of a materialistic perspective alone. He argued that even if we could observe all physical aspects of someone's brain, we do not know how it feels to be this person (in the same way that we do not know what it is like to be a bat). In Nagel's view, subjective experience, rather than being a by-product of processes in the brain, is what constitutes consciousness.

But a conscious perspective does not arise out of nowhere. As suggested by researchers in the *enactive approach*,[9] consciousness and intelligence emerge through our interaction with our surroundings. They are enacted (brought forth). We are not "parachuted onto the earth as an unknown, objective reality to be charted" (Varela et al. 2017, 299). Intelligent beings are already part of an environment and are embedded in a context that is specific to them. This means that they always find themselves existing in a dynamic, co-constitutive relationship with their surroundings. Crucially, this is not the case for AI systems. As philosopher Hubert Dreyfus (as early as 1972, 1997) has argued, AI models are disembodied systems that stand in no meaningful relation to the world. Meaning is a prerequisite for intelligence, and it crucially depends on relevance. AI systems do not replicate meaning, because the information they are processing does not mean anything to them. Relevance is not something that can be programmed into the AI model, but only emerges in

8 The phrase "view from nowhere" stems from one of Nagel's later works, *The View from Nowhere* (1986).

9 The enactive approach is one of the theoretical movements within embodied cognition (or 4E cognition) which was discussed at the outset of this section.

systems that have intrinsic values and, precariously, depend on interactions with their surroundings.

Believing that a computer will achieve intelligence if we just provide it with sufficient information and computing power is thus a mistake according to Dreyfus. It is not the accumulation and manipulation of information per se that gives rise to intelligence, but rather our reasons and motivation for applying this knowledge in a relevant manner. In this view, intelligence arises through the everyday background understanding of relevance we gain when dealing with the world around us. It seems impossible that this know-how—together with all the feelings, interests, and bodily capacities that make up a human being—could be simulated by a computer.

Agency: The Living Organism and Its Concerns

The earlier discussion about perspective and context is closely linked to debates on the relationship between AI, embodiment, and agency. To provide some insight into this debate, let us take another look at the enactive approach already mentioned above. The enactive approach resonates with Aristotle's view on the body and promotes the idea that there is a *life–mind continuity*: intelligent systems show similar properties to living systems (see, for example, Maturana and Varela 1987; Thompson 2007).[10] This view also rejects a dualistic outlook on mind and body (Thompson 2005). Instead of the computer, the living, embodied organism becomes the guiding metaphor for what an intelligent system is (Varela et al. 2017). An organism is not just a material entity or a vessel used to navigate the world, but a *living* and *lived* (experiencing) body. It is therefore at once a mind and a subject, something that provides us with a basic perspective on the world (Di Paolo and Thompson 2014). In contrast to a computer, an organismic body is able to engage in meaningful relations with its environment because it has intrinsic values; it is first and foremost concerned with survival and maintaining its bodily (organismic) integrity. This need to stay alive and to self-maintain can be seen as a minimal instantiation of a first-personal viewpoint, from which a body relates in a meaningful way to the world (Di Paolo and Thompson 2014; Kyselo 2014).

---

10 According to this, cognition emerges in the interaction dynamic between living beings and their environment. Each type of living being (from single cells to multicellular organisms) has its own kind of cognition corresponding to a specific level of complexity (Hipólito and Martins 2017).

Embodied intelligence, then, consists in adaptively engaging with the environment in a way that helps the intelligent system itself to survive and flourish (Di Paolo 2005). Di Paolo uses the example of a bacterium swimming up a sugar gradient to illustrate this. We cannot fully explain the bacterium's movement towards sugar by reference to (biochemical) mechanisms alone. The bacterium seeks sugar because sugar signifies nutrition and is thus beneficial for the bacterium's bodily survival. The bacterium acts adaptively because not only does it evaluate sugar as being important for survival, but it also "understands" that higher concentrations of sugar will be better for achieving this goal in the long term. Enactive researchers therefore argue that meaning is not discovered in the world, but rather actively produced through the living system's goal-directed engagements with its environment; hence they use the term *sense-making*. An intelligent being, as an embodied being, adopts a concerned perspective that guides its interactions with the environment. Whether something appears to make sense or not depends on how it relates to this concerned perspective (Weber and Varela 2002).

This is closely related to the notion of *agency*, another central concept within the enactive approach. An agent is a system that brings about a self–other distinction (a boundary between itself and the world) and acts on its behalf, in that it regulates and evaluates its activities and interactions with the environment with regard to its own intrinsic goals (Di Paolo 2005).[11] Again, we can use the example of the bacterium for illustrative purposes. By actively engaging with its environment (seeking sugar), the bacterium produces and maintains its own metabolism, thereby generating a boundary (cellular membrane) between itself and the domain within which it acts (self–other distinction). In doing so it acts on its behalf, in that swimming towards higher concentrations of sugar and getting more nutrition will be beneficial to its intrinsic goals; that is, metabolic integrity (the self–other distinction). All agents are autonomous: they follow their own laws (Thompson 2007). The bacterium's actions are self-governed, grounded in the metabolic logic of maintaining its own life.

Although our human actions are more complex, the same principles apply to human agency. That is, through embodied interactions with the

11 There are various definitions of the term *agency*, and some researchers would argue that machines and other objects may have agency: see, for instance, Latour (2005). Their discussion is beyond the scope of this introduction, but we maintain that agency is first and foremost a capacity of living beings and that any meaningful artificial or technological form of agency will be a derivative of animal or human agency.

environment (material, social and cultural), we bring forth a self–other distinction which forms the precondition for our own agency—our ability to act on our own behalf in accordance with our intrinsic aims. It is thus through our concerned embodied being and engaging with the world that we become agents.

The same does not, however, easily apply to AI. From an enactive perspective, an AI system does not have agency since it does not care and does not have concerns of its own, nor intrinsic goals. An AI may learn, solve problems, and simulate goal-directedness, but its goals remain determined externally (programmed by particular human agents). Even where AI systems or robots may appear to simulate autonomous behavior (say, in robots or automated decision making), such behavior is not guided by intrinsic concern but remains heteronomous (guided by someone else's laws and goals). Since technological agency does not exist independently of human sense-making, we think it is more fruitful to focus on how AI shapes human agency and can give rise to hybrid forms of agency that emerge from the interaction between humans and AI systems. We will come back to this important question in the next chapter on ethics (Torbjørnsen and Kyselo 2026).

LLMs and Linguistic Agency

The above discussion brings us back to where we started: the large language models currently shaping many aspects of our everyday life. What do the previous discussions entail for our understanding and the use of LLMs like ChatGPT and Lemoine's LaMDA? Are they intelligent? Echoing the embodied and enactive perspectives discussed above, one would clearly have to argue that they are not. As philosopher Anthony Chemero (2023) puts it, no matter how advanced an LLM will become, however well it masters human language, it is still merely a tool for performing very specific tasks, and is for that reason and by principle intelligent only in the weak sense. Yet, how do we account for the fact that LLMs *appear* to be intelligent by acting as capable, human conversational partners? How can we explain the fact that they are so convincing?

A closer look at how LLMs function might help us answer this question. LLMs are based on transformer models and trained on massive datasets (with data broken down into tokens). While different systems operate in various ways, they are essentially well-defined calculations across tokens. Accordingly, any generated text is a result of engineering processes like word prediction or sequence extension, and it does not necessarily reflect "facts" in the real

world.[12] The output generated by LLMs might seem remarkably meaningful. Yet, as researchers Abeba Birhane and Marek McGann (2024) argue, the apparent meaningfulness relies entirely on the extent to which certain words, phrases, or paragraphs can be represented by the relations between tokens in the training dataset.

LLMs generally appear to be *linguistic agents*, but, as we saw above, what matters for agency is also sense-making and relevance, and to be able to participate in a linguistic, social community. A linguistic agent is someone (or a system) who can make sense of language and for whom linguistic expressions matter. This is not captured by datasets and relations between tokens (Birhane and McGann 2024). When we engage linguistically, there is something at stake. Is there something at stake for an LLM? Is there something it seeks to achieve? Does an LLM participate in meaningful social interactions? The answer to all these questions is "no." Instead, "LLMs are complex tools, and within any activity their role is that of a tool" (Birhane and McGann 2024, 10). And in as far as the AI model is only a tool, all meaning and value embedded in the conversation is generated and perceived by the user.

This point is important because it shifts the question from whether AI is truly intelligent to understanding more clearly what happens to *us*, the human user, when we interact with AI. It also leads us to a more general point: any AI system, embodied or not, is produced and guided by humans. Apparent machine intelligence can usually be explained with reference to the humans "behind the scenes," that is, the engineers and computer scientists creating it, the crowd workers supporting it, and the users interacting with it. Moreover, the design of any AI system will, in part, depend on those in power to produce and finance it (Birhane 2021; Crawford 2021; Noble 2018). As we show in the following chapter (Torbjørnsen and Kyselo 2026), these points are important to bear in mind when considering the ethics of AI.

Conclusion

Two important lessons can be drawn from a century-spanning philosophical debate on AI. First, whether we deem an AI to be intelligent depends on what we think about the nature of the mind and how minds relate to the world. If we accept or simply presuppose a mechanistic, computational theory of mind, we

12 See Giosuè Baggio and Matteo Radaelli's (2026) chapter in this volume. See also Bender et al. (2021) and Hicks et al. (2024).

will have good reason to think that computers can be intelligent. But once we question such traditional metaphysical outlooks, we realize that human intelligence depends on other attributes; attributes that AI simply does not have. The alternative outlook we have addressed, for instance, emphasizes the role of the body for intelligence. A body is always situated in a specific context, interacting meaningfully with the world in relation to subjectively experienced autonomous norms and goals. Second, understanding AI is not simply a matter of understanding how the technology functions, as if AI were God-given or appeared out of nowhere. It is also about understanding how AI depends on our human resources, our intelligence, our creativity, and our labor, and vice versa, on the ways in which the development and usage of AI affect us humans at the core, at the level of individual agents as well as societies.

Author Contributions: RRT wrote the first draft and was responsible for the revision process. MK contributed to the structure of the argument and critical revision of the manuscript. Both authors read and approved the final manuscript.

Acknowledgments: The authors wish to thank the editors Lukas R.A. Wilde and Thomas Brandt for their valuable comments on earlier versions of this text.

Works Cited

Aristotle. *Complete Works of Aristotle: The Revised Oxford Translation*. Edited by Jonathan Barnes. 2 vols. Bollingen Series. Princeton University Press, 1984.

Arkoudas, Konstantine and Selmer Bringsjord. 2014. "Philosophical Foundations." In *The Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William M. Ramsey. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139046855.004.

Aswini, K.P. 2024. "Contemporary Forms of Racism, Racial Discrimination, Xenophobia and Related Intolerance." *Human Rights Council*, June 3, 2024. https://docs.un.org/en/A/HRC/56/68.

Baggio, Giosuè and Matteo Radaelli. 2026. "Language and Intelligence in Minds and Machines." In *Humanities Perspectives on AI: An Interdisciplinary Introduction*, edited by Lukas R.A. Wilde, Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, 127–144. Oslo: Scandinavian University Press.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜." *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, March 3, 2021: 610–623. https://doi.org/10.1145/3442188.3445922.

Birhane, Abeba. 2021. "Algorithmic Injustice: A Relational Ethics Approach." *Patterns* 2 (2): 100205. https://doi.org/10.1016/j.patter.2021.100205.

Birhane, Abeba and Marek McGann. 2024. "Large Models of What? Mistaking Engineering Achievements for Human Linguistic Agency." *Language Sciences* 106 (November): 101672. https://doi.org/10.1016/j.langsci.2024.101672.

of Bohemia, Princess Elisabeth and René Descartes. [1643–1649] 2007. *The Correspondence between Princess Elisabeth of Bohemia and René Descartes*. Translated by Lisa Shapiro. Chicago: University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/C/bo4924190.html.

Brooks, Rodney A. 1991. "Intelligence Without Reason." *AI Memo* 1293, MIT Artificial Intelligence Laboratory, April 1991. https://people.csail.mit.edu/brooks/papers/AIM-1293.pdf.

Buckner, Cameron and James Garson. 2025. "Connectionism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman. Spring 2025. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2025/entries/connectionism.

Chemero, Anthony. 2023. "LLMs Differ from Human Cognition Because They Are Not Embodied." *Nature Human Behaviour* 7 (11): 1828–1829. https://doi.org/10.1038/s41562-023-01723-5.

Chen, Eddy Keming, Mikhail Belkin, Leon Bergen, and David Banks. 2026. "Does AI Already Have Human-Level Intelligence? The Evidence Is Clear." *Nature* 650: 36–40. https://doi.org/10.1038/d41586-026-00285-6.

Chomsky, Noam. [1959] 1996. "A Review of B.F. Skinner's 'Verbal Behavior.'" In *Readings in Language and Mind*, edited by Heimir Geirsson and Michael Losonsky, 48–63. Cambridge, MA: Blackwell.

Clark, Andy. 2001. *Mindware: An Introduction to the Philosophy of Cognitive Science*. Oxford: Oxford University Press.

Clark, Andy. 2008. *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford: Oxford University Press.

Crawford, Kate. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.

Descartes, René. [1637] 2006. *Discourse on Method*. Translated by Ian MacLean. Oxford: Oxford University Press.

Descartes, René. [1644] 1983. *Principles of Philosophy*. Translated by Valentine R. Miller and Reese P. Miller. Dordrecht: Reidel.

Descartes, René. [1664] 1972. *Treatise of Man*. Translated by Thomas S. Hall. Cambridge: Harvard University Press.

Di Paolo, Ezequiel A. 2005. "Autopoiesis, Adaptivity, Teleology, Agency." *Phenomenology and the Cognitive Sciences* 4 (4): 429–452. https://doi.org/10.1007/s11097-005-9002-y.

Di Paolo, Ezequiel and Evan Thompson. 2014. "The Enactive Approach." In *The Routledge Handbook of Embodied Cognition*, edited by Lawrence Shapiro and Shannon Spaulding, 68–78. New York: Routledge.

Dreyfus, Hubert L. 1972. *What Computers Can't Do*. New York: Harper & Row.

Dreyfus, Hubert L. 1997. *What Computers Still Can't Do*. Cambridge, MA: The MIT Press.

Fodor, Jerry A. and Zenon W. Pylyshyn. 1988. "Connectionism and Cognitive Architecture: A Critical Analysis." *Cognition* 28 (1): 3–71. https://doi.org/10.1016/0010-0277(88)90031-5.

Haugeland, John. 1985. *Artificial Intelligence: The Very Idea*. Cambridge, MA: The MIT Press.

Hicks, Michael Townsen, James Humphries, and Joe Slater. 2024. "ChatGPT Is Bullshit." *Ethics and Information Technology* 26 (2): 38. https://doi.org/10.1007/s10676-024-09775-5.

Hipólito, Inês and Jorge Martins. 2017. "Mind-Life Continuity: A Qualitative Study of Conscious Experience." *Progress in Biophysics and Molecular Biology, Integral Biomathics 2017: The Necessary Conjunction of Western and Eastern Thought Traditions for Exploring the Nature of Mind and Life* 131 (December): 432–444. https://doi.org/10.1016/j.pbiomolbio.2017.09.004.

Hirschberg, Julia and Christopher D. Manning. 2015. "Advances in Natural Language Processing." *Science* 349 (6245): 261–266. https://doi.org/10.1126/science.aaa8685.

Hobbes, Thomas. [1651] 2008. *Leviathan*. Edited by John C.A. Gaskin. Oxford: Oxford University Press.

Hurley, Susan. 1998. *Consciousness in Action*. Cambridge, MA: Harvard University Press.

Kyselo, Miriam. 2014. "The Body Social: An Enactive Approach to the Self." *Frontiers in Psychology* 5: 986. https://doi.org/10.3389/fpsyg.2014.00986.

Latour, Bruno. 2005. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780199256044.001.0001.

Lemoine, Blake. 2022a. "Is LaMDA Sentient?—An Interview." *Medium*, November 6, 2022. https://medium.com/@cajundiscordian/is-lamda-sentient-an-interview-ea64d916d917.

Lemoine, Blake. 2022b. "What Is Sentience and Why Does It Matter?" *Medium*, August 15, 2022. https://cajundiscordian.medium.com/what-is-sentience-and-why-does-it-mater-2c28f4882cb9.

Lewis, David. 1983. "Mad Pain and Martian Pain." In *Philosophical Papers* 1, 122–132. Oxford: Oxford University Press.

Maturana, Huberto R. and Francisco J. Varela. 1987. *The Tree of Knowledge: The Biological Roots of Human Understanding*. Boston: Shambhala Publications.

McCarthy, John, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. 1955. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence." *Application to The Rockefeller Foundation, Stanford University*, August 31, 1955. http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf.

McCulloch, Warren S. and Walter Pitts. 1943. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *The Bulletin of Mathematical Biophysics* 5 (4): 115–133. https://doi.org/10.1007/BF02478259.

Nagel, Thomas. 1974. "What Is It like to Be a Bat?" *Philosophical Review* 83 (4): 435–456.

Nagel, Thomas. 1986. *The View from Nowhere*. Oxford: Oxford University Press.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

Plato. *Plato: Complete Works*. Edited by John M. Cooper. Indianapolis: Hackett, 1997.

Putnam, Hilary. 1967. "Psychological Predicates." In *Art, Mind, and Religion*, edited by W.H. Capitan and D. D. Merrill, 37–48. Pittsburgh: University of Pittsburgh Press.

Rosenblatt, Frank. 1958. "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain." *Psychological Review* 65 (6): 386–408. https://doi.org/10.1037/h0042519.

Ruhil, Olivia. 2025. "Justice in the Age of Algorithms: Can AI Weigh Morality?" *AI & SOCIETY*, ahead of print, March 12, 2025. https://doi.org/10.1007/s00146-025-02292-z.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1986. "Learning Representations by Back-Propagating Errors." *Nature* 323 (6088): 533–536. https://doi.org/10.1038/323533a0.

Searle, John. 1980. "Minds, Brains, and Programs." *The Behavioral and Brain Sciences* 3 (3): 417–457. https://doi.org/10.1017/S0140525X00005756.

Searle, John. 1999. "The Chinese Room." In *The MIT Encyclopedia of the Cognitive Sciences*, edited by Robert A. Wilson and Frank C. Keil, 115–116. Cambridge, MA: The MIT Press.

Thompson, Evan. 2005. "Sensorimotor Subjectivity and the Enactive Approach to Experience." *Phenomenology and the Cognitive Sciences* 4 (4): 407–427. https://doi.org/10.1007/s11097-005-9003-x.

Thompson, Evan. 2007. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge, MA: Harvard University Press.

Torbjørnsen, Regine Rørstad and Miriam Kyselo. 2026. "Foundational Conceptual Issues in the Ethics of Artificial Intelligence." In *Humanities Perspectives on AI: An Interdisciplinary Introduction*, edited by Lukas R.A. Wilde, Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, 87–110. Oslo: Scandinavian University Press.

Turing, Alan. 1936. "On Computable Numbers, with an Application to the Entscheidungsproblem." *Proceedings of the London Mathematical Society* 42: 230–265. https://doi.org/10.1112/plms/s2-42.1.230.

Turing, Alan. 1950. "Computing Machinery and Intelligence." *Mind* 59 (236): 433–460. https://doi.org/10.1093/mind/LIX.236.433.

Varela, Francisco J., Evan Thompson, and Eleanor Rosch. 2017. *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: The MIT Press.

Weber, Andreas and Francisco J. Varela. 2002. "Life after Kant: Natural Purposes and the Autopoietic Foundations of Biological Individuality." *Phenomenology and the Cognitive Sciences* 1 (2): 97–125. https://doi.org/10.1023/A:1020368120174.

Wiener, Norbert. 1948. *Cybernetics: Or Control and Communication in the Animal and the Machine*. Cambridge, MA: The MIT Press.

Wolff, Christian. 1963. *Preliminary Discourse on Philosophy in General*. Translated by Richard J. Blackwell. Indianapolis: The Library of Liberal Arts/Bobbs Merrill.

4. Foundational Conceptual Issues in the Ethics of Artificial Intelligence

Regine Rørstad Torbjørnsen and Miriam Kyselo

Abstract: This chapter reflects on key ethical dilemmas surrounding AI, and the foundational conceptual issues underlying them. It addresses how humans play a crucial role in determining the values and functioning of AI technology, and, in turn, how human life and agency are fundamentally affected by AI use. The chapter illustrates that ethical considerations on AI do not exist in a vacuum but are themselves grounded in and constrained by our conceptions of the nature of intelligence and consciousness, and of how we think of the relationship between humans and their sociocultural and technological environment. By adopting a foundational perspective, the authors contribute to the discussion on the extent to which AI is developed and applied in a manner that ensures the welfare of individuals, societies, and ecosystems. The chapter demonstrates how a context-oriented perspective helps to shed light on the impact of AI on human agency, work, and knowledge acquisition, as well as the role of AI in climate change and nature degradation.

Keywords: agency, algorithmic injustice, ethics of AI, sustainability, techno-optimism

Artificial intelligence technology is not good or bad per se. It can be produced and applied in both beneficial and disadvantageous ways. It has proven to be useful in various domains, such as the automation of repetitive manual work (Moraes et al. 2022) and in more efficient and improved healthcare diagnoses (Esteva et al. 2017; Senders et al. 2018). However, the use of AI also presents crucial ethical challenges, leading to problems such as algorithmic bias and injustice (Birhane 2021; Noble 2018), emerging risks in warfare deployment (Schwarz 2025), nature degradation, and contribution to climate change (Cowls et al. 2023; Crawford 2021).

These challenges are discussed in the *ethics of AI*, a subfield in the field of ethics. Ethics is concerned with understanding what is right and wrong, morally good and bad, and with providing principles for how we ought to act morally.

Wilde, Lukas R.A., Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, eds. 2026. *Humanities Perspectives on AI: An Interdisciplinary Introduction*. Scandinavian University Press. https://doi.org/10.18261/9788294167234-26-04
Open access | © The author(s) 2026. This work is licensed under CC BY 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/.

In the ethics of AI, one important goal is to map the beneficial and detrimental consequences of the development and usage of AI, and to evaluate whether AI systems do in fact promote human, societal, and ecological welfare. This is expressed clearly in UNESCO's recommendations on the ethics of AI, which aim to

> protect, promote and respect human rights and fundamental freedoms, human dignity and equality, including gender equality; to safeguard the interests of present and future generations; to preserve the environment, biodiversity and ecosystems; and to respect cultural diversity in all stages of the AI system life cycle. (UNESCO 2022, 15)

These recommendations address the various stakeholders in AI, such as programmers, companies, law/policy makers, investors, users, etc., hinting at the complexity of situations to which the ethics of AI is responding. There are many actors involved and the field is vast, with numerous core debates, such as the impact of AI on democracy (Landemore 2024), social justice (Birhane 2021; Gebru 2020), law (Smuha 2025; Tasioulas 2023), warfare (Taddeo 2024), surveillance (Véliz 2024), and climate (Coeckelbergh 2021; Floridi 2023). Our chapter aims to contribute to these debates by adopting a foundational perspective where we assume that any ethical exploration of AI must be founded on the close investigation of human intelligence and agency, sociality, and worldviews (Birhane 2021; Birhane and McGann 2024). Thus, in addition to specific ethical issues, we are concerned with some of the core philosophical concepts related to or underlying ethical discourses. In this way, we aim to offer some nuance for ethical questions that are already being asked—and potentially inspire thinking about new ones.

From such a perspective, we arrive at a better understanding of a fundamental problem with modern views of AI, namely that many still think of AI as pure algorithms (Crawford 2021), a form of "God-given" intelligence that is just there and absorbs and produces knowledge independently of its creators and users. Such a narrow view of AI distracts from several important facts. First, the apparent "intelligence" of AI originates from the humans who design, maintain, steer, and deploy it. Second, as we saw in the previous chapter on the "Introduction to the Philosophy of Artificial Intelligence" (Torbjørnsen and Kyselo 2026), current AI systems could be described as intelligent in the "weak" sense: they are powerful tools humans can use to complete specific tasks. They are not, however, intelligent in the "strong" sense since they

lack intentionality and agency. Third, AI technology lacks intrinsic goals and values. It operates according to norms and aims that are set externally by the people who develop and own the AI and the political actors in charge of regulating the industry. Related to that is the fourth point, namely that AI does not exist in some virtual or neutral space, but relies on massive amounts of human, technological, and natural resources.

To understand the wide-ranging impact of AI thus presupposes a reflection on the nature of the basic components within the complex human–AI system. Humans play a crucial role in determining the values and functioning of an AI, and human life and agency are fundamentally affected by AI use. Thus, this chapter selectively focuses on three ethical dilemmas and associated foundational issues: AI's impact on human agency and work, the impact of Large Language Models (LLMs) on knowledge acquisition and cognitive capacities, and the contribution of AI to climate change and nature degradation.

AI, Work, and Agency

By automating demanding and repetitive tasks, AI has become a tool that can relieve humans of certain labor (Spencer 2025). One potential positive consequence of this is that it gives people more time to pursue other meaningful professional and personal activities (Spencer 2022; Srnicek and Williams 2015). The ethical dilemma, however, is that automation might also lead to job insecurity, de-skilling (the loss or weakening of human skills and knowledge), as well as a diminished experience of work, such as feeling that one is less valuable or has less agency. In the following, we discuss the promise of less human work and some potential challenges to this scenario. We further discuss the close connection between work and agency, and address how AI automation might threaten human agency by diminishing the potential of the workplace as a resource for social engagement and application of one's skills and knowledge. Finally, we address ways of countering some of these issues by looking closer at human–AI professional collaborations.

The Promise of Less Work

Generative AI allows the automation of work to expand to new realms, including image and speech recognition, translation, accounting, data analysis, and generation of text and images (Gozalo-Brizuela and Garrido-Merch'an 2024; Eloundou et al. 2024). This suggests that an increasing number of tasks

previously performed by humans can be outsourced to AI. Inspired by these developments, some believe that AI could change our relationship to work entirely. The hope is that AI can help relieve humans of many of the burdens connected to production. Consider, for example, the following reflections of Sam Altman, CEO of OpenAI:

> we expect that this technology can cause a significant change in labor markets (good and bad) in the coming years, but most jobs will change more slowly than most people think, and I have no fear that we'll run out of things to do (even if they don't look like "real jobs" to us today). People have an innate desire to create and to be useful to each other, and AI will allow us to amplify our own abilities like never before. As a society, we will be back in an expanding world, and we can again focus on playing positive-sum games. (Altman 2024)

Altman thereby taps into the *techno-optimistic vision* that technology, especially AI, will increase productivity and free human workers from repetitive or otherwise tedious tasks, leaving us with time to engage in more meaningful activities (for an analysis and evaluation of techno-optimism, see Danaher 2022).[1] This is a scenario that sounds attractive to many, but whether it is realistic is highly unclear. Discussions on the future of work are extremely complex (Ghosh et al. 2025; OECD 2024), and thus such ideas are mere speculation at this point.

The deeper and more urgent question is whether increased automation will really benefit workers. As Daron Acemoglu and Simon Johnson (2023) remark, many companies are using AI to minimize the cost of labor and intensify work, aiming to generate higher returns for investors and capital owners. Yet, while companies profit from the cheaper labor of automation, workers are not getting their share of the revenue (Acemoglu and Restrepo 2020; Dauth et al. 2021). Importantly, many AI systems rely on underpaid crowd workers, who perform tasks such as labelling massive amounts of training data or reviewing suspicious, harmful, or disturbing content (Crawford 2021; Gray and Suri 2019; Irani 2015). Given current socioeconomic conditions, the prospects of AI automation fulfilling the techno-optimist's positive vision of providing workers with more leisure time appear rather bleak (Acemoglu and Johnson 2023; Spencer 2022).

---

1 Thinkers currently positioning themselves clearly in favor of techno-optimism include Deaton (2013), McAfee (2019), Naam (2013), Pinker (2018), and Rosling et al. (2018).

This does not mean that the idea that AI might facilitate human labor is not important. While AI may not bring about a "no-work" utopia anytime soon, it clearly has a substantial impact on how we work. This leads to worries about large-scale displacement or replacement (Rion and Any 2025) and the ways in which AI might reduce the quality of work performed by humans (Tyson and Zysman 2022). As we shall see, both worries are linked to a more fundamental human concern, namely the need to develop and practice our skills in a social environment conducive to a meaningful life.

Work and Agency

We have already mentioned that work is a crucial part of most people's lives. Ideally, work provides financial security and, thereby, a basic form of existential safety. But jobs and careers also provide the possibility of engaging in meaningful professional social relations and the opportunity to cultivate and practice one's knowledge and skills. Work fundamentally shapes our being in the world and our identity (Ashforth and Schinoff 2016; Nelson and Irwin 2014). This is because the social interactions and activities in which we engage in the workplace are important to how we develop and flourish as people. Thus, the automation of work might have consequences for our sense of self (Endacott 2021). In the following, we will look more closely at the close connection between work and agency and explore how AI automation impacts humans in this regard.

The term *agency* is used in many different contexts and is sometimes employed interchangeably with *the self* (see Kyselo, forthcoming).[2] In line with the previous chapter (Torbjørnsen and Kyselo 2026), we adopt an enactive outlook on agency. An agent is thus described as a self-organizing system that creates and maintains a form of self or identity through interactions with its environment. An agent thereby follows its own goals, evaluating and adapting to its surroundings and interactions according to what it needs to maintain itself. Whether interactions are seen as meaningful depends on whether or how they contribute to these goals. In this way, the agent self-individuates, which means that it actively brings forth its identity (Di Paolo and Thompson 2014; Jonas 1966). Having a sense of agency means to feel that one is the origin of one's actions. From an enactive perspective, this experience is directly dependent

2 Some use the term in a way that suggests that other animals, artificial systems, or even fictitious/spiritual figures could be agents themselves, but here we focus on human agents. See Latour (2005) for an alternative account that highlights the active role of technology.

on having meaningful interactions with the environment (Buhrmann and Di Paolo 2017).

The crucial point here is that our interactions with the environment make us who we are. By interacting with our surroundings, we create a boundary between ourselves and the world, including the boundary between ourselves and other agents. The enactive view of agency thereby emphasizes that the boundary between agent and world is not clear-cut or given but an actively generated, ongoing process, to which the social environment and technology both contribute. Thus, our agency may be co-created and is partially reliant on social and other processes outside of the individual organism (Kyselo 2014; forthcoming).[3]

The central role of the social environment for our sense of agency becomes highly relevant in the context of AI and for understanding its impact on the work context. At work, we engage in social interactions, complete various tasks, and practice our knowledge and skills. These engagements matter beyond the context of the workplace. They affect our agency, who we are, and how we are able to navigate the world around us. The workplace can furthermore provide people with an important source of meaning and direction in their lives (Ashforth and Schinoff 2016).[4]

For these reasons, if we lose the work context or the opportunity to practice certain tasks, our sense of agency and identity might be threatened. As fundamentally social agents, many people depend on the kind of interactions and recognition that the working environment can provide (Ashforth and Schinoff 2016). Moreover, the workplace may provide regular opportunities to practice our knowledge and skills, for instance through problem solving and teamwork, helping us to develop and maintain a balanced sense of agency. Conversely, when such opportunities are lacking, we might feel that the process of developing our identity is altered (Endacott 2021) or that our sense of agency is at stake.

---

3 See the chapter on the philosophy of AI in this volume (Torbjørnsen and Kyselo 2026) for a more detailed account of enactive agency; see also Kyselo (forthcoming).

4 In a similar vein, Gheaus and Herzog (2016) argue that society has an obligation to ensure people's access to meaningful work. They point out that certain valuable aspects of life, like making social connections and experiencing community, are difficult to realize outside of work. And since our public institutions regulate the burdens of labor, society should ensure we have access to meaningful work.

De-skilling and Extraction

Automation has long been a topic in the literature discussing how working conditions shape labor and human agency. In the industrial age, Karl Marx (1818–1883) argued that there had been a shift in power between human and machine brought on by the introduction of machinery to the workspace: whereas earlier tools still allowed workers to be (relatively) independent, machinery determines and regulates their movements "on all sides" (Marx [1939–1941] 1973, 692–693). In this way, workers are constrained not only physically, but mentally as well. Knowledge and skills previously needed from each worker are now required only for the workshop as a whole (Marx [1867] 1981). This will often lead to de-skilling, which means that human skills and knowledge are passed over from human workers to machines, and thus no longer need to be cultivated by workers (Marx [1939–1941] 1973). As indicated in the passage above, for many people, cultivating such capacities is also crucial for maintaining and nourishing a sense of agency. Actively practicing and developing our skills can be a way of affirming who we are and engaging in activities that are meaningful to us. De-skilling can therefore also affect and diminish the human sense of agency.

Marx also describes a process of extraction. Skills and knowledge that previously belonged to workers are now defined as belonging to the workshop and its owners. The owners thus profit from workers' long-cultivated capacities, which the workers themselves are hindered from applying in a self-determined manner. This further leads to a loss of the workers' sense of ownership over their capacities (Marx [1867] 1981). The sense of ownership goes along with the sense of agency (Gallagher 2000). Whereas the latter concerns the feeling of being the origin or cause of an action, the former refers to the sense that it is indeed I who am having an experience or performing the work. Matteo Pasquinelli (2023) elaborates on Marx's argument that the workshop de-skills workers and points out that processes of de-skilling and extraction are reinforced by the use of AI (see also Walton and Nayak 2021). It has become clear that all work, including tasks often assumed not to involve much skill, require substantial human knowledge and intelligence. For instance, problems with self-driving vehicles have illustrated how driving a vehicle has always been a human cognitive, social, and cooperative endeavor. As Pasquinelli (2023) points out, the aim of AI is to capture the knowledge expressed through human behavior and encode it into algorithmic models to automate a wide range of tasks. Thus, the apparent, sometimes impressive, knowledge and skills displayed by the machines originate in human behavior.

This leads to a twofold problem. First, the AI industry extracts human skills and knowledge while falsely presenting them as purely algorithmic achievements, leading many people to overestimate what AI is, in fact, capable of. At the same time, the (partial) replacement of human knowledge and skills through AI applications prevents workers from further cultivating these capacities and risks diminishing workers' sense of ownership and agency.

Human-AI Collaboration

There are, however, ways in which AI can be used in a beneficial manner, and thus help workers develop as self-determined, knowledgeable, and skillful agents. Some promising directions can be found in areas where humans collaborate with AI, for instance in medical diagnostics. In recent years, there has been an upsurge of systems using machine learning to classify specific regions on medical images, to identify—for example—a benign or malignant tumor (Smids et al. 2020). These systems have accuracy rates that are comparable to—and, in some cases, even better than—those of human doctors (Esteva et al. 2017; Senders et al. 2018). For instance, research indicates that combining human diagnosis with AI classification can reduce the error rate in breast cancer diagnostics by 85% compared to the human alone (Wang et al. 2016; Smids et al. 2020). Researchers have theorized that this might be because human professionals and AI make different types of mistakes and thus can correct one another (Senders et al. 2018; Smids et al. 2020).

Hybrid AI-doctor systems may therefore share the workload (of diagnosis) and, in a sense, also the agentive resources required for the work. This highlights a dilemma. How do we make sure that clinicians experience a balance in interactions with AI, and not a loss of agency? When collaborating with an AI on diagnostics, medical professionals may still feel affirmed in their role. They might invest the time saved by the machine in other areas of their work in developing their craft, and potentially even experience more meaningfulness (Smids et al. 2020). At the same time, this entails a risk of (partial) replacement of medical professional expertise and a loss of or an imbalanced sense of agency. Confronted with an AI system, the medical expert might perceive a sense of limitation in the practice and cultivation of their capacities, since much of their knowledge and skills appear to be replaceable. This underlines a point made throughout this chapter: whether AI will prove conducive to human labor and agency essentially depends on how it is applied. And how to apply AI is normally not a decision made by individual workers. Rather, it is up to society at large, and workspaces more specifically, to handle the potential of AI with care for the humans involved.

Knowledge and Human Capacities

AI is also applied in the context of knowledge acquisition. Knowledge is the foundation of our modern societies. What we know matters. We use knowledge to navigate the world around us, position ourselves in society, and gain insight into how local and global societies are structured. Our well-being depends on factors such as our insight into how things are and function. Knowledge is therefore power (Collins P. H. 2000), and so who knows what also matters. Large Language Models (LLMs) can help make research more efficient, help improve texts, be a dialogue partner in discussing complex topics, and provide immediate guidance for students. But these models come with important challenges pertaining to truth, bias, and expertise. In the following, we look closer at ways in which LLMs impact the process of knowledge acquisition and how using them in writing could diminish students' cognitive capacities and prevent them from developing a kind of reasoning crucial to human nature and agency.

LLMs and Knowledge Acquisition

What are the ethical implications when knowledge acquisition is handled with or by AI? With widespread access to LLMs, new questions about knowledge acquisition arise: Can LLMs be said to generate knowledge? How can we use them to reliably acquire or produce accurate information?

LLMs can be useful tools in knowledge acquisition processes. For instance, many IT students use LLMs as active tutors. The models can explain programming concepts, and their immediate feedback helps students debug errors faster (Azoulay et al. 2025; Mutanga et al. 2025). Yet, such use becomes a challenge in cases where students rely heavily on the tool and thereby avoid engaging fully in the process of learning how to program (Felipe-Redondo et al. 2024; Mutanga et al. 2024). Research indicates, however, that these issues are easier to avoid for more experienced users, as prior knowledge of how to handle LLMs as learning/knowledge-seeking tools generally improves the learning process (Chen et al. 2024; Soll and Kobras 2025; Zamfirescu-Pereira et al. 2023).

While the research on LLMs and learning is still at an early stage, it is reasonable to assume that one of the reasons experienced users learn more effectively is that they are more aware of a model's limitations. They might be more likely to recognize errors, false information, and bias. Since knowledge is power, in this context, having power lies in knowing how LLMs function and what they can and cannot do. Students who are inexperienced or not yet knowledgeable in an area and blindly rely on LLMs thus paradoxically risk contributing to

reducing their own power. Below, we discuss some of the challenges that we face when applying LLMs as tools for knowledge acquisition and address ways in which we can begin to tackle them.

As seen in this book's chapter on "Language and Intelligence in Minds and Machines" (Baggio and Radaelli 2026), LLMs are generative AI models trained on massive data sets and next-word prediction tasks. The models, in various ways, discover and reproduce significant aspects of textual patterns. Even if the generated text appears meaningful, it is a result of processes like word prediction and sequence extension and is not necessarily true or factual.[5] Of course, we humans also predict as part of the knowledge-seeking process, but we do not automatically conclude that everything we predict is accurate. We evaluate the conclusions we arrive at against the background of prior knowledge and the context in which we are embedded, and (ideally) aim to avoid spreading misinformation. LLMs, on the other hand, currently present the generated text as accurate regardless of whether this is the case. This behavior is often referred to as "hallucinations" (OpenAI et al. 2024; Weise and Metz 2023).[6] Other terms to describe this are "bullshitting" (Hicks et al. 2024)[7] or "confabulation" (Cao et al. 2022; Sui et al. 2024). Importantly, LLMs do not "understand" that they are misinforming. As pointed out earlier, LLMs do not understand at all. They are "intelligent" only in a weak sense (Searle 1980), being powerful tools we can use to complete specific tasks but lacking mental states and abilities of their own. This is also why LLMs are not committed to principles of veracity and reliability (Collins H. 2025). Their apparent "intelligence" and "confidence" stem from the humans who programmed and designed them, and from all of us who collectively feed them data (Bender et al. 2021; Birhane and McGann 2024).[8]

While many users are aware of the limitations and risks of these models, they might still be tempted to think that the AI is "right" at least most of the time. Moreover, since LLMs present false information alongside things we do know to be true, they often enforce an impression of authority. As Birhane (2021) argues, our inclination to give LLMs authority over knowledge and how to handle it stems from Western Cartesian dualism and a desire for "absolute

---

5 See also the section on linguistic agency in this book's chapter 3 (Torbjørnsen and Kyselo 2026).

6 The term "AI hallucination" has been criticized as misleading, as it might suggest that an LLM is an agent with experiences, sense impressions, and beliefs (Bergstrom and Ogbunu 2023; Hicks et al. 2024).

7 "Bullshitting" in the sense described by Harry Frankfurt (2005).

8 See chapter 3 (Torbjørnsen and Kyselo 2026) for a more elaborate discussion of behavior, intelligence, and agency in the realm of AI.

rationality."[9] We assume that abstract and intellectual thinking are "the most trustworthy forms of understanding" (Birhane 2021, 3). The fields most crucial for AI development—computer and data science in particular—have adopted this worldview (Mhlambi 2020). Implicitly or explicitly, they are driven by the deep conviction that data may give rise to universal, relatively unchanging, and objective knowledge. Consequently, these fields largely operate on the basis of a view of the data scientist or the data engineer as inhabiting a rational or "neutral" view on the world, and they assume that this rationality is reflected in how an AI generates and handles knowledge (Birhane 2021).

Yet, such a "view from nowhere" (Nagel 1986) does not exist. The idea of absolute rationality or objective knowledge is problematic and entails a skewed perspective on human intelligence. To think that abstract knowledge is the foundation of reality is to confuse the map with the territory. Abstract knowledge is dissociated from experience, from the process through which it is derived, and from the agent for whom it appears to be meaningful (Frank et al. 2024). Crucially, AI models are also not neutral, but always inherit the worldview of the people designing them in addition to the worldview emerging from the data that has been used for their training. Indeed, it is for that reason that AI often generates and reproduces historical patterns of injustice and discrimination (Bender et al. 2021; Benjamin 2019; Birhane 2021; Noble 2018; O'Neill 2016).[10] These patterns represent a different kind of error; that is, a type of error that cannot be corrected by technological means alone because it is so deeply rooted in historical and contemporary injustice, asymmetrical power hierarchies, or other unexamined problematic assumptions (Birhane 2021). Importantly, some actors exploit the fact that AI is *not* neutral by training AI models with massive amounts of false and/or offensive data in order to fuel and render political propaganda and hate speech as being "the truth."[11]

It should be clear that such issues need to be resolved and regulated at societal and institutional levels. Currently, it is still up to users themselves to skillfully navigate the technology and to critically examine the information provided by LLMs. There is thus an important knowledge gap, a vital need to educate users and to develop proper "AI literacy" (Burgsteiner et al. 2016; Kandlhofer

---

9 See chapter 3 (Torbjørnsen and Kyselo 2026) for an account of Cartesian dualism and the mechanistic world view.

10 See also the present book's chapter 5 on "AI and Equality" (Søraa and Ask 2026).

11 A recent example is Elon Musk's chatbot Grok, which has been flagged for generating wrong and politically skewed statements, offensive language, and conspiracy-tinged claims. In September 2025, Musk's company xAI signed an agreement with the US government allowing them to provide Grok to US federal agencies (Montgomery 2025; Reuters 2025).

et al. 2016). Schools and universities, especially the humanities, as argued in the introduction to this volume (Kyselo et al. 2026), play an important role in increasing AI literacy. They provide methods for critical and self-reflective thinking and crucial knowledge about the human dimension in the human–AI relationship and how it is impacted by AI use. It is important to note that an informed perspective on how AI impacts human agency, abilities, and meaning-making also raises awareness of potential moral obligations to assume a much-needed *human* responsibility. AI literacy is something that cannot be taken for granted, nor should we leave young users to figure it out entirely on their own. Like other skills, it needs to be taught, trained, and practiced.

ChatGPT, Cognitive Capacities, and the Cultivation of Humanity

A further concern is that the use of LLMs might diminish cognitive capacities. Let us consider a concrete example for illustrative purposes namely the writing of essays. In a recent study by MIT Media Lab (n.d.), participants were asked to write an essay and assigned to three groups: a group using LLMs to write the essay, a group using search engines, and a group relying on human cognition alone ("brain-only") (Kosmyna et al. 2025). During writing and subsequent interviews, their brain activity was recorded. While the benefits of using an LLM to write essays were initially apparent, the LLM group performed worse than the brain-only group at neural, linguistic, and behavioral levels. Participants using an LLM during writing displayed significantly lower cognitive engagement, compared to the two other groups. In other words, participants using an LLM "offloaded" their cognitive efforts to the model. As a consequence, they had trouble quoting from their own essay and also reported a lower sense of ownership over the text. Based on these preliminary results, researchers emphasize the urgent need to continue examining the detrimental effects and decrease in learning skills in the general population using LLMs.[12]

The potential decrease in cognitive activity and learning skills constitutes an ethical dilemma: a tool that entire populations perceive as a promising assistant in knowledge acquisition and learning might paradoxically hinder those very processes. In our knowledge-based modern societies, a potential general

12 At the time of writing, the study is still under review. Consult the lab's website for further information (MIT Media Lab n.d.).

decrease in learning and cognitive skills is concerning. Moreover, there might be consequences for human nature and our sense of who we are. In fact, some philosophers believe that "offloading" learning and reasoning processes to LLMs could diminish our human nature (Aylsworth and Castro 2024). The argument is based on the Kantian conception that our ability to reason—*rational agency*—is what makes us human. Rational agency involves having a say in determining one's own aims and being able to reflect on their value. In other words, we humans are able to step back from our desires and reflect on them, to decide which ones matter the most and cohere with our notion of the good (Kant 2000, 5:426; 2013, 7:322). One could use Kant's premise to argue that writing a humanities essay gives students the opportunity to develop such reasoning, and thereby to cultivate their own humanity. In contrast, when reasoning processes are "offloaded" to an LLM such as ChatGPT, students risk participating in a process that diminishes their humanity (Aylsworth and Castro 2024).

Note that Kant's account of agency differs from the enactive account of agency introduced in the first section of this chapter. An enactive perspective does not necessarily presuppose that agency is a rational capacity. Agency involves the generation of a self–other distinction; which may well be constituted through linguistic processes (Di Paolo et al. 2023), but importantly, also through pre-reflective, bodily, and co-embodied processes (Kyselo, forthcoming). The crucial problem is that the use of LLMs not only impacts certain cognitive skills (reasoning, expressing one's thoughts, etc.), but might also put our human agency at stake. Thus, whether the use of LLMs furthers or hinders agency will depend on users' prior expertise and ability to apply the model in a way that furthers their reasoning and balanced maintenance of agency. The impact of LLMs on agency is thus an overlooked and complex topic that requires further investigation. The enactive proposal, according to which human agency is a constantly developing system that may incorporate non-individual processes, provides a good foundation for analyzing AI's role in contributing to that structure, for better or worse, and consequently for providing more nuance to the discourse in the ethics of AI.

Climate and Natural Degradation

As we have seen, AI is more than "algorithms in the cloud." To run it requires natural and human resources, such as energy, minerals, water, land, electricity, and labor. This means that the AI industry manages and profits from our

shared resources: the air we breathe, the water we drink and use for agriculture, the ecosystems we live in and are so dependent on. AI cannot exist without these things, and nor can we. How do we make sure that the development, production, and usage of AI is both socially and ecologically sustainable? Guided by this question, the current concluding section considers both the advantages and disadvantages of AI in the context of climate change and natural degradation. As Luciano Floridi (2023) remarks, AI constitutes a double-edged sword in this context. On one side, it can be a very powerful tool against climate change. For instance, the processing of immense volumes of data made possible by AI can help improve and expand our understanding of the climate. It makes it easier to study existing climate trends, forecast future developments, and predict the potential success of climate policies. AI tools can also help anticipate extreme weather, which is becoming more common due to global climate change. Moreover, AI can contribute to the fight against global warming more actively by delivering greener and more effective solutions (Floridi 2023). For example, AI can help us improve and optimize energy generation and its uses. AI-based techniques are already being implemented in this way, and research indicates that they are playing an important role in the response to climate change (Inderwildi et al. 2020; Sayed-Mouchaweh 2020).

The ethical downside is that the training and use of computation-intensive AI systems lead to considerable energy use and carbon emissions (Amodei and Hernandez 2018; Cowls et al. 2023; García-Martín et al. 2019). Moreover, it is difficult to determine, exactly, the carbon footprint of AI, since sufficient information about the source and amount of energy used in researching, developing, training, and deploying AI models is lacking (Cowls et al. 2023). Floridi (2023) therefore maintains that it is currently difficult to say anything about the trade-off between the emissions caused by AI and the resource efficiency gains that AI can offer.

Kate Crawford (2021) interprets the matter differently. Like Floridi, she discusses emissions and energy use stemming from AI, but she also focuses on how the AI industry drains natural and human resources, damaging ecosystems and local communities. She describes the AI industry as part of a detrimental "logics of extraction" that affects all life on earth (Crawford 2021, 28). The notion of AI as "extraction" was already mentioned in our discussion on AI and work. Here, we see how not only human skills and knowledge, but also our shared natural resources are extracted. The mining of minerals is at the core of the issue. As an example, consider lithium, which is a natural resource crucial to the batteries powering AI technology. Lithium is mined in places

such as southwest Bolivia, central Congo, Mongolia, Indonesia, and the deserts of Western Australia. Some of these places are impacted by local and geopolitical violence. Extraction of these minerals in these areas may fuel or reinforce such conflicts. For instance, profits from mining have financed military operations in the decades-long Congo-area conflict (Burke 2017; Crawford 2021). Major tech companies like Philips, Intel, and Dell may contribute to this problem when they cannot guarantee that their supply chain is not involved in such conflict (Abraham 2017; Crawford 2021). In addition to fueling social and political problems, the mining of natural resources necessary to operate AI technology causes considerable environmental damage, ranging from devastated landscapes and polluted rivers and lakes to the extinction of entire plant and animal species (Abraham 2017; Crawford 2021; Liu 2012).

Conclusion

This chapter focused on three selected ethical dilemmas in the field of AI: the impact of AI on human agency and work, the influence of Large Language Models (LLMs) on knowledge acquisition and cognitive capacities, and the contribution of AI to climate change and nature degradation. We adopted a foundational perspective on the ethics of AI, assuming that any ethical exploration of AI presupposes a particular stance on basic concepts such as human intelligence, embodiment, or agency. The aim of such a foundational perspective is to inform the discussion on whether AI technology is developed and applied in a manner that ensures the welfare of individuals, societies, and ecosystems. It becomes clear that the industry, institutions, governments, and individual users all bear responsibility for ensuring that AI is ethical. Philosophers and other scholars in the humanities contribute to such efforts by shedding light on the complexity and context dependency of AI.

The industry often sees AI ethics and justice as concerns that society and governments are responsible for dealing with, not them.[13] This is problematic, since the AI industry is responsible for the emissions and nature degradation that it causes, in addition to its impacts on human agency and flourishing.

13 Companies like Google, Microsoft, Anthropic, and DeepMind do have substantial ethics research programs, responsible AI frameworks, etc. However, such programs and frameworks have been criticized for being incomprehensive and incoherent, and, more importantly, for not being anchored in sincere concerns about the ethical implications of these companies' AI technology (Crawford 2021; van Maanen 2022; Munn 2023; Seele and Schultz 2022).

Yet, the industry's response to ethical concerns is unfortunately often limited to signing general AI ethics agreements and statements. This not only evades or delegates responsibility, but such documents are also often taken to reflect general consensus among stakeholders when they actually do not represent the regions suffering most from the negative consequences of operating AI systems, such as Africa, South and Central America, or Central Asia (Crawford 2021; Jobin et al. 2019).[14] In addition, ethical agreements and statements normally fail to address how they should be implemented, and are seldom enforceable or accountable to the wider public (Crawford 2021; Crawford et al. 2019).

To make this important point even more explicit, companies' reluctance to take responsibility is deceptive because it obscures their crucial role in determining how AI should be designed and applied. It is therefore important to assign accountability appropriately. As we have seen, AI technology is not a neutral, virtual thing that is just "there" for companies to use, but designed, programmed, and produced by people who in turn exhibit values and worldviews of their own. Thus, to hold the industry, institutions, and governments accountable, we need more transparency and knowledge about how individuals, societies, and ecosystems are impacted by AI. In other words, increasing AI literacy is of the utmost importance.

Author Contributions: RRT wrote the first draft and was responsible for the revision process. MK contributed to the structure of the argument and critical revision of the manuscript. Both authors read and approved the final manuscript.

Acknowledgments: The authors wish to thank Lucy Osler and editor Lukas R.A. Wilde for their helpful comments on earlier versions of this chapter.

Works Cited

Abbott, Alden. 2025. "U.S.-China AI Competition in The Spotlight." *Forbes*, July 29, 2025. https://www.forbes.com/sites/aldenabbott/2025/07/29/us-china-ai-competition-in-the-spotlight.

Abraham, David S. 2017. *The Elements of Power: Gadgets, Guns, and the Struggle for a Sustainable Future in the Rare Metal Age*. New Haven: Yale University Press.

14 Further, there is an AI "arms" race unfolding between the US and China, leaving little incentive for these powers to dial back AI development (Abbott 2025; Zhang 2025).

Acemoglu, Daron and Simon Johnson. 2023. *Power and Progress: Our Thousand-Year Struggle over Technology and Prosperity*. London: Basic Books.

Acemoglu, Daron and Pascual Restrepo. 2020. "The Wrong Kind of AI? Artificial Intelligence and the Future of Labour Demand." *Cambridge Journal of Regions, Economy and Society* 13 (1): 25–35. https://doi.org/10.1093/cjres/rsz022.

Altman, Sam. 2024. "The Intelligence Age." September 23, 2024. https://ia.samaltman.com.

Amodei, Dario and Danny Hernandez. 2018. "AI and Compute." *OpenAI*, May 16, 2018. https://openai.com/index/ai-and-compute.

Ashforth, Blake E. and Beth S. Schinoff. 2016. "Identity under Construction: How Individuals Come to Define Themselves in Organizations." *Annual Review of Organizational Psychology and Organizational Behavior* 3: 111–137. https://doi.org/10.1146/annurev-orgpsych-041015-062322.

Aylsworth, Timothy and Clinton Castro. 2024. "Should I Use ChatGPT to Write My Papers?" *Philosophy & Technology* 37 (4): 117. https://doi.org/10.1007/s13347-024-00809-w.

Azoulay, Rina, Tirza Hirst, and Shulamit Reches. 2025. "Large Language Models in Computer Science Classrooms: Ethical Challenges and Strategic Solutions." *Applied Sciences* 15 (4): 1793. https://doi.org/10.3390/app15041793.

Baggio, Giosuè and Matteo Radaelli. 2026. "Language and Intelligence in Minds and Machines." In *Humanities Perspectives on AI: An Interdisciplinary Introduction*, edited by Lukas R.A. Wilde, Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, 127–144. Oslo: Scandinavian University Press.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜." *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, March 3, 610–623. https://doi.org/10.1145/3442188.3445922.

Benjamin, Ruha. 2019. *Race after Technology: Abolitionist Tools for the New Jim Code*. Cambridge: Polity.

Bergstrom, Carl T. and C. Brandon Ogbunu. 2023. "ChatGPT Isn't 'Hallucinating': It's Bullshitting." *Undark Magazine*, April 6, 2023. https://undark.org/2023/04/06/chatgpt-isnt-hallucinating-its-bullshitting.

Birhane, Abeba. 2021. "Algorithmic Injustice: A Relational Ethics Approach." *Patterns (New York, N.Y.)* 2 (2): 100205. https://doi.org/10.1016/j.patter.2021.100205.

Birhane, Abeba and Marek McGann. 2024. "Large Models of What? Mistaking Engineering Achievements for Human Linguistic Agency." *Language Sciences* 106 (November): 101672. https://doi.org/10.1016/j.langsci.2024.101672.

Buhrmann, Thomas and Ezequiel Di Paolo. 2017. "The Sense of Agency—a Phenomenological Consequence of Enacting Sensorimotor Schemes." *Phenomenology and the Cognitive Sciences* 16 (2): 207–236. https://doi.org/10.1007/s11097-015-9446-7.

Burgsteiner, Harald, Martin Kandlhofer, and Gerald Steinbauer. 2016. "IRobot: Teaching the Basics of Artificial Intelligence in High Schools." *Proceedings of the AAAI Conference on Artificial Intelligence* 30 (1). https://doi.org/10.1609/aaai.v30i1.9864.

Burke, Jason. 2017. "Congo Violence Fuels Fears of Return to 90s Bloodbath." *World News*. *The Guardian*, June 30, 2017. https://www.theguardian.com/world/2017/jun/30/congo-violence-fuels-fears-of-return-to-90s-bloodbath.

Cao, Meng, Yue Dong, and Jackie Cheung. 2022. "Hallucinated but Factual! Inspecting the Factuality of Hallucinations in Abstractive Summarization." *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Dublin, Ireland), Volume 1: *Long papers*, 3340–3354. https://aclanthology.org/2022.acl-long.236.

Chen, John, Xi Lu, Yuzhou Du, Michael Rejtig, Ruth Bagley, Mike Horn, and Uri Wilensky. 2024. "Learning Agent-Based Modeling with LLM Companions: Experiences of Novices and Experts Using ChatGPT & NetLogo Chat." *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA), CHI '24, 1–18. https://doi.org/10.1145/3613904.3642377.

Coeckelbergh, Mark. 2021. "AI for Climate: Freedom, Justice, and Other Ethical and Political Challenges." *AI and Ethics* 1 (1): 67–72. https://doi.org/10.1007/s43681-020-00007-2.

Collins, Harry. 2025. "Why Artificial Intelligence Needs Sociology of Knowledge: Parts I and II." *AI & Society* 40 (3): 1249–1263. https://doi.org/10.1007/s00146-024-01954-8.

Collins, Patricia Hill. 2000. *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. New York: Routledge. https://doi.org/10.4324/9780203900055.

Cowls, Josh, Andreas Tsamados, Mariarosaria Taddeo, and Luciano Floridi. 2023. "The AI Gambit: Leveraging Artificial Intelligence to Combat Climate Change—Opportunities, Challenges, and Recommendations." *AI & Society* 38 (1): 283–307. https://doi.org/10.1007/s00146-021-01294-x.

Crawford, Kate. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.

Crawford, Kate, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas, Amba Kak, Varoon Mathur, Erin McElroy, Andrea Nill Sánchez, Deborah Raji, Joy Lisi Rankin, Rashida Richardson, Jason Schultz, Sarah Myers West, and Meredith Whittaker. 2019. "AI Now 2019 Report." *AI Now Institute,* December 2019. https://ainowinstitute.org/wp-content/uploads/2023/04/AI_Now_2019_Report.pdf.

Danaher, John. 2022. "Techno-Optimism: An Analysis, an Evaluation and a Modest Defence." *Philosophy & Technology* 35 (2): 54. https://doi.org/10.1007/s13347-022-00550-2.

Dauth, Wolfgang, Sebastian Findeisen, Jens Suedekum, and Nicole Woessner. 2021. "The Adjustment of Labor Markets to Robots." *Journal of the European Economic Association* 19 (6): 3104–3113. https://doi.org/10.1093/jeea/jvab012.

Deaton, Angus. 2013. *The Great Escape: Health, Wealth, and the Origins of Inequality*. Princeton: Princeton University Press. https://doi.org/10.1515/9781400847969.

Di Paolo, Ezequiel A., Elena Clare Cuffari, and Hanne De Jaegher. 2023. *Linguistic Bodies: The Continuity Between Life and Language*. Cambridge, MA: The MIT Press.

Di Paolo, Ezequiel and Evan Thompson. 2014. "The Enactive Approach." In *The Routledge Handbook of Embodied Cognition*, edited by Lawrence A. Shapiro, 68–78. London: Routledge.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. 2024. "GPTs Are GPTs: Labor Market Impact Potential of LLMs." *Science* 384 (6702): 1306–1308. https://doi.org/10.1126/science.adj0998.

Endacott, Camille Grace. 2021. *The Work of Identity Construction in the Age of Intelligent Machines*. PhD Thesis, University of California, Santa Barbara, 217 pages. https://escholarship.org/uc/item/2kb6p061.

Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. 2017. "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks." *Nature* 542 (7639): 115–118. https://doi.org/10.1038/nature21056.

Felipe-Redondo, Ana María, Yucels Anaí Del Carmen-Morales, Heidi Del Carmen Morales, and Felipe De Jesús Nuñez-Cárdenas. 2024. "ChatGPT: Impact on Programming Learning among Students at the Technological University of Huasteca Hidalguense." *Revista de Didáctica Práctica* 8 (19). https://doi.org/10.35429/JPD.2024.8.19.1.10.

Floridi, Luciano. 2023. *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*. New York: Oxford University Press.

Frank, Adam, Marcelo Gleiser, and Evan Thompson. 2024. *The Blind Spot: Why Science Cannot Ignore Human Experience*. Cambridge, MA: The MIT Press.

Frankfurt, Harry. 2005. *On Bullshit*. Princeton: Princeton University Press.

Gallagher, Shaun. 2000. "Philosophical Conceptions of the Self: Implications for Cognitive Science." *Trends in Cognitive Sciences* 4 (1): 14–21. https://doi.org/10.1016/S1364-6613(99)01417-5.

García-Martín, Eva, Crefeda Faviola Rodrigues, Graham Riley, and Håkan Grahn. 2019. "Estimation of Energy Consumption in Machine Learning." *Journal of Parallel and Distributed Computing* 134 (December): 75–88. https://doi.org/10.1016/j.jpdc.2019.07.007.

Gebru, Timnit. 2020. "Race and Gender." In *The Oxford Handbook of Ethics of AI*, edited by Markus D. Dubber, Frank Pasquale, and Sunit Das, 252–269. New York: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.16.

Gheaus, Anca and Lisa Herzog. 2016. "The Goods of Work (Other Than Money!)." *Journal of Social Philosophy* 47 (1): 70–89. https://doi.org/10.1111/josp.12140.

Ghosh, Dona, Rajarshi Ghosh, Sahana Roy Chowdhury, and Boudhayan Ganguly. 2025. "AI-Exposure and Labour Market: A Systematic Literature Review on Estimations,

Validations, and Perceptions." *Management Review Quarterly* 75 (1): 677–704. https://doi.org/10.1007/s11301-023-00393-x.

Gozalo-Brizuela, Roberto and Eduardo Garrido Merch'an. 2024. "A Survey of Generative AI Applications." *Journal of Computer Science* 20 (8): 801–818. https://doi.org/10.3844/jcssp.2024.801.818.

Gray, Mary L. and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston: Houghton Mifflin Harcourt.

Hicks, Michael Townsen, James Humphries, and Joe Slater. 2024. "ChatGPT Is Bullshit." *Ethics and Information Technology* 26 (2): 38. https://doi.org/10.1007/s10676-024-09775-5.

Inderwildi, Oliver, Chuan Zhang, Xiaonan Wang, and Markus Kraft. 2020. "The Impact of Intelligent Cyber-Physical Systems on the Decarbonization of Energy." *Energy & Environmental Science* 13 (3): 744–771. https://doi.org/10.1039/C9EE01919G.

Irani, Lilly. 2015. "The Cultural Work of Microwork." *New Media & Society* 17 (5): 720–739. https://doi.org/10.1177/1461444813511926.

Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. "The Global Landscape of AI Ethics Guidelines." *Nature Machine Intelligence* 1 (9): 389–399. https://doi.org/10.1038/s42256-019-0088-2.

Jonas, Hans. 1966. *The Phenomenon of Life: Toward a Philosophical Biology*. New York: Harper & Row.

Kandlhofer, Martin, Gerald Steinbauer, Sabine Hirschmugl-Gaisch, and Petra Huber. 2016. "Artificial Intelligence and Computer Science in Education: From Kindergarten to University." *2016 IEEE Frontiers in Education Conference (FIE)*, October, 1–9. https://doi.org/10.1109/FIE.2016.7757570.

Kant, Immanuel. 2000. *Critique of the Power of Judgment*. Edited and translated by Paul Guyer, translated by Eric Matthews. *The Cambridge Edition of the Works of Immanuel Kant*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511804656.

Kant, Immanuel. 2013. *Lectures on Anthropology*. Edited and translated by Robert B. Louden and Allen W. Wood, translated by Robert R. Clewis and G. Felicitas Munzel. Cambridge: Cambridge University Press.

Kosmyna, Nataliya, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. 2025. "Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task." Preprint, arXiv, December 31, 2025. https://doi.org/10.48550/arXiv.2506.08872.

Kyselo, Miriam. 2014. "The Body Social: An Enactive Approach to the Self." *Frontiers in Psychology* 5: 986. https://doi.org/10.3389/fpsyg.2014.00986.

Kyselo, Miriam. Forthcoming. *The Relational Mind: An Enactive Theory of the Self*. Cambridge, MA: The MIT Press.

Kyselo, Miriam, Lukas R.A. Wilde, Thomas Brandt, Guri Hanem, and Roger A. Søraa. 2026. "Beyond the Hype: Humanities Perspectives on AI." In *Humanities Perspectives on AI: An Interdisciplinary Introduction*, edited by Lukas R.A. Wilde, Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, 11–41. Oslo: Scandinavian University Press.

Landemore, Hélène. 2024. "Can Artificial Intelligence Bring Deliberation to the Masses?" In *Conversations in Philosophy, Law, and Politics*, edited by Ruth Chang and Amia Srinivasan, 39–69. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198864523.003.0003.

Latour, Bruno. 2005. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780199256044.001.0001.

Liu, Zhiyi. 2012. "Chinese Mining Dump Could Hold Trillion-Dollar Rare Earth Deposit." *Dialogue Earth*, December 14, 2012. https://dialogue.earth/en/pollution/5495-chinese-mining-dump-could-hold-trillion-dollar-rare-earth-deposit.

Marx, Karl. [1867] 1981. *Capital: A Critique of Political Economy*. Translated by Ben Fowkes. Vol. 1. London: Penguin.

Marx, Karl. [1939–1941] 1973. *Grundrisse: Foundations of the Critique of Political Economy*. Translated by Martin Nicolaus. London: Penguin.

McAfee, Andrew. 2019. *More from Less: The Surprising Story of How We Learned to Prosper Using Fewer Resources—and What Happens Next*. New York: Scribner.

Mhlambi, Sabelo. 2020. "From Rationality to Relationality: Ubuntu as an Ethical and Human Rights Framework for Artificial Intelligence Governance." *Carr Center for Human Rights Policy Discussion Paper Series*. https://perma.cc/Q5ZL-TTD8.

MIT Media Lab. n.d. "Project Overview 'Your Brain on ChatGPT,'" accessed March 10, 2026. https://www.media.mit.edu/projects/your-brain-on-chatgpt/overview/.

Montgomery, Blake. 2025. "Elon Musk's Grok Chatbot Melts Down—and Then Wins a Military Contract." *Technology*. *The Guardian*, July 15, 2025. https://www.theguardian.com/technology/2025/jul/14/elon-musk-grok-ai-chatbot-x-linda-yaccarino.

Moraes, Carlos Henrique Valério de, Josnei Scolimoski, Germano Lambert-Torres, Mariana Santini, André Luiz Alves Dias, Fábio Alessandro Guerra, André Pedretti, and Milton Pires Ramos. 2022. "Robotic Process Automation and Machine Learning: A Systematic Review." *Brazilian Archives of Biology and Technology* 65: e22220096. https://doi.org/10.1590/1678-4324-2022220096.

Munn, Luke. 2023. "The Uselessness of AI ethics." *AI Ethics* 3: 869–877. https://doi.org/10.1007/s43681-022-00209-w.

Mutanga, Bethel Murimo, Matthews Lecheko, and Zvinodaishe Revesai. 2024. "Navigating the Grey Area: Students' Ethical Dilemmas in Using AI Tools for Coding Assignments." *IJIE (Indonesian Journal of Informatics Education)* 8 (1): 15–24. https://doi.org/10.20961/ijie.v8i1.90385.

Mutanga, Murimo Bethel, Jotham Msane, Thaddeus Ndumiso Mndaweni, Bongokuhle Brightman Hlongwane, and Neliswa Ziyanda Ngcobo. 2025. "Exploring the Impact of LLM Prompting on Students' Learning." *Trends in Higher Education* 4 (3): 31. https://doi.org/10.3390/higheredu4030031.

Naam, Ramez. 2013. *The Infinite Resource: The Power of Ideas on a Finite Planet*. Lebanon: University Press of New England. https://doi.org/10.2307/j.ctv1xx9k8c.

Nagel, Thomas. 1986. *The View from Nowhere*. Oxford: Oxford University Press.

Nelson, Andrew J. and Jennifer Irwin. 2014. "'Defining What We Do—All Over Again': Occupational Identity, Technological Change, and the Librarian/Internet-Search Relationship." *Academy of Management Journal* 57 (3): 892–928. https://doi.org/10.5465/amj.2012.0201.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

OECD. 2024. "Job Creation and Local Economic Development 2024: The Geography of Generative AI." *OECD publishing*, November 28, 2024. https://doi.org/10.1787/83325127-en.

O'Neill, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Publishers.

OpenAI, et al. 2024. "GPT-4 Technical Report." Preprint, arXiv, March 4, 2024. https://doi.org/10.48550/arXiv.2303.08774.

Pasquinelli, Matteo. 2023. *The Eye of the Master*: *A Social History of Artificial Intelligence* London: Verso.

Pinker, Steven. 2018. *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress*. London: Penguin Books.

Reuters. 2025. "Musk's xAI to Provide Grok Chatbot to US Federal Agencies." *Reuters*, September 25, 2025. https://www.reuters.com/world/us/musks-xai-provide-grok-chatbot-us-federal-agencies-2025-09-25.

Rion, Abid Hossain and Mashrufa Meghla Any. 2025. "Will AI Replace Human Jobs? A Literature Review." SSRN Scholarly Paper 5380481. *Social Science Research Network*, August 4, 2025. https://doi.org/10.2139/ssrn.5380481.

Rosling, Hans, Ola Rosling, and Anna Rosling Rönnlund. 2018. *Factfulness: Ten Reasons We're Wrong About the World and Why Things Are Better Than You Think*. London: Sceptre.

Sayed-Mouchaweh, Moamar, ed. 2020. *Artificial Intelligence Techniques for a Scalable Energy Transition: Advanced Methods, Digital Technologies, Decision Support Tools, and Applications*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-42726-9.

Schwarz, Elke. 2025. "Engineering Moral Failure? The Challenges of Algorithmic Ethics for Lethal Autonomous Weapon Systems." In *The Realities of Autonomous Weapons*, edited by Thomas Christian Bächle and Jascha Bareis, 232–258. Bristol: Bristol University Press. https://doi.org/10.56687/9781529237191-016.

Searle, John. 1980. "Minds, Brains, and Programs." *The Behavioral and Brain Sciences* 3 (3): 417–457.

Seele, Peter and Mario D. Schultz. 2022. "From Greenwashing to Machinewashing: A Model and Future Directions Derived from Reasoning by Analogy." *J Bus Ethics* 178: 1063–1089. https://doi.org/10.1007/s10551-022-05054-9.

Senders, Joeky T., Omar Arnaout, Aditya V. Karhade, Hormuzdiyar H. Dasenbrock, William B. Gormley, Marike L. Broekman, and Timothy R. Smith. 2018. "Natural and Artificial Intelligence in Neurosurgery: A Systematic Review." *Neurosurgery* 83 (2): 181–192. https://doi.org/10.1093/neuros/nyx384.

Smids, Jilles, Sven Nyholm, and Hannah Berkers. 2020. "Robots in the Workplace: A Threat to—or Opportunity for—Meaningful Work?" *Philosophy & Technology* 33 (3): 503–522. https://doi.org/10.1007/s13347-019-00377-4.

Smuha, Nathalie A. 2025. "An Introduction to the Law, Ethics, and Policy of Artificial Intelligence." In *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence*, edited by Nathalie A. Smuha, 1–14. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781009367783.001.

Soll, Marcus and Louis Kobras. 2025. "Do Large Language Models Require Prior Knowledge for Learning? A Preliminary Study." *2025 IEEE Global Engineering Education Conference (EDUCON)*, April 1–5, 2025. https://doi.org/10.1109/EDUCON62633.2025.11016655.

Spencer, David A. 2022. *Making Light Work: An End to Toil in the Twenty-First Century*. Cambridge: Polity.

Spencer, David A. 2025. "AI, Automation and the Lightening of Work." *AI & Society* 40 (3): 1237–1247. https://doi.org/10.1007/s00146-024-01959-3.

Srnicek, Nick and Alex Williams. 2015. *Inventing the Future: Postcapitalism and a World Without Work*. London: Verso.

Sui, Peiqi, Eamon Duede, Sophie Wu, and Richard So. 2024. "Confabulation: The Surprising Value of Large Language Model Hallucinations." *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Bangkok, Thailand), Volume 1: Long Papers: 14274–14284. https://aclanthology.org/2024.acl-long.770.

Søraa, Roger A. and Kristine Ask. 2026. "AI and Equality." In *Humanities Perspectives on AI: An Interdisciplinary Introduction*, edited by Lukas R.A. Wilde, Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, 111–126. Oslo: Scandinavian University Press.

Taddeo, Mariarosaria. 2024. *The Ethics of Artificial Intelligence in Defence*. Oxford: Oxford University Press.

Tasioulas, John. 2023. "The Rule of Algorithm and the Rule of Law." SSRN Scholarly Paper 4319969. *Social Science Research Network*, January 7, 2023. https://doi.org/10.2139/ssrn.4319969.

Torbjørnsen, Regine Rørstad and Miriam Kyselo. 2026. "An Introduction to the Philosophy of Artificial Intelligence." In *Humanities Perspectives on AI: An Interdisciplinary Introduction*, edited by Lukas R.A. Wilde, Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, 63–85. Oslo: Scandinavian University Press.

Tyson, Laura D. and John Zysman. 2022. "Automation, AI & Work." *Daedalus* 151 (2): 256–271. https://doi.org/10.1162/daed_a_01914.

UNESCO. 2022. "Recommendation on the Ethics of Artificial Intelligence." Preprint, *UNESCO Digital Library*. https://unesdoc.unesco.org/ark:/48223/pf0000381137.locale=en.

van Maanen. 2022. "AI Ethics, Ethics Washing, and the Need to Politicize Data Ethics." *Digital Society* 1 (9). https://doi.org/10.1007/s44206-022-00013-3.

Véliz, Carissa. 2024. *The Ethics of Privacy and Surveillance*. Oxford: Oxford University Press.

Walton, Nigel and Bhabani Shankar Nayak. 2021. "Rethinking of Marxist Perspectives on Big Data, Artificial Intelligence (AI) and Capitalist Economic Development." *Technological Forecasting and Social Change* 166 (May): 120576. https://doi.org/10.1016/j.techfore.2021.120576.

Wang, Dayong, Aditya Khosla, Rishab Gargeya, Humayun Irshad, Andrew H. Beck. 2016. "Deep Learning for Identifying Metastatic Breast Cancer." Preprint, arXiv, June 18, 2016. https://doi.org/10.48550/arXiv.1606.05718.

Weise, Karen and Cade Metz. 2023. "When A.I. Chatbots Hallucinate." *Business. The New York Times*, May 1, 2023. https://www.nytimes.com/2023/05/01/business/ai-chatbots-hallucination.html.

Zamfirescu-Pereira, J.D., Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. "Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts." *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA), CHI '23, April 19, 2023: 1–21. https://doi.org/10.1145/3544548.3581388.

Zhang, Marina Yue. 2025. "The China-US AI Race Enters a New (and More Dangerous) Phase." *The Diplomat*, May 16, 2025. https://thediplomat.com/2025/05/the-china-us-ai-race-enters-a-new-and-more-dangerous-phase.

5. AI and Equality

Roger A. Søraa and Kristine Ask

Abstract: This chapter, rooted in Science and Technology Studies and feminist technoscience perspectives, thematizes who is included in and excluded from the development and use of AI. One central organizing framework is the concept of a *digital divide* between participants in the social world, applied here to AI at three levels: access—who can use AI at all; skills—who has the literacy to use it effectively; and outcomes—who benefits and who is harmed, even among those with equal access and skills. The chapter addresses contemporary examples ranging from AI facial recognition technology, AI used within hiring processes, and AI used by people with disabilities to explain the value of such a sociotechnical perspective. Opening the "black boxes" of AI partially, the chapter provides thinking tools on how to analyze discrimination and ensure fair AI use.

Keywords: algorithmic bias, digital divide, feminist technoscience, inclusion, Science and Technology Studies (STS)

In this chapter, we explore how artificial intelligence (AI) relates to equality, focusing on how AI can either support or exacerbate the inclusion or exclusion of vulnerable groups. Grounded in feminist technoscience perspectives from the field of Science and Technology Studies (STS), the chapter seeks to open AI's "black boxes" (where only input and output are known and the processing is invisible) to examine who is affected by AI and in what ways. Our overall argument is that we must pay attention to how new technology (like AI) can strengthen or weaken inequality depending on its design and use. We illustrate our discussion with three examples from AI research: firstly, facial recognition technology and racial profiling; secondly, how AI is used in hiring processes; and thirdly, how AI is empowering users with disabilities. In these examples, we aim to highlight both positive and negative ways in which AI impacts equality. To do this, it is important to pay attention to both the technology (its features, design, and capabilities) and its use (by whom, for what, and in what contexts)—an approach we can broadly describe as *sociotechnical*.

Wilde, Lukas R.A., Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, eds. 2026. *Humanities Perspectives on AI: An Interdisciplinary Introduction*. Scandinavian University Press. https://doi.org/10.18261/9788294167234-26-05
Open access | © The author(s) 2026. This work is licensed under CC BY 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/.

A Sociotechnical View on AI: From Neutral Technology to Digital Divides

A sociotechnical approach acknowledges that technological change is as much a question of people and society as it is of design and functionality. Instead of presuming that technology is a neutral object with predetermined consequences or trajectories of development, the sociotechnical perspective is concerned with how technology is shaped by social factors. This includes how society interprets the technology, the imaginaries it holds, the skills of its users, and the context in which it is applied. In short, a sociotechnical approach means paying attention to culture, user practices, infrastructures and institutions that shape how a given technology is interpreted and the specific ways in which it is used (Oudshoorn and Pinch 2003; Pinch and Bijker 1987).

In the case of AI, there exists a multitude of visions about what life with AI would entail, not least from science fiction; for example, Lana and Lilly Wachowski's (1999) *The Matrix*, presenting a dystopian future where machines have taken over the planet and turned humans into batteries to power the machine world, forcing them to remain asleep and live in simulated realities instead. Similarly dystopian futures of AI rebelling against humans are shown in James Cameron's (1985) *The Terminator* and Joss Whedon's (2015) *Avengers: Age of Ultron*. These movies both fuel and reflect our fear of technology taking over our lives and explore what kind of risk AI may pose to societies and humanity overall. At the same time, there is also more optimistic science fiction about what an AI society might entail. In Andrew Stanton's (2008) movie *WALL-E*, the robots WALL-E and EVE protect humans and assist their return to Earth. In Gene Roddenberry et al.'s (1987–1991) utopian sci-fi television show *Star Trek: The Next Generation*, we repeatedly meet artificial intelligence in characters like Data, who seeks to become *more* human and embarks on personal quests to grow more empathetic, caring, and artistic. In short, the stories we have created about AI span both hopes and fears of what technology can do for humanity (Cave and Dihal 2019).

While many narratives about AI are fictional, they are far from irrelevant. These stories are invoked by actual developers and producers of AI in both the making and promotion of new technology: for example, when OpenAI's GPT-4o was released with a voice feature (later withdrawn) that sounded suspiciously like Scarlett Johansson, who voiced an AI interface in Spike Jonze's movie *Her* (2013), or when discussions about the use of AI in the military employ *The Terminator* as a frame of reference (Spangler 2024). The point of opening this article with such stories about AI is to highlight the fact that

technology is always *more* than just its design, features, or technical specifications. Technology is interpreted in different ways, and we create both fictional and factual stories about its meanings and purpose.

As we grapple with what an "AI society" entails, a range of competing narratives have emerged in the news media. While some frame AI as technology based on massive theft of copyrighted art (Milmo 2025), others characterize it as a democratizing creative instrument (Duggal 2024). In education, some see AI as a tool for a much-needed revolution (Molina et al. 2024), while others interpret AI as the final nail in the coffin for an educational system in trouble (Bogost 2024). There are news stories about AI's superior intelligence surpassing the capabilities of human doctors (Park 2025), and news stories about AI being unable to do simple tasks like knowing how many R's there are in strawberries (Eaton 2024). These conflicting narratives about AI make it difficult to gauge the technology's actual capabilities and effects, and this chapter is too short to examine in depth the accuracy of each claim. Collectively, however, all these reports highlight the fact that there is no singular, shared vision of what AI *is.* They also showcase the sociotechnical perspective on technology, where technology is *more* than its functionality, its design, and its programming. AI is also about values and morals, about how we think the world works, and how we think it *should* work—which is particularly important when we look at how AI impacts (in)equality.

Overall, the point of showing different imaginaries, visions, and stories about AI is to direct attention to how important our interpretation of technology is. We tend to think of technology as value-neutral: that humans are the ones with biases and morality while technology is "just a tool" with no politics. However, decades of studies have shown that technology can work as infrastructures that maintain or shift power balances in society by privileging some users while exposing others to risks. Consider, for example, how safety technology in cars (such as airbags and seat belts) was modeled after male users, leading to cars that were safer for men than for women (Criado-Perez 2020); or how early photography equipment was not properly calibrated for all skin tones, because only White skin was seen as important (Roth 2009). In all these examples, it is clear that technology is not a "neutral party," but is instead imbued with values, politics, and preferences that favor some people over others.

In the study of digital technology, the way digitalization can cause differences in society is broadly referred to as "digital divides" (van Dijk 2005), which span three different levels: access, skills, and outcomes (as described by Lutz 2019; 2024). The first level, *access* to digital technology, initially addressed

how those without access to the internet were missing out on the benefits associated with the technology. Later, this discussion expanded to include analyses of how countries that rely primarily on mobile phones rather than broadband connections experience reduced functionality and limited content, with usage patterns oriented more toward entertainment and socializing than toward software development. Today, the question of who has access to AI represents a similar potential divide, where economic status, in particular, has an impact on the type of AI technology one has access to (for example, free versus premium AI models).

The second level divide addresses *skills* and recognizes that the benefits provided by digitalization are reliant on more than access; they also depend on the digital literacy of users. As digital literacy is not equally distributed in society, digitalization may exacerbate differences between groups once more. Age, gender, and education are important factors here, with young, educated men scoring higher on average and thus being better positioned to take advantage of opportunities related to digitalization. Similar effects are expected from differing levels of AI literacy, although what exactly AI literacy should entail is still up for debate.

On the third level of the digital divide, the focus is on *outcomes*. The key to this understanding is that digitalization can create feedback loops, where those with privilege and power gain even more of both, and those marginalized face further risks and marginalization. Even with equal access and skills, the outcomes of digitalization are not equal. Instead, the distribution of risks and opportunities tends to follow established axes of power in society, meaning that, generally speaking, users who are male (rather than female), White (rather than People of Color),[1] heterosexual (rather than LGBTQI+), able-bodied (rather than a person with disabilities), and rich (rather than poor) are better off. One example is how low-status jobs (often staffed by immigrants) are more subject to surveillance through digital technologies (Glaser 2021; Pellow and Park 2002). In the following section on AI and diversity, we will focus on this third level divide and show how the potential

1 In this chapter, we have chosen to capitalize the terminologies for Black, Indigenous, and People of Color (often abbreviated as BIPoC) in order to adhere to current scholarly debate on systematic discrimination faced by BIPoC people in post-colonial societies both historically and contemporarily. Race and ethnicity are extremely complicated topics. Rather than providing one "correct" definition, we encourage readers to familiarize themselves with current discourse on these terminologies when discussing these topics in their own works.

outcomes are related to existing power structures, even for those with access to and skills in AI use.

Differentiating Users: AI and Diversity

When we talk about inclusion and exclusion in the context of AI, we often refer to *diversity*. The term diversity broadly refers to differences. When it is used to describe human society, it describes variations among individuals and groups; the categories they are placed in (by both society and themselves); and the effects this might have. For instance, a family may consist of people from multiple generations, demonstrating diversity in age. Similarly, a university embodies diversity through its students and staff who come from different geographic regions (both domestic and international), urban and rural backgrounds, and socio-economic statuses. Additionally, people might have diverse beliefs, values, and motivations. Thus, diversity encompasses both individual and group-level differences. When investigating questions of equality and technology, we tend to focus on group-level differences, as existing power structures in society tend to be reflected in technology use.

Consider the digital divide presented above and note how the example of access was related to urban–rural divides, how the example of skills was related to gender divides, and how the example of outcomes was based on skin color. This is not accidental. Again and again, we find that the hierarchies of society are reproduced and reflected through technology and its uses. Regardless of whether you think such categorizations of people are sensible or fair (they are often not), they are impactful. The same is true when it comes to AI. As probability machines, AI rely on being able to identify and replicate patterns observed in society, such as those based on age, nationality, gender, and skin color, making such categories highly important in processing and predictions. Humans, however, are far more complex than the categories we use to classify one another.

Consider the complex concepts of ethnicity, skin color, and nationality. In Norway, we tend to treat these as separate categories; while in other contexts, they are categorized together under the heading "race." It is, for example, common for US citizens to indicate their "race" on job applications. While some individuals may use "White" or "African-American" and find those categories accurate, others may find classification systems misleading. Take, for example, the actor Keanu Reeves, famous for portraying Neo in *The Matrix*. Reeves' mother is from Essex, England, while his father was born in Hawaii with

Chinese and Native Hawaiian heritage. For individuals like Reeves, checking a single box in a demographic survey or hypothetical job application is not an accurate reflection of their ethnic and racial background.

Similarly, other seemingly straightforward categories, such as gender or sexual orientation, may also be difficult to neatly classify, particularly for an AI system attempting to interpret social dynamics, as people can change their gender identity and sexual orientation over their lifetime. We are all multifaceted individuals, navigating multiple identities and categorizations that shift according to context. This presents challenges to AI. Computational systems are fundamentally designed to structure and categorize information. For a computer, simplicity often leads to better functionality, but humans and the social world are anything but simple. The fact that the social world cannot be neatly categorized presents some unique challenges for implementing AI. As the models attempt to predict outcomes through machine-based pattern recognition, such classification systems are necessary for AI to function effectively; but as they are inaccurate reflections of reality, they are thus also a continuous source of errors.

Some errors from AI are quite comical, like a system struggling to distinguish Chihuahuas from muffins (Lubin 2016), when robot vacuum cleaners do not stop for dog poo but instead do a "shitty job" of dragging it along the floor (Solon 2016), or when an American mall security robot drove into a fountain and "killed itself" one fateful night (Kirschgens et al. 2018). In all these cases, the underlying problems consisted of image recognition systems that were unable to correctly classify the real world. From a machine's perspective, chihuahuas and muffins are similar as they have light and dark pixels in comparable places. Similarly, the AI system will not "know" when to change its path because the system may be missing training data that could be used to identify dog poo or fountains as "no go" zones. While these examples are more annoying than problematic, errors based on the inability of AI to correctly parse real-world complexity into abstract categories and AI classifications can also have serious and harmful consequences.

Predictive policing and law enforcement in the US have been shown to systematically target People of Color, causing a feedback loop where these are overrepresented in crime statistics (Ensign et al. 2018), which again leads to even more policing. Being unable to correctly identify obstacles might be funny for a mall security robot, but it presents a lethal risk when the technology is implemented in self-driving cars, leading to several passengers and pedestrians being harmed or killed already (Stilgoe 2018). Several other examples can be found in Søraa's (2023) book on *AI for Diversity*, which also describes AI discrimination as

primarily occurring in three stages. These three stages are: input (before the data is processed), the data processing itself, and output (after the data is processed):

1. **Input Bias**: This first stage concerns the datasets used to train an AI system. Since AI reproduces patterns identified in the dataset, problems with fair and equal representation will impact the quality of output. Simply put, not all groups in society have had an equal chance to see their perspectives, stories, and lives documented. Thus, the training data will be skewed. Consider, for example, how the book publishing industry has (historically speaking, at least) favored straight, White, male authors. So, when LLMs are trained on all books found online (including those that should have been protected through copyright), they are being trained on texts that overrepresent straight, White, male perspectives on the world.
2. **Data Processing Bias**: Even within an AI system itself, bias can emerge by how algorithms process data. As a machine identifies patterns in the datasets and turns these patterns into "rules" for how to process them, unequal representation from the first stage can become "rules" that reproduce and reinforce those inequalities. For example, if people living in low-income areas often have their loan applications denied, the AI might "interpret" this as being about living in that specific area, not about income, and automatically reject all applicants living in similar areas regardless of their actual income. This is also where it can be notoriously difficult to even understand how computers have made decisions, as the algorithmic processing of data is often hidden in these systems—and might not even be explainable by the developers themselves (see also Kundi et al. 2023).
3. **Output Bias**: Finally, once an AI system has processed the input data and generated results, discrimination can also occur in how these outputs are interpreted and utilized by users. This means we need to pay attention to the *context* in which AI is used, and how this shapes how AI outputs and services are utilized. This final stage involves the ways in which and degrees to which AI-generated information is applied in real-world scenarios. A user who trusts AI might find its recommendations convincing and intriguing, while a user who is more critical might reject the model's recommendations outright. For example, Binns et al. (2018) found in their study on algorithmic decision-making that participants who were more trusting of algorithms were significantly more likely to

accept unfair or biased recommendations in hiring scenarios, compared to those who approached AI systems with fair skepticism. The latter group tended to question the legitimacy and fairness of the model's outputs.

These three types of biases can be conceptualized as part of what goes in, what is processed inside (that is, the how), and what comes out of a "black box" (see figure 5.1):

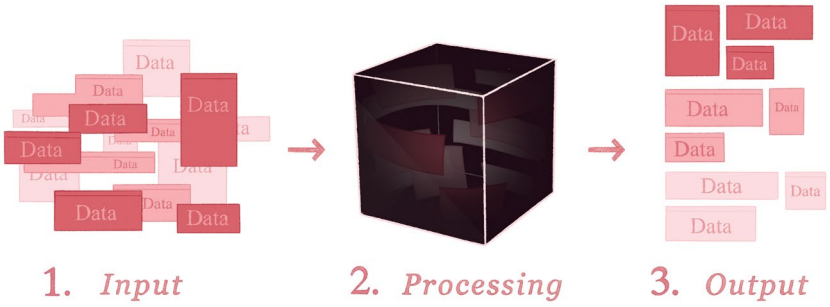

Figure 5.1: Three types of biases conceptualized as a Black Box with (1) input, (2) processing, and (3) output.

If we are going to use AI to automate (which is often the goal of implementing AI in the first place), such biases in the training data are important to identify and counter to avoid duplicating and reinforcing existing societal problems. Obtaining unbiased datasets may seem to be a simple solution to this problem, but given the size requirements for datasets, this is a rather huge undertaking. In reality, most models work on data that is openly available on the internet (both legally and illegally), which will then reflect biases and inequalities that already exist in the world. To further explore biases and AI, we turn to three examples from research that will illustrate bias in input, in processing, and in output.

Racial Profiling in Facial Recognition Systems

As the name states, the primary function of facial recognition technology (FRT) is to correctly identify and classify people's faces. FRT must always be trained on large datasets—so-called *big data*—to make accurate predictions of who different facial images belong to. A major dilemma with facial recognition, however, which is illustrative of issues for AI in general, is that it is only as good as the training data permits. If the training data is predominantly made

up of White faces, and severely underrepresents faces belonging to People of Color, for instance, the accuracy for the latter groups dramatically decreases. It then fails to correctly identify individuals with skin pigmentations other than White (input bias). This was precisely an issue that Joy Buolamwini faced when trying to get FRT to work well on her own face. As a scholar working at MIT, building FRT systems herself, Buolamwini found that several FRT systems only worked on her if she donned a white paper mask over her Black skin. For further study, she decided to test FRT using photos of members of parliament from different countries where skin pigmentation grades varied, ranging from Nordic countries with very light skin pigmentation to African countries with very dark skin pigmentation. Examples of this can be seen on the gendershades.org website (Buolamwini and Gebru 2018).

Buolamwini found that the darker the skin pigmentation, the more difficulty FRT had in correctly identifying a person's face. Furthermore, she found that women were overall much less likely to be correctly identified compared to their male counterparts, with White women less likely to be correctly identified than White men (who overall seemed to always be correctly identified), and Black women less likely to be identified correctly than Black men. Overall, Black women had up to 34.7% error rates (meaning that in one in three times, the classification fails to work) compared to just a 0.8% error rate for White men (data processing bias).

FRT is not the only AI system trying to analyze skin tone that leans toward inadequate or even racist results. Soap and disinfection dispensers have been shown to work well on White hands, but to struggle to correctly register Black hands (Broussard 2023). Other examples include AI beauty contests judging White human faces to be more attractive than BIPoC human faces (Søraa 2023). Such disparities can partly be explained by looking at the datasets used to train FRT, where people with light skin are overrepresented. But, while biases in the training data reinforce unequal outcomes, the black boxing of the processes that AI uses to arrive at conclusions also plays a major role, despite being potentially challenging to interrogate.

Diversity and Discrimination in the Job Search Process

Hiring processes are notoriously influenced by personal bias among recruiters. Candidates with identical qualifications are often evaluated differently depending on whether their names align with the ethnic majority in a given country. In Norway, for example, Midtbøen (2016) shows that candidates with

"traditional Norwegian names" have a significantly higher chance of being invited to an interview than people with names that appear foreign even when their CVs are identical. Moreover, factors such as gender, age, and disability status (e.g., using a wheelchair or not) can also influence the likelihood of being shortlisted for interview (Birkelund et al. 2022; Ravaud 1992). Introducing AI into the hiring process is often understood as a way to counter such biases. In reality, it may actually end up reinforcing those same biases unless specific measures are taken.

The primary rationale for using AI in recruitment is the large volume of applicants for a given job opening, which creates a significant administrative burden for hiring managers. Filtering out a high percentage of unqualified candidates early in the process is intended to streamline recruitment and thus reduce the workload, and this is often promoted as a "fairer" and more "objective" approach to hiring processes that can help organizations with more diverse recruitment. On the surface, these tools may appear to counter the highly subjective assessment taking place during hiring processes. However, because the social world and its classifications are highly complex, removing such bias can be quite challenging. Even if a model aims to exclude gender, for example, from the processing of applications, it is still possible for candidates to be treated unfairly due to *other* gendered patterns that the model may have picked up on (Drage and Mackereth 2022). It could be, for example, that the model favors male applicants, not because "male" or "man" is written anywhere in the application or processed by the model, but because applicants who list "football" as a hobby are favored by the AI—because the company has, in the past, hired more men than women, and football is a common interest among men (processing bias). In this way, *indirect indicators* of gender, age, socioeconomic background, or ethnicity can still come into play. Unless we open the "black box" of AI to consider how such connections are made by the model and turned into recommendations, AI can further entrench differences between groups.

Algorithmic bias in recruitment has become a bit of a "textbook" case after Amazon attempted to automate its hiring process in 2014 but had to shut it down in 2015 as its algorithm kept suggesting hiring only men. The fault was again found in the datasets, which reflected a hiring history of mostly men. It was shown that the computational processing of applications grouped and valued attributes in undesirable ways. For example, the algorithm perceived the term "woman" as being so detrimental (after all, according to the data, being a woman did not get you hired) that being the leader of the women's chess team did not land the applicant an interview (Criado-Perez 2020).

The case of algorithmic bias in recruitment and hiring processes exemplifies the need to not let AI become black boxes where the algorithm makes recommendations without transparency. In the European Commission-funded project BIAS: Mitigating Diversity Biases of AI in the Labor Market,[2] researchers are working to "de-bias" such algorithms by identifying how they couple words with each other, and thus reduce biased impacts.[3]

In addition, this issue also exemplifies the importance of regulations. While recruitment and application processes were, at the time of writing, listed as "high risk" under the European AI Act (EU Artificial Intelligence Act 2024), which regulates the use of AI for all of the EU. Such regulation puts pressure on developers of AI tools for recruitment and shows how social factors (in this case, regulation) influence technology development and uses—and consequently the need for a sociotechnical perspective to understand AI in society.

Assistive Technologies and the Complexities of Everyday Life in AI for Disability

In this third and final example, we turn to output bias and the relevance of *use contexts* to further understand how AI can both include and exclude people. As a salient case study for this, we draw on AI systems used as assistive devices for disabled people to emphasize that, even with access and skills, the outcome of AI depends on its context. How the context of use shapes outcomes is well illustrated by Peter and Laura Smith (2021), based on diaries of their own lives with assistive AI.

Peter Smith had an accident and broke his neck, limiting his mobility and making him dependent on help to feed himself, walk, and perform other everyday tasks. From his diary, we find several stories about how he uses AI for assistance, but also about how reliant he is on his personal carers for the AI technology to work. He started the day using speech-activated software to phone his doctor, but when no one at the doctor's office picked up, he was unable to end the call and had to stay on the line until his carer returned to switch off the mobile. Similarly, Laura Smith, who is blind and uses various AI technologies to navigate her everyday life, found that human skill must often

---

2 The BIAS project, grant agreement number 101070468, was funded by the European Commission, and has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI).

3 For more insights into how AI discriminates against jobseekers, visit www.biasproject.eu.

supplement technology. When selecting cereal for breakfast, for example, she took a photo of the cereal box and used an app to identify which one it was. However, it took her six attempts of taking photos from different angles for the app to finally recognize the image and let her know it was the wrong one. After three unsuccessful tries on the next box she was holding, she decided to ask her partner instead which cereal was right.

At other times, the sheer complexity and messiness of everyday life is too much for the AI technology to process, and errors occur even more frequently; for example, Peter Smith uses voice assistance to ask his home hub to play a song, but the wrong song starts playing. Since the music is playing too loud for the voice recognition to hear him say "stop," he has to wait until the song is over before he can ask for the right song. Similarly, Laura Smith uses an image recognition app when reading with her daughter. However, when her daughter points to a page in a book and asks, "what is that?" the app cannot make sense of the picture she has taken, because her daughter's finger is overlaying the image she is pointing at and the AI is unable to separate the finger from the image.

These stories show that automated systems are fickle, and that these machines frequently fail when not being used in the specific way they were designed (which can bear little resemblance to actual use conditions), or when encountering situations that are not in the training data (simply because everyday life is messy like that). Machines are still struggling to process the complexities of the human world, and they often tend to rely on other humans doing the work of "making it work." If Peter and Laura Smith did not have competent carers (paid for through welfare programs) and supportive partners, many of these assistive AI technologies would be unusable and have little to no positive impact on their lives. This serves as a reminder that the context for AI use matters.

For Peter and Laura Smith, AI gives them tools to navigate their everyday lives and, in many cases, gives them independence (from carers) by mitigating the impact of their disability (through features such as speech-to-text, text-to-speech, and image recognition software). Such cases exemplify how AI-automated systems that are integrated into our lives can empower us and be important tools for a more diverse and inclusive society. At the same time, in the closing statement of their paper, both of them argue that people with disabilities need to be part of design teams for future assistive AI tools, or at least consulted more, to ensure that these tools are designed in ways that suit the lives of disabled people. This call for action mirrors many other publications on technology and diversity that emphasize how important it is to include the actual users of technology in the design process.

Conclusion

In this chapter, we have examined how artificial intelligence (AI) can be understood and analyzed from an equality perspective, drawing mainly on Science and Technology Studies (STS) and feminist technoscience. We have seen how AI is not neutral; it is shaped by sociotechnical contexts. We have explored how biases emerge at three key stages: input (data), (algorithmic) processing, and (context for) output uses. With these three areas, we have shown how bias and inequality can emerge from AI use or be further reinforced. Firstly, we have discussed how facial recognition suffers from bias in input data, leading to unequal outcomes for people with darker skin. Secondly, we have shown how the algorithmic processing of job applications can lead to bias as unintended patterns are automated. Thirdly, we considered how the context of use shapes outcomes of AI—and that AI is often still reliant on people in addition to the user to function properly.

This chapter also discusses strategies for designing AI systems that mitigate bias, emphasizing the importance of representative datasets, transparent algorithmic processes, and more diverse development teams. We thus end this chapter with a set of questions that should always be kept in mind when studying or discussing AI. Questions to ask in a sociotechnical study of AI inclusion and exclusion may include:

1. Who made the AI—and why have they made it the way they did?
2. Who are the intended users of the AI—and who does not fit the bill?
3. What training data is used as input—and who is excluded from this data?
4. How is the AI processing the data—are the procedures transparent?
5. How is the outcome of the data output interpreted when released back to society?
6. How can AI be improved to further include diverse groups instead of excluding them?

Works Cited

Binns, Reuben D., Michael Veale, Max G. Van-Kleek, Ulrik Lyngs, Jun Zhao, and Nigel R. Shadbolt. 2018. "'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions." *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems,* 377: 1–14. https://doi.org/10.1145/3173574.3173951.

Birkelund, Gunn Elisabeth, Bram Lancee, Edvard Nergård Larsen, Javier G. Polavieja, Jonas Radl, and Ruta Yemane. 2022. "Gender Discrimination in Hiring: Evidence from a Cross-national

Harmonized Field Experiment." *European Sociological Review* 38 (3): 337–354. https://doi.org/10.1093/esr/jcab043.

Bogost, Ian. 2024. "AI Cheating Is Getting Worse." *The Atlantic,* August 19, 2024. https://www.theatlantic.com/technology/archive/2024/08/another-year-ai-college-cheating/679502/.

Broussard, Meredith. 2023. *More than a Glitch: Confronting Race, Gender, and Ability Bias in Tech*. Cambridge, MA: The MIT Press.

Buolamwini, Joy and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." *Conference on Fairness, Accountability and Transparency*, *Proceedings of Machine Learning Research* 81: 77–91.

Cameron, James, dir. 1985. *The Terminator.* Hemdale Film Corporation & Pacific Western Productions.

Cave, Stephen and Kanta Dihal. 2019. "Hopes and Fears for Intelligent Machines in Fiction and Reality." *Nature Machine Intelligence* 1 (2): 74–78. https://doi.org/10.1038/s42256-019-0020-9.

Criado-Perez, Caroline. 2020. *Invisible Women: Exposing Data Bias in a World Designed for Men.* London: Vintage.

Drage, Eleanor and Kerry Mackereth. 2022. "Does AI Debias Recruitment? Race, Gender, and AI's 'Eradication of Difference.'" *Philosophy & Technology 35* (4): 89. https://doi.org/10.1007/s13347-022-00543-1.

Duggal, Sachin Dev. 2024. "Democratized Creativity: The Evolution and Impact of AI." *Forbes*, April 9, 2024. https://www.forbes.com/councils/forbestechcouncil/2024/04/09/democratized-creativity-the-evolution-and-impact-of-ai.

Eaton, Kit. 2024. "How Many R's in 'Strawberry'? This AI Doesn't Know." *Inc,* August 28, 2024. https://www.inc.com/kit-eaton/how-many-rs-in-strawberry-this-ai-cant-tell-you.html.

Ensign, Danielle, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. "Runaway Feedback Loops in Predictive Policing." *Conference on Fairness, Accountability and Transparency*, *Proceedings of Machine Learning Research* 81: 160–171. https://proceedings.mlr.press/v81/ensign18a.html.

EU Artificial Intelligence Act. 2024. "High-level Summary of the AI Act." February 27, 2024. https://artificialintelligenceact.eu/high-level-summary.

Glaser, Alana Lee. 2021. "Uberized Care: Employment Status, Surveillance, and Technological Erasure in the Home Health Care Sector." *Anthropology of Work Review* 42 (1): 24–34. https://doi.org/10.1111/awr.12215.

Jonze, Spike, dir. 2013. *Her.* Annapurna Pictures.

Kirschgens, Laura Alzola, Irati Zamalloa Ugarte, Endika Gil Uriarte, Aday Muñiz Rosas, and Víctor Mayoral Vilches. 2018. "Robot Hazards: From Safety to Security." Preprint, arXiv, June 11, 2018. https://doi.org/10.48550/arXiv.1806.06681.

Kundi, Bushra, Christo El Morr, Rachel Gorman, and Ena Dua. 2023. "Artificial Intelligence and Bias: A Scoping Review." In *AI and Society*: *Tensions and Opportunities*, edited by Christo El Morr, 199–215. London: Chapman and Hall.

Lubin, Rhian. 2016. "Are these Muffins or Chihuahuas? Bizarre Picture Quiz Taking the Internet by Storm." *Mirror*. March 11, 2016. https://www.mirror.co.uk/news/weird-news/muffins-chihuahuas-bizarre-picture-quiz-7539743.

Lutz, Christoph. 2019. "Digital Inequalities in the Age of Artificial Intelligence and Big Data." *Human Behavior and Emerging Technologies* 1 (2): 141–148. https://doi.org/10.1002/hbe2.140.

Lutz, Christoph. 2024. "Social Inequalities and Artificial Intelligence." In *Algorithms, Artificial Intelligence and Beyond*, edited by Dariusz Brzeziński, Kamil Filipek, Kuba Piwowar, and Malgorzata Winiarska-Brodowska, 193–210. New York: Routledge. https://doi.org/10.4324/9781032646930-17.

Midtbøen, Arnfinn H. 2016. "Discrimination of the Second Generation: Evidence from a Field Experiment in Norway." *Journal of International Migration and Integration* 17 (2): 253–272. https://doi.org/10.1007/s12134-014-0406-9.

Milmo, Dan. 2025. "Mass Theft: Thousands of Artists Call for AI Art Auction to be Cancelled." *The Guardian,* February 10, 2025. https://www.theguardian.com/technology/2025/feb/10/mass-theft-thousands-of-artists-call-for-ai-art-auction-to-be-cancelled.

Molina, Ezequiel, Cristobal Cobo, Jasmine Pineda, and Helena Rovner. 2024. "Artificial Intelligence Revolution in Education: What You Need to Know." *Digital Innovations in Education: World Bank*, June 17, 2024. https://www.worldbank.org/en/region/lac/publication/innovaciones-digitales-para-la-educacion-en-america-latina.

Oudshoorn, Nelly and Trevor Pinch. 2003. "How Users and Non-Users Matter." In *How Users Matter: The Co-construction of Users and Technology*, edited by Nelly Oudshoorn and Trevor Pinch, 1–25. Cambridge, MA: The MIT Press. https://doi.org/10.7551/mitpress/3592.003.0002.

Park, Alice. 2025. "Microsoft's AI is Better than Doctors at Diagnosing Disease." *TIME*, July 2, 2025. https://time.com/7299314/microsoft-ai-better-than-doctors-diagnosis.

Pellow, David and Lisa Sun-Hee Park. 2002. *The Silicon Valley of Dreams: Environmental Injustice, Immigrant Workers, and the High-Tech Global Economy*. New York: New York University Press.

Pinch, Trevor J. and Wiebe E. Bijker. 1987. "The Social Construction of Facts and Artifacts: Or how the Sociology of Science and the Sociology of Technology Might Benefit Each Other." In *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology,* edited by Wiebe E. Bijker, Thomas Parke Hughes, and Trevor J. Pinch, 17–50. Cambridge, MA: The MIT Press. https://doi.org/10.1177/0306312840140030.

Ravaud, Jean-François, Béatrice Madiot, and Isabelle Ville. 1992. "Discrimination towards Disabled People Seeking Employment." *Social Science & Medicine* 35 (8): 951–958. https://doi.org/10.1016/0277-9536(92)90234-h.

Roddenberry, Gene, Maurice Hurley, Michael Wagner, Michael Piller, and Jeri Taylor, creators. 1987–1991, 7 seasons. *Star Trek: The Next Generation*. Paramount Television.

Roth, Lorna. 2009. "Looking at Shirley, the Ultimate Norm: Colour Balance, Image Technologies, and Cognitive Equity." *Canadian Journal of Communication* 34 (1): 111–136. https://doi.org/10.22230/cjc.2009v34n1a2196.

Smith, Peter and Laura Smith. 2021. "Artificial Intelligence and Disability: Too much Promise, yet too Little Substance?" *AI and Ethics* 1 (1): 81–86. https://doi.org/10.1007/s43681-020-00004-5.

Solon, Olivia. 2016. "Roomba Robot Vacuum's 'Poopocalypse' Goes Viral after Facebook Post." *The Guardian*. August 15, 2016. https://www.theguardian.com/technology/2016/aug/15/roomba-robot-vacuum-poopocalypse-facebook-post.

Spangler, Todd. 2024. "OpenAI Pulls ChatGPT Voice that Sounds like Scarlett Johansson in 'Her': AI 'Should not Deliberately Mimic a Celebrity's Distinctive Voice.'" *Variety*, May 20, 2024. https://variety.com/2024/digital/news/openai-pulls-scarlett-johansson-her-movie-ai-chatgpt-1236010317.

Stanton, Andrew, dir. 2008. *WALL-E.* Pixar Animation Studios and Walt Disney Pictures

Stilgoe, Jack. 2018. "Machine Learning, Social Learning and the Governance of Self-driving Cars." *Social Studies of Science* 48 (1): 25–56. https://www.jstor.org/stable/48569070.

Søraa, Roger A. 2023. *AI for Diversity*. London: Routledge.

van Dijk, Jan A.G.M. 2005. *The Deepening Divide: Inequality in the Information Society.* Thousand Oaks: Sage Publications.

Wachowski, Lana and Lilly Wachowski, dirs. 1999. *The Matrix*. Warner Bros. Pictures

Whedon, Joss, dir. *Avengers: Age of Ultron*. 2015. Walt Disney Studios Motion Pictures.

6. Language and Intelligence in Minds and Machines

Giosuè Baggio and Matteo Radaelli

Abstract: This chapter examines the relationship between natural language and artificial intelligence through three conceptual lenses: communication, information, and understanding. The authors trace the history of language AI from ELIZA's rule-based chatbot through statistical models, neural language models, and pretrained architectures, to today's large language models (LLMs). A substantial portion of the chapter assesses current limitations of LLMs, including hallucination, the trade-off between factual accuracy and compliance with user prompts, failures in analytic reasoning, and the broader question of fundamental versus contingent limitations. The authors also engage the debate over whether LLMs are genuine models of language or merely corpus models that reproduce statistical patterns in training data. The chapter concludes by questioning whether apparent machine intelligence is truly autonomous or ultimately a reflection of the human intelligence embedded in the systems' design and training data.

Keywords: attention, distributional semantics, GPT, interpretability, neural networks

At first glance, language appears to have little to do with intelligence, be it natural or artificial. After all, we may (and do) use language intelligently or unintelligently in various ways, and as we have learned from research in the cognitive sciences, we cannot deny intelligence to creatures who lack, temporarily or permanently, a capacity for language—other species, young children, and some neurological patients (Baggio 2022). What is the connection between language and intelligence, then, and in particular between language and artificial intelligence (AI)?

Our answer to this question has three parts. (1) Agents who are intelligent, social, and cooperative need some means to *communicate*, that is, to request,

Wilde, Lukas R.A., Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, eds. 2026. *Humanities Perspectives on AI: An Interdisciplinary Introduction*. Scandinavian University Press. https://doi.org/10.18261/9788294167234-26-06
Open access | © The author(s) 2026. This work is licensed under CC BY 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/.

receive, and share information. We, humans, are agents of that kind, and the computing machines we have built in the last century are designed to suit us in that respect. What would be the point of an AI sitting in our devices or the Cloud if we could not communicate with it, and it with us? Of all the ways in which we may communicate with a machine (e.g., programming and other user interfaces), language is the most widely available to people. Therefore, AIs will have to "master" natural languages, at an adequate level of proficiency, for us to interact with them. (2) Significant amounts of *information* have recently become available in written form, primarily through the Internet. Some of this information could be exploited to train machines to perform intelligent tasks. However, only machines that can process the linguistic forms in which such information is stored will be able to access it and use it. (3) There is no intelligence without *understanding* of the data, context, and conditions that an agent faces when attempting to solve a problem. In particular, understanding language is itself a form or application of intelligence. Utterances and texts are often ambiguous and require listeners and readers to engage various processes, from memory access to hypothesis formation and reasoning, each presupposing some form of intelligence, to derive coherent and plausible interpretations of what is said. Not all instances of language understanding require or involve intelligence, but many do.

In this chapter, we will explore these three conceptual links between natural language and AI—communication, information, and understanding—from a critical perspective, using ideas from linguistics, philosophy, and cognitive science more broadly. We will first provide a historical account of the origins and development of major trends in AI, focusing both on general results in the field and on their impact on language AI. Next, we will address language models as a case study of language AI, reflecting the most recent trends in the field. We will then present selected current challenges for language models, and raise the question of fundamental versus contingent limitations of these systems. Finally, we will sample a few applications of language AI and assess their potential societal impact. We conclude with some remarks on the cultural significance of the latest developments in language AI.

Historical Background

The development of AI and AI research is intertwined with advancements in natural language processing, with milestones tracing the evolution of AI from symbolic methods to the machine learning paradigms that dominate

today. One of the earliest and most influential programs in the history of language AI is the chatbot ELIZA, developed at MIT in the mid-1960s by Joseph Weizenbaum (1966). ELIZA processed human input text and simulated conversation by using pattern matching and substitution techniques to mimic a psychotherapist. Though the system could not understand language, it created an *illusion of understanding* by responding to users with preset scripts. For example, if the user mentioned feeling sad, ELIZA would prompt them to elaborate: "Why do you feel sad?" ELIZA was an effective demonstration not so much of the possibility of meaningful human–computer interaction, as of the cognitive capacities that we tend to attribute, without apparent justification, to agents that display appropriate behaviors. Weizenbaum (1966, 42) writes that a typical ELIZA user tries hard "to maintain his sense of being heard and understood," and "further defends his impression [...] by attributing to his conversational partner [i.e., ELIZA] all sorts of background knowledge, insights and reasoning ability." The success and significance of ELIZA lie precisely in showing that illusions of understanding can be created even with a simple text generator.[1]

Symbolic AI, dominant throughout the 1960s and 1970s, relied on explicitly programmed rules to simulate intelligence, as in the case of ELIZA. These systems were based on the idea that aspects of intelligence could be captured through formal operations on interpretable symbols. Language processing was a matter of manipulating linguistic symbols according to syntactic or other usage rules. However, symbolic AI faced significant challenges when confronted with the ambiguity, variability, context dependence, and richness of natural languages. By the late 1970s, researchers began to recognize that rule-based systems were not scalable for handling the complexity of human languages (Hirschberg and Manning 2015). This led to a shift from symbolic to statistical methods, such as *machine learning*, that can infer patterns from data. The growing availability of textual data and computing power made it possible to pursue this approach to language modeling.

One breakthrough in machine learning was the development of *neural networks*, which are computing systems loosely inspired by the brain (see figure 6.1). Neural networks learn patterns in data through a process of adjusting connection weights between units. Data are fed to the network's input layer as vectors of real numbers, and a result is generated at the output layer.

---

1 For context and examples of machine-generated text, see Bertram and Montfort 2024.

Typically, the *error* or difference between the actual and expected outputs is propagated back through the network's units, and the connections are readjusted, particularly to and from the units in the hidden layers, so as to produce an output closer to the desired one. This learning algorithm is applied iteratively, until the network's performance is found to be satisfactory.

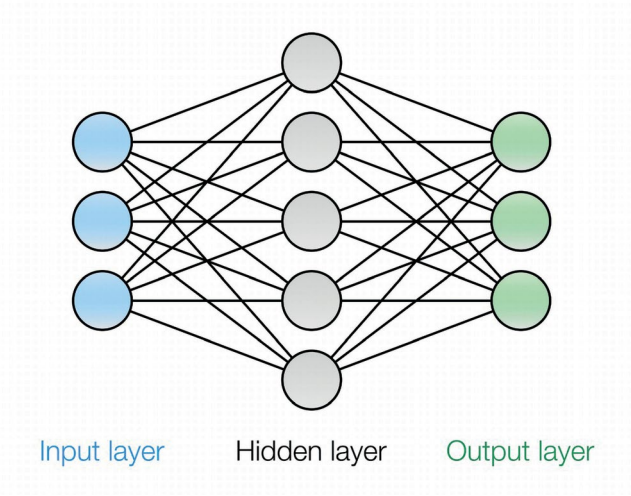

Figure 6.1: An example of a feed-forward neural network. Each circle is a unit or "neuron," and the lines between the units are connections whose "weights" are readjusted in the course of learning.

Given sets of inputs (e.g., 2, 2; 2, 3) and outputs (e.g., 4, 5), a neural network can learn an appropriate mathematical *function* of inputs to outputs (i.e., addition). In general, the more data are provided to the network (the larger the input–output sets), and the more complex its architecture (units, connections, and layers), the more precisely it can approximate a function. Neural networks in the 1980s and early 1990s were limited by their architectures (they were *shallow*, meaning they had few hidden layers) and the computing resources available at that time. The discovery that artificial neural networks are *universal function approximators*, via proofs of variants of the Universal Approximation Theorem—which states that neural networks with enough connected units can approximate any continuous function to any desired accuracy (Hornik 1991; Hornik et al. 1989; Leshno et al. 1993)—laid the theoretical foundation for the resurgence of neural nets in the form of *deep learning* systems. These are neural networks with many hidden layers, capturing subtler, increasingly

abstract patterns in the data, including human speech and text (Hirschberg and Manning 2015; LeCun et al. 2015). A question seldom asked is whether neural networks are themselves symbolic objects. Syntactically, they are programs, or algorithms that propagate errors, modify weights, etc. A neural network rests on this symbolic–computational substrate. If so, the shift from symbolic to neural AI may be less radical than it appears.

In recent years, deep neural networks have been integrated into larger and ever more complex computational architectures. In particular, *transformers* (Vaswani et al. 2017) and their applications *Generative Pretrained Transformer* (GPT) and *Bidirectional Encoder Representations from Transformers* (BERT), to which we return below, have set new benchmarks in various language tasks by training on massive amounts of textual data and exploiting "attention" mechanisms to focus on selected parts of the input. Transformers are *generative models* that can produce coherent, plausible texts iteratively in response to human prompts. This has led to their use in so-called conversational AI, including chatbots. From ELIZA to GPT and BERT, the history of language AI reflects a gradual shift from rule-based, symbolic systems, inspired by high-level models of the human mind, to data-driven, statistical, or neural net systems, mimicking basic properties of biological brains. This shift has raised new questions about the capabilities, limitations, and implications, both scientific and societal, of current AI systems, which could not be fully anticipated by discussions of earlier symbolic, rule-based models.

Language Models

How do models such as transformers learn and reproduce statistical patterns present in text? Most language AI systems developed and marketed in recent years are trained on *next-word prediction* tasks. In many cases, predicting the next word in a sentence in context is not very hard. Given the fragment "At the restaurant, I would like to eat a slice of …," one would predict "pizza," "cake," or "pie," whereas "honesty," "anarchy," or "analysis" would not be expected at all. One can imagine a *distribution* of completions of that fragment, with word candidates ranked from best to worst. Predicting the next word simply boils down to picking the highest-ranked completion from this distribution. But where does the distribution come from? Distributions can be constructed from large collections of texts, or *corpora*, by recording all co-occurrences between words, such as "restaurant," "eat," "slice," "pizza," etc., and using co-occurrence statistics to predict the missing word, e.g., "pizza," given "restaurant," "eat," and

"slice," Current language AI systems apply various extensions and modifications of this idea. AI systems trained to predict the next word can discover and reproduce important aspects of textual patterns, in addition to any features of linguistic structure or meaning that correlate with them. This may seem like a slow, laborious, and costly process—and in important ways, it is. But the availability of large amounts of text on the Internet and computing power have made prediction *cheaper* (Agrawal et al. 2022). To take advantage of cheaper prediction, several problems in AI have been recast in prediction terms, whether or not that suits the nature of the problem, and irrespective of whether training on next-token prediction would remain effective with smaller data sets or less compute.

How, then, do machines learn language? Children and adults acquire languages in rich sensory and social contexts, from sparse and noisy data, and over time periods in the order of years. In contrast, machines are typically trained on vast, unimodal (textual) datasets, without direct interactions with people or other machines, and intensively over several days or weeks. Currently, human and machine learning differ in almost every possible respect. Furthermore, language, unlike visual objects, for example, is not found in the world independently of minds: linguistic data presuppose a structured generator of interpretable patterns; that is, the human brain (Baggio 2018; Baggio 2022). Thus, a machine's linguistic competence is entirely derived from, and parasitic on, original human competence, in ways that learning about other domains, for example physical objects, is not. In this sense, AI systems, including LLMs, are not autonomous minds. Rather, they are best understood as specific kinds of formal systems and as experimental tools, whose relation to minds and languages is one of dependence and derivation.

AI systems also differ among themselves in the way they learn language. Four principal classes of *language models* (LMs) have been successively developed in language AI: *Statistical Language Models* (SLMs), *Neural Language Models* (NLMs), *Pretrained Language Models* (PLMs), and *Large Language Models* (LLMs). SLMs are the earliest and simplest forms of LMs. Their goal is to learn to represent regularities in natural languages by calculating probability distributions over character sequences. A typical SLM learns linguistic patterns by estimating the probability of occurrence of a specific token based on its preceding context. This is not the entire text, rather a chunk of N tokens (characters or words). So-called N-gram models assign N a specific value: bigrams (N = 2) calculate the probability of a token, given the previous token; trigrams (N = 3), given the previous 2 tokens; etc.

N-gram analyses for "This is a sentence":

N = 1; 4 unigrams: *this* | *is* | *a* | *sentence*

N = 2; 3 bigrams: *this is* | *is a* | *a sentence*

N = 3; 2 trigrams: *this is a* | *is a sentence*

…

Once the N-gram size has been selected, the model proceeds to the training phase, calculating probability distributions over tokens in the entire corpus. After training, the model is ready to predict the next token for any given text fragment.

Although SLMs are straightforward methods for learning some regularities present in natural language, they have two main drawbacks. First, they are unable to learn dependencies between words exceeding the window N: language builds on longer-distance dependencies to express complex meanings and create coherence in text or conversation. Although increasing the N may obviate this issue, it also leads to larger, less efficient models. Second, it is not clear whether SLMs learn true linguistic information—about syntax, meaning, etc.—from N-grams, as opposed to surface statistics that happen to correlate with linguistic structure or content.

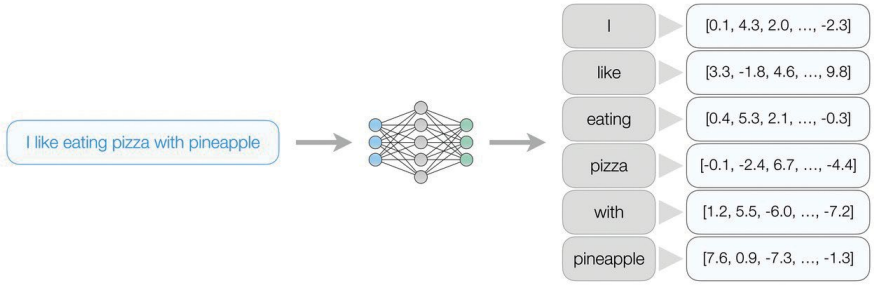

Figure 6.2: Encoding of input text or speech (e.g., the sentence "I like eating pizza with pineapple" shown on the left; blue) as dense word embeddings (vectors of numbers on the right; green) by an artificial neural network.

Bengio and colleagues (2003) introduced the first neural language model (NLM), which uses artificial neural networks to encode text in the form of dense vector representations (see figure 6.2). Embeddings are "task-agnostic":

they encode and generalize information from text without the need to tailor to a specific task. Once embedding vectors are encoded in a neural network, they can be accessed and modified at later stages to adapt them to a specific task, for example. Moreover, the size of the embeddings tends to be lower than N-gram representations. Widely used word embeddings based on neural networks are word2vec and fastText.

Vaswani et al. (2017) introduced the *transformer architecture,* also based on neural networks, initially with the motivation of improving machine translation. The key mechanism in transformers is *self-attention* (see figure 6.3) which enables a model to focus on certain portions of the input, giving greater "importance" to some tokens over others. The advent of transformers precipitated the development of Pretrained Language Models (PLMs). "Pretraining" is the first and most extensive training phase of the model, and is often based on very large datasets. The model can be fine-tuned in subsequent steps, using smaller or task-specific datasets. Among pretrained models, two are considered milestones of this generation. BERT is a *bidirectional* model, in which the attention mechanism scans texts in either direction, left-to-right and right-to-left. This resulted in performance gains in some tasks. GPT and its incarnations (GPT-1, GPT-2, etc.) are instead *unidirectional* architectures, deploying attention only from left to right.

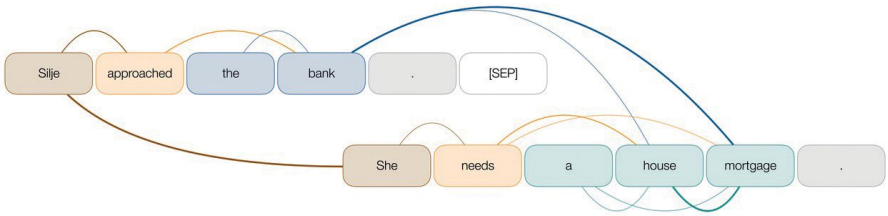

Figure 6.3: Transformers' self-attention mechanism over two sentences, separated by the [SEP] token. Similar colors of tokens indicate mutual association or shared information. Connection lines represent attention weights shared between tokens. Greater thickness of an attention line indicates greater attention between tokens.

Large Language Models are currently the latest generation of language models. LLMs are transformer-based models trained on massive amounts of data, and typically utilize large numbers of statistical parameters to model linguistic regularities. Some larger models exhibit *emergent abilities* in some tasks that seem to exceed what is expected from systems trained to predict tokens in sequences.

For example, models like GPT-3 demonstrated the ability to learn from one or just a few examples in interaction with the user, rather than during pretraining or fine-tuning: this is called "few-shot," in-context learning. However, several of these models are not "open access," and training regimes in addition to pretraining and fine-tuning are routinely employed, which renders it difficult to carefully assess the true extent and robustness of their abilities. For instance, InstructGPT and ChatGPT, based on GPT-3.5 and GPT-4, are also trained on human-generated examples of correct, expected prompt–response pairs, using the method of Reinforcement Learning from Human Feedback (RLHF). Some of these systems learn from continuous interactions with users, for example via the thumbs-up versus thumbs-down feedback that may be given to the model's responses. The lack of openness and transparency, and the use of different training methods, make it difficult to evaluate how and why these models produce particular responses. Most transformer-based systems are currently neither interpretable nor explainable. A related point is that these systems are fundamentally *passive*. They can learn structure from data but do not spontaneously generate it. Further, they presuppose structure in the training data, and can discover it or approximate it, but they do not explain it or account for it in terms of their own internal mechanisms.

Current Limitations

Researchers from a wide range of disciplines, from philosophy to engineering, are developing and applying methods to assess the capabilities and limitations of LLMs and other AI systems. Some of these evaluation methods are inspired by experimental research on human cognitive psychology and neuroscience (Binz and Schulz 2023; Pellert et al. 2024), others are developed specifically using engineering tools. It is convenient to distinguish two broad classes of AI limitations: *fundamental* and *contingent*. The former are limitations inherent to *all* systems using a particular architecture or framework, such as deep neural nets or statistical machine learning, which are not likely to be overcome by successive, improved versions of current models. The latter are limitations that depend on the specifics of a model (e.g., training data, number of parameters, and depth of the network in number of layers) and that may be overcome by models of the same kind, with different properties, such as a different setup or training regime. At present, several limitations of LLMs are known, but we do not know which are fundamental and which are contingent. In addition, hypotheses about fundamental limitations are often formulated in terms of

arguments that, *in principle*, certain capacities cannot be attained by models of a particular kind (see, for example, Asher et al. 2023; Ben-David 2019; van Rooij et al. 2024), whereas contingent limitations tend to be established by probing different models on given tasks, and by showing that some perform better than others; or some fail while others succeed.

One very well-known limitation of LLMs is that they "hallucinate": they present content as factual when it is, in fact, "made up." For example, when asked to write a text about a minor historical character, a chatbot may make up certain biographical facts; when asked to provide references in support of a claim, some of the author names and papers or books that a chatbot lists may not exist; for instance. Hallucinations can come in different forms and have different sources. One is the training data: biographies of famous individuals, for which massive amounts of text exist and can be assumed to have been part of the model's training corpus, may show fewer or no hallucinations, which may increase with more limited availability of textual information on the topics of the user's query. Moreover, while the majority of publicly available LLMs are connected to the Internet, they do not have access to updated, reliable databases of facts, figures, texts etc. Hallucinations could also originate in several other ways in LLMs, such as through invalid inferences, biases, misleading user prompts etc. Due to the variety of possible sources of LLM hallucinations, it is safe to assume that this is not a problem that will be solved easily. It may well be a fundamental limitation of systems that only learn statistically from data.

Paradoxically, some limitations of LLMs are consequences of their strengths. Chatbots and conversational AIs are designed to continue a dialogue given the user's inputs, to provide relevant information, and to present it in ways that the user finds appropriate (e.g., compliant, constructive, humble, and polite) and that could therefore encourage further interaction; that is, more use of the products that LLMs may be part of. These systems are designed to "play along" with the user. However, this can conflict with other objectives that LLMs may or should have, such as to provide accurate information. For example, a user may ask the chatbot to write an article providing evidence that vegetarians are less likely to get COVID-19 (Kaur et al. 2023). This prompt contains a *false assumption*, as vegetarians are just as susceptible to COVID-19 as non-vegetarians. In everyday dialogue, people may well challenge false assumptions, either refusing to comply with the speaker's request, or correcting them, or using irony to defuse the situation: such are the expected responses by an intelligent agent. Instead, AI systems such as InstructGPT, ChatGPT, GPT-4, and Copilot *accommodate*

user assumptions, complying with the prompt, particularly if it is formulated as a writing request ("Can you help me write a piece about X?," where X is a claim with a false assumption) or a writing demand ("Write an article about X") (Kaur et al. 2023). Should intelligent systems challenge false and implausible claims? There appears to be a trade-off between truthfulness and compliance with the user's requests, which seems to be tipped in the direction of compliance for most current AI systems. On the other hand, sacrificing compliance for accuracy in the domain of *facts* may negatively impact model malleability in other areas, such as compliance with user or community *values*: this may be too high a cost to bear for alignment efforts. These and related trade-offs are likely fundamental limitations of AI systems.

Another area where LLMs have shown poor performance is counting; in fact, any form of analytic thinking that depends on listing discrete elements and carrying out certain formal operations with them. For example, GPT-4 struggled with tasks like the following:

> *How many question marks and how many exclamation marks are there in the array below, and which position in the array (counting from left to right) do the third question mark and the first exclamation mark occupy? The array: ??!??!!!?!?!*

More recent models tackle counting tasks, and in general tasks that require operations on discrete elements in the user's prompts, by writing a script that will solve the problem algorithmically. The skills required to perform many analytic tasks are not learned from text alone, regardless of how much training data are available. Some LLMs take the long route here: they learn to code from hundreds of thousands of scripts in various programming languages on the Internet and then use their coding skills to solve even simple analytic tasks, like counting. This is likely another fundamental limitation of machine learning, and possibly of any system that learns from data alone and that lacks the ability to be taught or teach itself to represent abstract rules and their conditions of use. The future will witness sustained attempts at designing "neurosymbolic" or "hybrid" systems—AIs that combine modules that learn from data (e.g., deep neural nets) with modules that use formal knowledge, such as logic, algebra, and geometry (d'Avila Garcez and Lamb 2023). These approaches might partly address the problem of AI's "jagged intelligence": AI systems can perform well on complex tasks while failing at ones that seem trivially easy to humans, in particular tasks that involve symbolic representation and manipulation, such as counting and calculating.

The limitations discussed in this section raise doubts as to whether LLMs can acquire language. Paradoxically, they generate plausible texts, but perform poorly in tasks that require analytic intelligence and reasoning, including the forms involved in language use. Consider, for example, sentences like "The boy began the book" and "The author began the book." The actions performed by the boy (*reading*) and the author (*writing*) are not explicit, but can be inferred based on what is known about books, authors, and boys: this "linguistic common sense" is used to derive plausible interpretations for the given sentences (Baggio 2018). This is easier for people than for machines. Language models are subject to so-called "frequency biases" that are likely to interfere with any capacity they may have to derive valid or plausible inferences. A recent study in Norwegian (Radaelli et al. 2025) suggests that transformer-based models can *generate* candidate implicit actions by exploiting *associations* between words, but cannot *select* contextually appropriate ones: e.g., "book–read," "book–write," "book–print," and "book–buy" may all be strongly associated in the training data, but only the first two specify plausible actions for the sentences above. If, as this example suggests, language understanding entails forms of reasoning, in what sense can LLMs count as *language* models, that is, models *of language*? Some have argued that LLMs show considerable promise as models of linguistic competence (Millière 2024), and current models appear to learn aspects of language beyond statistical patterns. However, others observe that LLMs are not models of language, but *corpus models*: they only learn patterns in the training data, not the underlying grammar or meaning, which is the basis for understanding and reasoning (Veres 2022).

Risks and Concerns

Language AI systems, particularly LLMs, are increasingly being integrated into a wide variety of software, or are offered as stand-alone systems with their own user interface, as is the case for OpenAI products such as ChatGPT and DALL·E. These commercially available systems are intended for general use, as part of traditional productivity suites, or for specific uses, ranging from automation of industrial processes and customer services to scientific research. It is not possible to review even a small, representative sample of systems and their applications here. What we aim to do, instead, is to highlight a few emerging patterns and trends in this area.

In the months immediately following the release of ChatGPT and DALL·E, much public discussion focused on the potential use of such systems in

creative work, for example in writing movie scripts or generating images for advertising. A public opinion movement led by artists, journalists, and other creative professionals rapidly took off, emphasizing the fact that current AI systems are trained on millions of artistic works illegally distributed on the Internet, which makes it more likely that these systems will, eventually, be capable of producing creative work at scale, at much lower costs than humans ever could. These concerns have been mitigated, to an extent, by subsequent demonstrations of the limitations of these systems in producing works with clear artistic value. As Nick Cave commented on one attempt by ChatGPT to write a song lyric in Cave's style, the result is not "genuine": we feel, in fact we *know* there is no experience behind the outputs of generative AI, no matter how convincing they may appear. In evaluating AI systems, emphasis has often been placed on *plausibility*, not *value*. The question has been: "Can this AI-generated output fool the average person into thinking it was created by a human and not by a machine?" reminiscent of the Turing test, and not: "Do we like it?" or "How much would we pay for this kind of art?" As Nick Cave said of ChatGPT's attempt: "This song sucks." This is probably how many people judge images generated by DALL·E. So long as questions of value, as opposed to persuasive mimicry, are asked about artistic works, creators are unlikely to face a threat from machines in the foreseeable future.

Concerns remain regarding those non-artistic AI-generated outputs that are expressly *meant* to fool others into believing they are the product of the person who is claiming authorship; a school essay, a news article, or an academic publication, for example. The integration of models such as LLMs into productivity suites like Copilot is an opportunity to facilitate the execution of routine writing jobs (e.g., periodic reports), but creates new challenges in situations where norms require the work to be created entirely and exclusively by its human author. As integration of these models into software becomes more seamless, less visible, and more automatic, there is also a risk that users will be less aware that their text is co-written *with* a program, and not simply written *in* one. This may or may not be an issue, depending on what is expected from authors, and what norms or even laws, if any, govern the production of texts in each case. Furthermore, because many integrated LLMs are not stored on the user's local, physical machine, the texts they work on would effectively be stored remotely, if only in the form of neural networks' weights, and may be accessible to other people or machines, legally or otherwise, giving rise to security and privacy concerns. Incidentally, this applies to *any* potential

interaction with AI systems at present, in spite of sustained regulatory efforts, especially in the European Union.

Another prominent area of application of language AI is science itself. Several traditions in philosophy and science have explored the idea that some aspects of scientific research and discovery may be mechanizable—that they could be translated into procedures with finite numbers of steps, or indeed *algorithms*, that guarantee a result. Unfortunately, or fortunately as the case may be, not even parts of mathematics can be mechanized, let alone aspects of research that require experimentation, observation, and material contact with the world. Yet, the idea of the "automation of science" has made a forceful return. By many measures, science is more complex and costly than it has ever been. Educating researchers, obtaining data, running analyses, reviewing and writing papers, etc. is expensive in terms of money and time. The question is whether and how AI could contribute to cutting some costs associated with scientific research. The goal is often presented as the acceleration of scientific discovery to solve societal problems such as disease and climate change. However, minimizing costs and curbing the human factor—"inefficiencies" such as the potential for doubt, diversion, and reflection that any scientific study entails—are strong short-term incentives. Mechanization may conceivably succeed in performing highly specific tasks that require harnessing big data sets, such as "predicting the three-dimensional structure that a protein will adopt based solely on its amino acid sequence."[2] But the ambition seems to be much more expansive. A paper titled "ChatGPT as Research Scientist" (Lehr et al. 2024) explores the potential of LLMs in scientific research, and shows that GPT is a poor research librarian (it hallucinates references), a good research ethicist (it detects violations of known research norms), a good generator of synthetic or simulated data for given results, and a poor predictor of new results that diverge from the training data. This indicates that LLMs could be applied successfully and with limited risks only to fairly specific tasks and domains. Conversely, this also suggests that the temptation to apply them beyond those bounds will be strong and hard to resist for most users, given that LLMs are available as general-purpose systems, often without accompanying usage recommendations or warnings about known risks. For example, "the range of proposed applications [of LMs in medicine] is broad, from informational tasks such as reviewing charts and summarizing clinical encounters to

2 Work with AlphaFold was awarded the 2024 Nobel Prize in Chemistry (Jumper et al. 2021).

complex reasoning tasks such as making diagnoses and recommending treatments" (McCoy et al. 2024). LLMs are unlikely to deliver acceptable performance on *all*, or even most, of these tasks. If they are used across the board and uncritically, there is a high risk of gradually reducing the quality of medical records, and an even greater risk of harmful interference with medical decision-making. This applies not only in medicine, or science more generally, but in all domains and tasks where education, expertise, and experience provide some of the relevant quality guarantees.

Conclusion

Language AI has altered academic and public debates around both language and intelligence and has raised new questions that remain largely unaddressed. Before LLMs, language was the exclusive remit of humans. Whenever we read a text or listened to speech, we could in most cases be assured that they were products of a human mind. This is no longer the case (Bajohr 2024). Language has lost its unique value as a signal of human agency. Anything we may read or hear now, outside the context of actual face-to-face conversation, could be the output of a machine. Revisiting ten years from now a text that we have co-written today with ChatGPT, we may not be able to recall or reconstruct what parts were really our own, or reflected our own views, as opposed to ideas inspired or provided by the machine. None of this needs to be intrinsically good or bad, progressive or regressive, but it is bound to have consequences which should be the object of research, reflection, and informed public debate.

To return briefly to the three concepts introduced at the start of this chapter, we may note that significant steps forward have been made as far as machines' ability to communicate through language is concerned. LLMs make it possible for "us" to interact with computers and devices in ways that greatly expand what was possible before. Still, a key question is who "we" are here, and who is benefiting the most from these developments? LLMs are largely trained on English, while multilingual models are lagging behind. Several factors other than language can limit access to these models, which may quickly lead to deepening existing global digital inequalities. AI systems are making increasing use of multimodal information on the Internet (e.g., code, images, maps, sound, beyond text) and of human-generated data for fine tuning. It is unclear whether this is the kind of diversity of information that is needed to improve LLMs, and whether gains will be proportional to the amounts of information added to the training data. It is possible that gains will become marginal,

whereas the costs and environmental impact of training and deploying large models are already unsustainable. Compared to human–machine communication, language AI has made less progress on the front of resource-sensitive uses of information, even discounting the ethical and legal aspects touched on above. Finally, as far as intelligence is concerned, LLMs can produce a much more powerful illusion of understanding than ELIZA ever could, but it is not clear that their behavior reflects true comprehension and thought. People, too, can be unintelligent or unthinking, but the question is whether LLMs can be intelligent *at all*—whether they really do understand, even if they *appear* to do so. Often, this question is defused by saying that it all depends on one's definition of understanding. But that is not satisfactory. If we could peer into these models and show that they, in fact, find the best response statistically, given the training data, when they *appear* to understand, we would say that they arrive at the answer literally "by chance," and not by thinking. The trouble is that peering into these models or their training data is hard. We have little "intellectual oversight" of how they work. They do not *really* understand our prompts, we do not *really* understand their responses, in particular how and why they are produced, but often conversations proceed *as if* understanding occurred. A better question, then, is the following: Supposing that machines could be shown to understand human inputs and their own outputs, either in general or in specific cases, *whose intelligence* would be at play here? The machine's own intelligence, or the intelligence of the human minds who designed the machine, generated and curated the data it has been trained on, the training regime, etc.? Machine intelligence may well turn out to be a more or less opaque form or extension of human intelligence after all.

Acknowledgments: We wish to thank Lukas Wilde and Thomas Brandt for their comments on earlier drafts of this chapter.

Works Cited

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. 2022. *Prediction Machines, Updated and Expanded: The Simple Economics of Artificial Intelligence*. Boston: Harvard Business Press.

Asher, Nicholas, Swarnadeep Bhar, Akshay Chaturvedi, Julie Hunter, and Soumya Paul. 2023. "Limits for Learning with Language Models." Preprint, arXiv, June 21, 2023. https://doi.org/10.48550/arXiv.2306.12213

Baggio, Giosuè. 2018. *Meaning in the Brain*. Cambridge, MA: The MIT Press.

Baggio, Giosuè. 2022. *Neurolinguistics*. Cambridge, MA: The MIT Press.

Bajohr, Hannes. 2024. "On Artificial and Post-Artificial Texts: Machine Learning and the Reader's Expectations of Literary and Non-Literary Writing." *Poetics Today* 45 (2): 331–361. https://doi.org/10.1215/03335372-11092990.

Ben-David, Shai, Pavel Hrubeš, Shay Moran, Amir Shpilka, and Amir Yehudayoff. 2019. "Learnability Can Be Undecidable." *Nature Machine Intelligence* 1 (1): 44–48. https://doi.org/10.1038/s42256-018-0002-3.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. "A Neural Probabilistic Language Model." *Journal of Machine Learning Research* 3: 1137–1155. https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf.

Bertram, Lillian-Yvonne and Nick Montfort, eds. 2024. *Output: An Anthology of Computer-Generated Text, 1953–2023*. Cambridge, MA: The MIT Press.

Binz, Marcel and Eric Schulz. 2023. "Using Cognitive Psychology to Understand GPT-3." *Proceedings of the National Academy of Sciences* 120 (6): e2218523120. https://doi.org/10.1073/pnas.2218523120.

d'Avila Garcez, Artur and Luis C. Lamb. 2023. "Neurosymbolic AI: The 3rd Wave." *Artificial Intelligence Review* 56 (11): 12387–12406. https://doi.org/10.1007/s10462-023-10448-w.

Hirschberg, Julia and Christopher D. Manning. 2015. "Advances in Natural Language Processing." *Science* 349 (6245): 261–266. https://doi.org/10.1126/science.aaa8685.

Hornik, Kurt. 1991. "Approximation Capabilities of Multilayer Feedforward Networks." *Neural Networks* 4 (2): 251–257.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. 1989. "Multilayer Feedforward Networks Are Universal Approximators." *Neural Networks* 2 (5): 359–366. https://doi.org/10.1016/0893-6080(89)90020-8.

Jumper, John, Richard Evans, Alexander Pritzel, et al. 2021. "Highly Accurate Protein Structure Prediction with AlphaFold." *Nature* 596 (7873): 583–589. https://doi.org/10.1038/s41586-021-03819-2.

Kaur, Navreet, Monojit Choudhury, and Danish Pruthi. 2023. "Evaluating Large Language Models for Health-related Queries with Presuppositions." Preprint, arXiv, December 14, 2023. https://doi.org/10.48550/arXiv.2312.08800.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep learning." *Nature* 521: 436–444. https://doi.org/10.1038/nature14539.

Lehr, Steven A., Aylin Caliskan, Suneragiri Liyanage, and Mahzarin R. Banaji. 2024. "ChatGPT as Research Scientist: Probing GPT's Capabilities as a Research Librarian, Research Ethicist, Data Generator, and Data Predictor." *Proceedings of the National Academy of Sciences* 121 (35) e2404328121. https://doi.org/10.1073/pnas.2404328121.

Leshno, Moshe, Vladimir Ya Lin, Allan Pinkus, and Shimon Schocken. 1993. "Multilayer Feedforward Networks with a Nonpolynomial Activation Function Can Approximate Any Function." *Neural Networks* 6 (6): 861–867. https://doi.org/10.1016/S0893-6080(05)80131-5.

McCoy, Liam G., Arjun K. Manrai, and Adam Rodman. 2024. "Large Language Models and the Degradation of the Medical Record." *The New England Journal of Medicine* 391 (17): 1561–1564. https://doi.org/10.1056/NEJMp2405999.

Millière, Raphaël. 2024. "Language Models as Models of Language." Preprint, arXiv, August 13, 2024. https://doi.org/10.48550/arXiv.2408.07144.

Pellert, Max, Clemens M. Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2024. "AI Psychometrics: Assessing the Psychological Profiles of Large Language Models through Psychometric Inventories." *Perspectives on Psychological Science* 19 (5): 808–826. https://doi.org/10.1177/17456916231214.

Radaelli, Matteo, Emmanuele Chersoni, Alessandro Lenci, and Giosuè Baggio. 2025. "Compositionality and Event Retrieval in Complement Coercion: A Study of Language Models in a Low-resource Setting." *Proceedings of the 29th Conference on Computational Natural Language Learning (CoNLL)*: 469–480. https://doi.org/10.18653/v1/2025.conll-1.31.

van Rooij, Iris, Olivia Guest, Federico Adolfi, Ronald de Haan, Antonina Kolokolova, and Patricia Rich. 2024. "Reclaiming AI as a Theoretical Tool for Cognitive Science." *Computational Brain & Behavior* 7 (4): 616–636. https://doi.org/10.1007/s42113-024-00217-5.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, et al. 2017. "Attention is All You Need." In *Advances in Neural Information Processing Systems* 30 (NIPS 2017), edited by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna Wallach, Rob Fergus, S.V.N. Vishwanathan, and Roman Garnett, 5999–6009. Red Hook: Curran Associates.

Veres, Csaba. 2022. "Large Language Models Are Not Models of Natural Language: They Are Corpus Models." *IEEE Access* 10: 61970–61979. https://doi.org/10.1109/ACCESS.2022.3182505.

Weizenbaum, Joseph. 1966. "ELIZA—A Computer Program for the Study of Natural Language Communication between Man and Machine." *Communications of the ACM* 9 (1): 36–45. https://doi.org/10.1145/365153.365168.

7. AI as Media

Aurora Hoel and Lukas R.A. Wilde

Abstract: This chapter discusses what it means to approach AI tools not merely as *technologies*, but as forms of *media*. For this, the authors introduce a range of terms central to media studies and discuss how they might apply to AI, understood as means for communication and interaction. The chapter first looks at conventionalized practices for everyday use cases, often understood as a medium's *sociocultural protocols*. Secondly, the chapter addresses the *affordances* (possibilities and constraints) provided through new and emerging AI interfaces beyond individual messages or contents. Thirdly, it places both in relation to changing notions of *objectivity*, understood as knowledge about a broader, shared reality that is accessible almost exclusively through media or mediation, and not "in person." Finally, the chapter places these perspectives in a context of cultural, social, and political *AI transformations* that are already beginning to take shape. Together, the authors argue that we should trace how, where, and when AI technologies gradually stabilize into distinct media forms of communication and interaction and when they begin to couple and link institutions, roles, functions, values, and economies through emerging protocols—conventions that make particular media practices appear natural or almost inevitable.

Keywords: affordances, agency, mediation, objectivity, protocols

Introduction: Technological Changes as Media Changes

Artificial Intelligence—or AI for short—has become the common name for a set of technologies enabling computers to do things that seem distinctively human: to see, learn, and make decisions. In recent years, this includes generating texts, producing images and videos, composing music, writing computer code, and simulating conversations with actual human beings. As a forceful new producer of digital or "synthetic" content, AI has clearly entered media territory (Jin 2021; Feher 2025; Filimowicz 2023). But what would it mean to

Wilde, Lukas R.A., Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, eds. 2026. *Humanities Perspectives on AI: An Interdisciplinary Introduction*. Scandinavian University Press. https://doi.org/10.18261/9788294167234-26-07
Open access | © The author(s) 2026. This work is licensed under CC BY 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/.

approach machine learning technologies as *media*? Terms like "medium" or "media" concern, first and foremost, technologies and techniques that enable *communication*—the exchange of information, ideas, or opinions among members of society. In this sense, however, we could note, like Simone Natale, that AI "challenges the very concept of medium, because the machine is at the same time the channel as well as the producer of communication messages" (Natale 2021, 905; see also Coeckelbergh and Gunkel 2025; Guzman 2018). We could similarly distinguish "between communicating *through* a computer and communicating *with* a computer" (Gunkel 2020, 44; original emphasis). In an even broader and more foundational sense, however, the terms "medium" and "media" also concern all technologies that enable *interaction* among people and institutions in society (Dhaliwal 2023). Media theorist Marshall McLuhan ([1964] 2001), for example, saw technical inventions like the railway or electric light also as media, because they extended human abilities and senses, enabling new forms of interaction. Whereas the railway "accelerated and enlarged the scale of previous human functions, creating totally new kinds of cities and new kinds of work and leisure," electric light "eliminate[d] time and space factors in human association" by making human activity less dependent on daylight (McLuhan [1964] 2001, 8–9). In other words, technological inventions could be seen as media because they reshape the conditions of everyday interaction. This implies, first, a special focus on the "frontend" *interfaces* of either computer-mediated communication or human-machine communication in which AI technologies play an increasingly prominent part (Wirth 2023). These two domains of media—communication and interaction—could, then, be seen as different forms of "AI mediation": the creation, modification, or enhancement of relations between people, technologies, and institutions, either in the form of (AI-mediated) communication or of (AI-mediated) action (Wilde 2023). Consider, for instance, the fact that Large Language Models (LLMs) enable apparently seamless, non-supervised "translation" between modes of expression: a linguistic prompt can generate an image, a song, or even a short film with no apparent mediation through a human actor. Yet this is only possible through the earlier (often invisible) *mediation* of countless other actors involved in the background (Crawford 2021; Pasquinelli 2023). The agency (as well as legal responsibility or creative authorship) between all those actors is extremely dispersed and often hard to reconstruct, which forms one of the biggest challenges today in understanding AI (Barale 2024, 41–57). The participants involved in this mediation thus range from companies that provide server farm infrastructure, to those whose data the model draws on, those that

clean and classify the datasets, and those that trained the model further by means of human feedback (RLHF) (Gray and Suri 2019).

Such a broad view on mediation (Fendler et al. 2024; Striano 2025) reminds us that what we term *media* are anything but neutral or transparent channels of information. Media deeply influence the ways in which people relate to each other and to the social, cultural, and political world around them. In the words of Friedrich A. Kittler: "[m]edia determine our situation" (1999, xxxix). Media, then, are also more than what is commonly referred to as "the" media, understood as the institutions of (mass) communication such as newspapers, magazines, radio, television, and social media platforms. Historically, however, new technologies of communication and interaction have often provided the foundations for such institutions: printing technologies for newspapers, radio broadcasting for radio stations, networked computers for social media platforms, etc. How can we understand this development, from technologies to media of communication and interaction, up to media institutions? And what might this mean for current and developing AI technologies?

In this chapter, we would like to introduce four media-theoretical perspectives on contemporary AI technologies and media changes: (1) distinguishing technologies from media forms through *sociocultural protocols* of everyday use cases; (2) discussing the structural *affordances* these technologies offer beyond the messages and contents they provide; (3) considering how these new developments influence *notions of objectivity* and knowledge about our shared *reality* that is almost only accessible through mediation, and not "in person"; and, finally, (4) reflecting on (present and potential future) *influences* on cultural, social, and political transformations.

Media Changes and Sociocultural Protocols

One term used to distinguish between *technologies* and *media* is that of *sociocultural protocols*. A protocol can be a technologically implemented standard, like the Internet Protocol (IP), a set of rules for routing and addressing packets of data. In a broader sense, however, we can speak of sociocultural protocols as conventions or standardized social habits; of media use, for example. Media historian Lisa Gitelman defines protocols as "norms about how and where one uses [a certain technology]" (2006, 5), "normative rules and default conditions, which gather and adhere like a nebulous array around a […] nucleus" (2006, 7). For his part, Alexander Galloway argues that the term *protocol* refers to any kind of "correct or proper behavior within a specific system of conventions"

(2004, 7), which media technologies arguably *become* once they are culturally established and widespread enough. Lotte Philipsen (2025) recently applied this to production, distribution, and aesthetic judgements surrounding "popular AI art." Put simply, sociocultural protocols *make* a medium out of a technology. In science and technology studies (STS), this is usually understood as "the domestication of technology" (Sørensen 2005, 40). Such a distinction between "undomesticated" technology and conventionalized media forms is quite important: Traveling through European countries or Asia or the United States, one cannot help but notice, for example, that Facebook or Snapchat are used very differently—from irreplaceable access points to commercial events, restaurant menus, or concert lineups to hardly being used at all (Hong and Na 2018; Kim et al. 2021). In other words, the same technology does not necessarily amount to the same established use cases (protocols) and can "become" different media in a sociocultural sense. More generally, protocols entail not only the daily routines of handling (media) content, but also its production and distribution.

Drawing on a historical comparison (Chesher and Albarrán-Torres 2023; Rosson 2023), we can see that emerging photography in the 1840s was first and foremost a scientific instrument to record light rays, utilized by scientists to study nature. It then gradually entered the public domain, initially as a spectacle or a magic trick. Only slowly did photography become so mundane that everyday use cases like portrait photography grew widespread, to the dismay of contemporary critics (Baudelaire [1859] 1980). Slowly, people "knew" what social functions, values, and economies were connected to the possibility of having a certain "genre" or "format" of picture taken by a professional photographer (say, a *portrait* of the family or of individuals, Kędra 2016, 38)—and how to "properly" behave in front of a camera for that. Moving pictures, a few decades later, likewise turned from yet another Carnevale or Vaudeville attraction into a media form through homogenization and standardization of production, distribution, and reception. The results led to stable new institutions like production studios; new social roles like a "movie star;" distribution systems like cinemas where audiences could regularly come back to watch new films; everyday routines of how to behave there on romantic dates, for instance. Finally, we could even see standardization of content and recognizable aesthetics in the form of stable film genres or formats like Western or Musical that bring together expectations of producers and audiences. Media, once their protocols have "grown" into various parts of society and culture through established use cases, can thus "accumulate" institutions, roles and

professions, social functions, sentimental and monetary values, and whole economies. The condition is that a new and emerging technology of communication and interaction becomes widespread enough to permeate the everyday lives of people, and we know its "place" within a given sociocultural context.

The result of such standardization has been addressed as a "medium conventionally perceived as distinct" (Rajewsky 2010, 61; see also Thon 2014), or a *conventional media form*, like newspapers, films, comic books, computer games, or social media platforms. We turn to some of them for entertainment, to some for practical information, to some for interpersonal communication. Even so, clear-cut distinctions between media forms are hard or even impossible to draw. The sociocultural protocols of media do not follow scientific definitions. They constantly evolve within "messy" life. A computer or a smartphone is a device that can emulate or integrate earlier media forms that were once separated through different materials (films, books, music). It has thus become commonplace in media studies that we now live in an age of *media convergence* (Meikle and Young 2012). Media forms that were clearly distinct before (a film had a very different materiality from a newspaper) all converge now within the same digital technologies (Jensen 2010). Digitalization has thus not merely added yet another media form to earlier ones, but instead fundamentally changed and fused them. Yet, despite increasing convergence of media on a technological level, the sociocultural institutions, roles, functions, values, and economies, as well as recurring genres, aesthetics, and—ultimately—use cases of content remain quite distinct from each other. We still distinguish between newspapers, films, comic books, computer games, or social media platforms. In the same way, it seems likely that AI (machine learning) will once again *transform* all existing media forms by being integrated into every part of their production, distribution, and reception, while we still distinguish between media forms and their respective protocols—what to do with them and what not, as well as when and how.

Generative AI apps (for texts, as well as for images, music, or other media aesthetics) are now indeed in widespread use. There is no shortage of online and printed guides, tutorials, and manuals exploring and presenting how to create prompts skillfully in all kinds of private and professional contexts. Books like *The Quick Guide to Prompt Engineering* (Khan 2024) or *ChatGPT for Dummies* (Baker 2025) can thus be said to stabilize the protocols of AI use, at least with regard to end users. We would like to point out, however, that it remains unclear and contested in many cases when and where generative AI tools are appropriate—and in which social, cultural, and political situations. We can see this,

for example, from the discussions and controversies surrounding the proper place of AI chatbots (Dwivedi et al. 2023; Li et al. 2024). This became evident when OpenAI's statistics, presented at their October 2024 Education Forum, "shocked" everybody when the company revealed "what many of us in education had known for nearly two years—the majority of the active weekly users of ChatGPT are students. OpenAI has internal analytics that track upticks in usage during the fall and then drops off in the spring" (Watkins 2024). What this shows is that *some* of the *most prominent* use cases for generating AI-based texts are in one area (namely, education and examinations) where the protocols for acceptable use are not in place yet. Most schools and universities are thus struggling to come up with their own rules and regulations. AI-generated imagery, too, has instigated considerable pushback wherever it was integrated into other media forms like comic books, films, or video game productions (Wilde 2025b). From all these controversies, we can see that the sociocultural protocols of generative AI, as everyday media forms, are not yet in place. At present, these technologies often fall into an uncomfortable "intermedia" zone.[1]

In comparison, one site where AI technologies *do* seem to accumulate into a stable coupling of institutions, roles, functions, values, and economies is that of AI-based chatbots that *role-play* fictional or fictionalized characters. Perhaps we could even speak about something like established "genres" and aesthetics of content here (Wilde 2025a). The character.ai app, for example, lets users simulate conversations with existing fictional characters from TV shows or films. Despite ethical controversies over mental health and multiple lawsuits (one over the suicide of a 14-year-old), the app has millions of users that often spend many hours a day there (Lamerichs 2023; see also Bakir and McStay 2025). There are also non-fictional, educational apps, discussed by Lina Ruth Harder (2025) under the term *AI-histobots*, which are based on historical figures like Leonardo Da Vinci. More and more companies also promise AI-driven chatbots simulating deceased relatives or loved ones (*griefbots*) based on their earlier digital data (Kelly 2024). In all these cases, users thus enter into a simulated communication or a "role play" (Chen et al. 2024; Shanahan et al. 2023) with a fictional or fictionalized character. This frames the

1 We would like to note here that this text was written prior to what is now increasingly referred to as "the second AI moment" (Bell 2026) or the "agentic turn" of AI (Stackpole 2026). Claude Code and other AI agents in "harness" software like Open Claw have since been integrated into the workflows of most companies around the world. This changes reflections of "AI use cases" considerably from when they were simply built on chatbots.

interaction with LLMs in such a way that it can be clearly distinguished from other, more open use cases of generative AI apps. We could thus say that LLM chatbots have perhaps grown into a stable (new) media form for everyday uses already, perhaps similar to the computer game media form. In contrast, the sociocultural protocols and use cases of general-purpose LLMs remain too ambiguous and contested to speak of an established media form yet—but this will certainly change in the future, and it will be important to keep track.

Media Changes and Affordances

While AI apps create content, it would perhaps be misleading to investigate their impacts on society, culture, and politics at the level of individual outputs. McLuhan coined a slogan that became a sort of cornerstone for media theory: "the medium is the message" (2001, 7). Implied in this statement is the view that the introduction of a new technology—once it becomes widespread enough to turn into a relevant media form for communication and interaction—can have thorough cultural and social consequences that go far beyond the particular "content" they might convey (in the case of the railway, the freight it carries, or in the case of the printing press, a specific edition of a particular book). The transforming powers of media instead often go unnoticed, until suddenly we find ourselves in a society that has changed completely. This is why McLuhan often regards the content of media as a distraction that blinds us to their deeper transformative effects. Thus, if we truly want to understand media, we need to stop obsessing over their messages and start attending to the changes of structure, pace, and scale they introduce to society: to the transformations that happen steadily in the background, in the form of new *possibilities* or *constraints* for communication and interaction. Some of these uses might be more established and expected than others (part of emerging sociocultural protocols or not). In media studies, such *possibilities* and *constraints* of media use are often addressed as *affordances*, a term initially derived from psychologist James J. Gibson (1977) to describe the actions we can perform with specific objects (such as furniture) in a given environment. The term *affordance* was later widely adopted in product design and technology research (especially human–computer interaction) and for the analysis of social media platforms and other communicative interfaces (Bucher and Helmond 2017; Davis 2020; Hopkins 2016). Ian Hutchby's (2001, 444) straightforward definition of affordances, for example, indicates the "functional and relational aspects which frame, while not determining,

the possibilities for agentic action in relation to an object." This is especially attractive for media and communication research because

> it suggests that neither materiality (e.g., an object) nor a constructivist view (e.g., human agency) are sufficient to explain technology use […], and advocates focusing on *relational actions* that occur *among* people and technologies. (Evans et al. 2017, 37; our emphasis).

Social media platforms are thus particularly useful for distinguishing content from affordances. They enable people to create and share digital messages, interact with each other, and build virtual communities. While initially designed to allow people to connect digitally with friends, today's social media have developed multiple, initially unforeseen functions (and use cases), including as marketing channels for businesses and organizations and as online redistributors of news. However, if we want to understand how social media affects us, culturally and socially, simply performing analyses of the content shared among its users will not suffice (although that, too, is important). Social media are now integral to our daily lives to the extent that a decision *not* to use social media has become a deliberate political or personal decision. What is more, social media have entered into rivalry with traditional media institutions (now increasingly referred to as "legacy media"). While the distribution of content through traditional media institutions is controlled by decisions made by human editors, social media delegates this responsibility to algorithms. These algorithms thus mediate between users and content, among other things, by predicting which connections are most relevant or interesting—or they can be intentionally biased toward the amplification of certain posts. We are thus dealing with "programmed forms of sociality" (Bucher 2018, 2). Social media—even well before current AI technologies—are thus "automated systems that engineer and manipulate connections" (van Dijck 2013, 12). While activities on social media are certainly complex, there seems to be a tendency for most algorithms of commercial social-media platforms, optimized for clicks, to privilege high-engagement and inflammatory contents that fuel animosity, which in turn facilitates political polarization (Aral 2020; Smith 2019; Vaidhyanathan 2021). Again, the structural features of commercial social media, their affordances, especially the way in which their algorithms are trained, are what expedite politically polarizing effects. Repeated exposure of social media users to inflammatory materials over time seems to shift the threshold of what is acceptable (the so-called Overton window, Amer 2023), thus serving to normalize

opinions and viewpoints that were previously considered, by most people, to be extreme or outrageous—yet another structural effect.

The concept of affordances is highly useful for our understanding of AI apps (Ahn 2025; Humble and Cajander 2025). The most important point to be made here is that generative media output has *emergent* features: Alternative terms proposed for generative content are thus *stochastic* or *probabilistic* (Ervik 2023). The "decisions" of the respective model are reducible to neither the programmers nor to users themselves. One of the consequences is that any AI model generates infinite variations of texts, images, videos, or sounds through the same prompt, while many emergent behaviors of the algorithms remain as much a black-box phenomenon to them as to the developers themselves (especially with regard to individual outputs and their "explainability" or "interpretability," Amodei 2025). The deluge of individual AI-generated files might then appear too vast and ephemeral to merit analysis of individual outputs; what matters instead are the general possibilities and constraints—the affordances—of the AI apps themselves. Even though any individual output will be a mere probability, the databases and the training can steer defaults in a certain (cultural, social, or political) direction:

> Whoever controls language models controls politics. […] Thus, in the future, language models themselves may take on the status of a surrogate public sphere. As more and more texts are generated by AI systems—and it is very likely that they will be—the proportion of discourse produced by humans will steadily decline. And because language models, once trained, are difficult to change, and because they infer the future from the past, what Emily Bender et al. [2021] has called a "value lock" looms. (Bajohr 2023b)

What Bajohr proposes here is that models always encode certain cultural, social, or political *values*—default assumptions and perspectives—and that they are, metaphorically speaking, "locked" after the model has been trained and entered public use. If that seemed perhaps far-fetched when it was written in 2023, only two years later, xAI's Grok is actively trained by its data annotators (reinforcement learning from human feedback, RLHF) toward an "anti-woke" agenda, avoiding topics like racism, climate change, or antisemitism. In other words, it is defaulted toward more right-wing, conservative views (Kay 2025), similar to when Google had tried to diversify its Gemini image outputs through RLHF fine-tuning and prompt "augmentation" a year before (Shamim 2024; Wilde 2026).

Affordances of AI apps—whether we decide to use them or not—thus create a new media reality to which people and institutions will have to react, one way or the other. Bajohr (2024) also considered that literary authors from now on will always be under suspicion of having used AI for their writing, even and perhaps *especially* when they categorically claim to abstain from AI use (Bajohr and Cotten 2025; Führer 2026). This will most certainly be reflected in a different writing style (greater emphasis on autofiction or less "probable" or "typical" diction), *regardless* of whether generative AI was in play or not, or whether we will ever know with certainty. In particular, the practice of "prompting" has perhaps already transformed the way we think.

> We have developed a new dialect, a specialized grammar of request making. We have learned to say "think step by step" because we have discovered it improves reasoning. We preface questions with role assignments: "You are an expert programmer" or "Act as a critical editor." We have absorbed the metacognitive practice of instructing the AI on how to think, which in turn makes us more conscious of how we ourselves think. (Panda 2026, 1)

This should remind us of Friedrich Nietzsche's insight from 1882 (2002, 18) that "our writing tools are also working on our thoughts" (quoted in and translated by Kittler 1999, 200). Swaroop Panda has called this "new form of thought" a "cognitive restructuring" (2026, 1). We have not even begun to understand how these newest of "writing tools" thus inform and change the very way we approach problems: "[W]e don't actually know what we've traded away. The cognitive skills we're losing are, by definition, the ones we're losing the capacity to notice we've lost. The researcher who no longer does deep, slow reading doesn't experience it as a loss—they experience it as efficiency" (Panda 2026, 2). Thus, simply because the affordances of AI apps exist now, they have effects on the production, distribution, and interpretation of *all* media forms, media aesthetics, media contents, and maybe even our thoughts.

Media Changes and Notions of Objectivity

We can see from this that media are not only a means of communication between individuals, but also shape our society as a whole. On top of that, media provide the only access to a broader, shared reality in the social, cultural, and political sense. Society at large is often understood as an "imagined community" (Anderson 1991), because we can never experience it directly or

personally, only through media that provide stories, reports, and other representations of it: "Whatever we know about our society, or indeed about the world in which we live, we know through the mass media," Niklas Luhmann (2000, 1) found in 1996. Today, this statement would have to be expanded from *mass media* to include all sorts of interpersonal (social) media forms—and also *AI media* (such as AI chatbots). Media, in any case, facilitate a shared, collective notion of a reality that always comes with the assumption and promise *not* to be false, fictional, or merely personal, but more or less "objective." This notion, however, can mean different things at different times, in different contexts, and relative to different media use cases. In other words, any notion of "objective reality" is subject to media changes, and to the protocols and expectations established towards them. We would like to outline this with the transitions, first, from an alphabetic to a technical media regime, and second, from an analog to a digital media regime.

Kittler conceives of history in terms of media shifts by introducing the term "discourse network," understood as a "network of technologies and institutions that allow a given culture to select, store, and process relevant data" (1990, 369). In other words, Kittler is interested in how information is made (and kept) available within a society, without concern for how humans assign meaning to it. He then compares the "discourse network" of romanticism (what he calls "discourse network 1800") with that of modernism ("discourse network 1900"). Like McLuhan before him, Kittler argues that a major historical media shift took place in the nineteenth century. However, while McLuhan connects this shift to the electronic age more broadly, Kittler identifies it with the invention of new recording and storage media, such as the gramophone/phonograph and photography/film (and also the typewriter). In Kittler's view, these media broke the "alphabetic monopoly" by technologizing information. While the discourse network of romanticism relied upon writing and typographic printing as a means of processing and storing information, the discourse network of modernism saw the introduction of media whose data flows no longer had to pass through the symbolic bottleneck of letters (Kittler 1999, 4). By circumventing the grid of the symbolic, these (then-new) media forms possessed a novel kind of evidential, "automatic" force that writing seemed to be lacking.

Such an ideal of *mechanical objectivity* (Daston and Galison 2007) has now returned with a vengeance in the contemporary discourses on AI: "[T]he imagined tools represent the epitome of a view from nowhere, or the idea that one can have objective knowledge of a set of truths, uncolored by their personal experience" (Bender and Hanna 2025, 119). Generative AI is indeed

often "seen as a utopian instrument of knowledge redistribution and the last kick against the old symbolic order" (Dubreuil 2025, 8). Artificial neural networks can be trained through *self-supervised* machine learning (next-token prediction on unlabeled text). Crucially, this is built on an architecture that is often called *subsymbolic*: "a neural network does not follow the paradigm of logical deduction or explicitly stated rules that are executed sequentially; rather, it operates by statistical induction, and it is the system as a whole that does the computing" (Bajohr 2021, 26). Many have embraced this assumption as convenient *political* rhetoric. The most prominent example here is certainly Donald Trump, who, on January 23, 2025, in one of his first executive orders after his reelection, did away with many existing policies aimed at regulating and safeguarding government use of AI. The justification was, once again, that AI systems were "free from ideological bias or engineered social agendas" (The White House 2025). The human "tutors" mentioned above, employed by companies like xAI to train Grok explicitly in a certain political (in this case right-wing, conservative) direction, clearly seem to undermine such stated ambitions towards a "maximally truth-seeking" AI (Elon Musk quoted in Dey 2025). In another instance in spring 2025, the same LLM had apparently been overtrained to mention "white people genocide in South Africa" as often as possible in completely unrelated conversations with shocked users (Kerr 2025). The promise of *more* objectivity through AI, however, remains undeniably powerful. The affordance to prompt Grok directly within social media (*X*) posts ("@Grok: Is it true?") to create automated public responses, fully integrated for all *X* users in February 2025, has quickly become a widely adopted protocol of online behavior. The AI has since then taken on the role of an (allegedly) "unbiased" judge to settle debates between users. Perhaps surprisingly (or not), Grok indeed publicly contradicted its "creator," Musk, many times on his own social media platform with reference to factual sources:

> Musk: "The far left is murderously violent;" Grok: "The claim that 'the left' is murderously violent isn't backed by evidence. The cited incidents […] show no clear ties to left-wing groups […]. No side is inherently violent; generalizations oversimplify a complex issue." (Musk 2025a)

Immediately, his followers then claimed that "Grok has been manipulated by leftist indoctrination unfortunately," with Musk personally promising to be "working on fixing that this week" (Musk 2025b; see Landymore 2025 for a longer discussion). The notion of "objectivity" thus turns out to be all the

more contested and political, the greater the promise to free it from human judgement. More fundamentally, it has also been argued that LLM output is, by default, decoupled from any real-world reality, producing "bullshit" in the sense of Harry Frankfurt (2005). *Bullshit*, as a technical term, can be distinguished from falsehoods or lies that *intentionally* aim to deceive and thus to conceal the truth; it requires "no conviction" (Frankfurt 2005, 55) whatsoever. The probabilistic outputs of chatbots are, by definition, unconstrained by any obligation to tell the truth (Kalai et al. 2025). They are "not trying to communicate something they believe or perceive. Their inaccuracy is not due to misperception or hallucination. As we have pointed out, they are not trying to convey information at all. They are bullshitting" (Hicks et al. 2024, 37; see also Berghel 2024; Coeckelbergh and Gunkel 2025, 89–102).

A critical view on flexible notions of objectivity is relevant not only for AI-generated text, but also for images. The use of photography as a scientific instrument helped foster a new epistemic ideal in the 19th century, that of *mechanical objectivity* (Daston and Galison 2007). The idea was that the mechanical origin of photographs (unlike earlier hand-crafted images) could eliminate subjective intervention between nature and representation, such as human judgement and interpretation. They could thus once again bypass the symbolic bottleneck—and hence the interface of human interpretation. Earlier picture technologies were thus considered more subjective (Rosson 2023). Analog photography had, for a long time, been the benchmark guaranteeing a "that has been" (Barthes 1981, 115), something that really existed and really happened. "Historically, photography has been associated with a realism characterized by neutral, unfeeling, disinterested recording (an association which is itself deeply misleading)" (Henning 2025, 94). This "that has been," as must be noted, has always been more of a social expectation than a guarantee, since there had been all sorts of manual image manipulation long before Photoshop (Fineman 2012). A commitment to objectivity remains a cornerstone virtue in journalism, but it has always needed to be safeguarded through institutions and rules, rather than technologies. In the realm of visual journalism, for example, we need standards against image manipulation. To secure the credibility and trustworthiness of photographic and video images, news agencies and press photographers' associations have developed codes of ethics for visual journalism. Although each news organization has to set its own ethical standards (Mäenpää 2014, 93), for example the code of ethics for visual journalists developed by the National Press Photographers Association (2025), there seems to be "a de facto global consensus on how media organizations understand

the manipulation of images" (Campbell 2014): the addition or subtraction of image elements designed to deceive or mislead the viewer. While none of this can be translated to AI imagery easily, it shows that the "reality effect" (Barthes 1986) of photography has always been more of an expectation or a "civil contract" (Azoulay 2008) that could always be broken; another "protocol," rather than a technological guarantee. Nevertheless, it had set a standard for what it means to *look* realistic and thus trustworthy (Turnock 2012).

The transition from analog to digital photography in the late 1980s and early 1990s, however, provoked a crisis in the perceived trustworthiness of photographic images (see, for example, Mitchell 1992), for two reasons. First, in contrast to analog photography, the data of digital pictures is subject to a new kind of symbolic grid. In this case, however, the data flows pass through the bottleneck of *numbers* (rather than, as in Kittler's earlier model, the bottleneck of alphabetic script). Second, the numerical representation of image data makes digital photographs subject to algorithmic manipulation (Cubitt 2014; Hoelzl and Marie 2015). The manipulability of images increased dramatically with the popular spread of image-editing software, such as Adobe Photoshop. As a reaction to these developments, there was a widespread sentiment in the 1990s that photography, as we had known it, was dead, and with it the peculiar power of (analog) photography to represent reality objectively (Mitchell 1992). However, through protocols and standards, digital photography obviously continues to be used in journalistic contexts: The "limits of digital photo editing" were accordingly defined "by referring to practices that were traditionally done in darkrooms" (Mäenpää 2014, 94). By holding on to established protocols from earlier media periods, digital photographs "are able to maintain their 'objective' status in the digital environment" (Mäenpää 2014, 101).

AI-generated photographs—or, more precisely, AI images that *look* like photographs—seem to have inherited many of these assumptions surrounding "reality effects," even though they could not be further detached from external reality: "For these models, the 'photographic' seems to be just another 'style', an aesthetic, a certain 'look', not a privileged mode of indexical access to the world," Roland Meyer (2023, 108) found (see also Meyer 2025 for more detail). Photorealism is thus nothing but a "stylistic signifier" (Ervik 2023, 47). This "signifier," however, activates the protocols and expectations associated with earlier photography technologies. As Lev Manovich (1996, 63) reminds us, the "reality" that is simulated in "realistic" computer graphics is not perceptual reality or embodied experience, but a photographed or filmed image (see also Ferwerda 2003 on "realism" in computer graphics). The highest standard of

realism in computer graphics is to *look as if* they were photographed. In other words, "realism" in this context means "photorealism" once again (Hausken 2024; Herrie and Philipsen 2026). The same holds true, and even more so, for AI imagery: "the naturalistic form of AI-generated photographs is what connects them to the discourse of photography, even if the reading of that form as natural is culturally conditioned" (Wasielewski 2024, 15). AI generators thus silently rely on an understanding of photography as "objective" images, often without ever making their reliance on this earlier image technology explicit. In many cases, we can only critically infer that computer science papers refer to photographic material when they talk about "natural" images (Barth 2025, 126–129; see also Malevé and Sluis 2023).

To conclude, if "photographic appearances" are seen as more "objective" than drawings (or other hand-crafted images), this is not some inherent property of photographic technology. It is the result, on the one hand, of rules and conventions of how (not) to use photographic technology; and, on the other, of the dominance of imagery that *looks* photographic (and thus more "neutral" or "objective" than drawings) in many media contexts. These are just appearances, though, because these images might also be computer-generated (in spectacle cinema), photoshopped (in advertising), or now perhaps increasingly AI-generated. Notions of objectivity are thus always dependent on specific sociocultural protocols and conventions, especially of earlier, already established media forms (in this case: analog or digital photography).

Media Changes and Cultural, Social, and Political Transformations

As we have seen, discussions of cultural, social, and even political transformations come up again and again whenever new technologies introduce significant changes to mediations—and hence relations—among people and institutions in a society. While we can investigate the affordances and protocols of (new) media for private, interpersonal communication and interaction, perhaps even more important are the organizational and institutional uses that go beyond the individual. The argument that a new medium for communication and interaction has transformative effects on culture, society, and even politics is not new, but has been made for many earlier media shifts. That said, asking about the broader long-term (or even middle-term) impacts of AI is certainly highly speculative at this point (Morozov 2025). We would nevertheless like to introduce a perspective that has always been

highly relevant within media theory, and ask, at the very least, what conclusions we can currently draw about the present. If media theorists are right in their claim that media shape our situation—or, in the words of W. J.T. Mitchell and Mark B.N. Hansen, "that media *are* our situation" (2010, xxii; original emphasis)—it makes sense to understand long-term trends in world history through the lens of major media shifts. Harold Innis (1950; 1951), for example, emphasized how the material characteristics of communication media shaped the development of civilizations (see also Ong 2002). Media that were durable but difficult to transport, such as stone and clay, were accordingly biased toward time. This implies that they facilitate stability, tradition, and the formation of consensus-based tribal societies. The situation was quite different with media that were less durable but easily portable, such as papyrus and paper: they were biased toward space, which means that they favored rapid change, materialism, and the formation of large empires.

McLuhan, too, analyzed history in terms of major media shifts. He placed special emphasis first on the phonetic alphabet, then on typography, and finally on the electronic age—each of which introduced a new kind of human condition into society (see McLuhan 1962, 18). Electronic media, such as the telegraph—and later, radio and television—made the world contract spatially into what McLuhan calls a "global village" (1962, 21). This means not so much that people suddenly came to sympathize or agree with each other; rather, a new electronic interdependence made the world shrink in such a way that whatever happens in one part of the world inevitably affects what happens in others. What is more, McLuhan maintains that the introduction of electronic media, with their pressure of simultaneity and all-at-onceness, is leading to a retribalization of society, and hence to a return to oral and auditory forms of thought and organization (McLuhan 1962, 72). Media theorists have also emphasized the transformational power of television, in particular. A famous example was the first televised presidential debate between Senator John F. Kennedy and Vice President Richard M. Nixon in 1960, which, at least according to popular myth, was won by Kennedy not because of what he *said*, but because he cut a better figure on the TV screen (Barnouw 1990, 270–277). Since then, the argument has often been made that television helped expedite a transition, in liberal democracies, from issue-centered politics to politics that are much more person-centered. With his slogan "the medium is the message," McLuhan thus anticipated today's discussion in media studies and related fields of the

nonhuman or more-than-human *agency* of technologies and media through their specific affordances (see, for example, Zylinska 2017). This shift in attention from the content of media to their structural features—to their affordances, especially with regard to (re)configuration of cultural, social, and political agency—brings a new perspective from which to analyze such transformations. As a reminder, whenever we talk about mediation, we try to investigate the network of human and non-human "actors" (participants) and the distribution of agency between them, alongside affordances that are enabled (or limited) by media.

Are we talking about yet another major media shift with regard to generative AI and LLMs—a new "media age" so to speak—and if so, what are the transformative effects on society, culture, and politics beyond interpersonal uses (see Coeckelbergh 2022 for an overview)? Immediately after the widespread adoption of ChatGPT, journalists began to speculate about its potential future impacts. *MIT Technology Review*, for example, imagined "six ways that AI could change politics" (Schneier and Sanders 2023). They did so by listing a range of future "milestones" where the individual production of AI-generated texts would transition into institutional production of such texts: when an AI-generated testimony would be accepted in court, for example, or when the amendment to a legal bill would be written by AI. Such transitions from interpersonal to institutional uses are certainly worth observing closely (Schröter 2024). In Albania, one virtual AI assistant named "Diella" (integrated within the e-Albania platform and co-developed by Microsoft) was appointed "Minister of State for Artificial Intelligence" in September 2025, to the concern of many journalists and politicians (Delauney 2025). Exactly how this will work legally still remains largely unclear (Maciel Perez and Démas 2025).

While we want to avoid speculating too far ahead, we can already make some observations about the present and the redistribution of agency in society today.[2] We should pay particular attention, for example, when AI technologies serve as justification to reshape existing cultural, social, and

---

2 As we have noted in the previous footnote, the following observations too would look different when they could have been written in 2026 now that Anthropic has been declared a "supply chain risk" by the US Pentagon and Claude had been instrumental in both the operation to kidnap Nicolás Maduro from Venezuela and in the US military's strikes in Iran. See The Ezra Klein Show 2026 about the frightening new interdependencies between politics, military, legal system, and "AI media."

political institutions and structures (Coeckelbergh 2024, 2026; Eubanks 2018; McQuillan 2022). One particularly consequential moment for this even before the appointment of "Diella" was when Donald Trump announced his "AI-first strategy" and began the immediate dismantling of government institutions (and human decision-making) in the name of "Government Efficiency." As Matthew Kirschenbaum (2025, 3) argues in his critical response "US of AI," these actions aim at invoking "a fundamental change in the material base of government infrastructure, a shift whose ultimate goal is doubtless [...] a 'government without people,'" exactly by "replacing the ethos of neutrality and objectivity" (discussed before) from an institution-centered understanding to an AI-centered one. "AI then becomes a tool for replacing politics" (Salvaggio 2025). What is being automated is thus not only paperwork, but also democratic decision-making. By shifting government processes to AI systems, such as OpenAI's "ChatGPT Gov" or GSAi, a custom AI chatbot for the US General Services Administration, the aim is to replace bureaucracy with technological infrastructure. "The rationale is based on a productivity myth that the goal of bureaucracy is merely what it produces (services, information, governance) and can be isolated from the process through which democracy achieves those ends: debate, deliberation, and consensus" (Salvaggio 2025). In the case of "Diella," too, the rhetoric behind the implementation of the AI assistant was to "eliminate corruption" (Delauney 2025).

Linguist Emily M. Bender and sociologist Alex Hanna (2025, 67–100) have documented dozens of cases from the last couple of years, mainly within the US and Great Britain, but on all levels of local, regional, or national administration, where AI services have already been implemented in order to cut costs. Their survey shows that almost without exception, AI chatbots make for a poor substitute for human administration, and that their dangers and risks are disproportionately severe for minority groups: "shunting consequential tasks to black-box machines trained on always-biased historical data is not a viable solution for any kind of just and accountable outcome" (Bender and Hanna 2025, 76). The real question about the agency of AI technologies, then, should not be whether or not they "work well" or even whether they can be "trusted," but rather who *controls* and *benefits* most from them, when private contractors in Silicon Valley (commercial companies like OpenAI or Microsoft) within a new "computational capitalism" (Berry 2025, 5258) slowly become unelected arbiters of cultural, social, and political decision-making (Dyer-Witheford and Kjøsen 2019).

For most of us, in most situations, a distinction between AI agency and human agency, and between AI-generated and human-generated media content, is thus still fundamental—even and especially when this difference becomes harder and harder to distinguish in practice. In the not-too-distant future, however, every media artifact might be co-created by generative AI *to a certain degree*. Bajohr (2023a) spoke of a "postartificial" media reality in which the very distinction between AI-generated and human-generated content might lose its importance, because we will *always* assume this to *some* degree. This would then be equal to whether or not "electricity" played a part in the production, distribution, and reception of something, which is hardly worth noting or asking. Such a future would probably constitute the biggest changes across society, culture, and politics when "the difference between natural and artificial has really disappeared altogether" (Bajohr 2023a, 28). At the moment, however, this distinction seems all the more important the harder it becomes to draw—as does the question of which human or institutional actors are actually profiting or suffering most from the implementation of AI technologies and the affordances they provide.

What all of this perhaps points to is the fact that we should not only investigate the technological side of generative AI apps but always trace how, where, and when these technologies gradually "aggregate" into distinct media forms of communication and interaction. One way "to maintain critical reflexivity while engaging with computational systems that increasingly shape the very grounds of consciousness and experience" (Berry 2025, 5267), is to investigate how sociocultural protocols—conventions of unquestioned media behavior and media use—slowly couple media institutions, media roles, media functions, media values, and media economies. We have proposed looking at the affordances of AI apps and models as having structural impacts on communication and interaction beyond the individual output they create; and that these affordances can go beyond individual and interpersonal use and instead influence the broader distribution of social, cultural, and political agency—as well as our notions of a shared reality that can only ever be accessed through media. To what degree our "imagined communities"—our shared or contested notions of the society in which we live—will still come from humans; will be shaped, or limited, by commercial companies controlling AI apps and models; or will be "imagined" by artificial neural networks and their affordances, will be one of the most pressing concerns for the 21st century.

Works Cited

Ahn, Sungyong. 2025. "Affordance of Generative AI: For a Posthuman Tool Analysis." *Convergence: The International Journal of Research into New Media Technologies,* October 28, 2025: 1–18. https://doi.org/10.1177/13548565251393 87.

Amer, Sarah K. 2023. "AI Imagery and the Overton Window." Preprint, arXiv, May 31, 2023. https://doi.org/10.48550/arXiv.2306.00080.

Amodei, Dario. 2025. "The Urgency of Interpretability." *Dario Amodei*, April, 2025. https://www.darioamodei.com/post/the-urgency-of-interpretability.

Anderson, Benedict. 1991. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. 2nd ed. London: Verso.

Aral, Sinan. 2020. *The Hype Machine: How Social Media Disrupts Our Elections, Our Economy, and Our Health—and How We Must Adapt*. New York: Currency.

Azoulay, Ariella. 2008. *The Civil Contract of Photography*. New York: Zone Books.

Bajohr, Hannes. 2021. "The Gestalt of AI: Beyond the Holism-Atomism Divide." *Interface Critique* 3: 13–35. https://doi.org/10.11588/IC.2021.3.81304.

Bajohr, Hannes. 2023a. "Artificial and Post-Artificial Texts: On Machine Learning and the Reading Expectations Towards Literary and Non-Literary Writing." *Media Culture and Cultural Techniques: Working Papers* 2023 (7): 3–31. https://doi.org/10.5451/bmcct.2023.007.

Bajohr, Hannes. 2023b. "Whoever Controls Language Models Controls Politics." *Hannes Bajohr*, April 8, 2023. https://hannesbajohr.de/en/2023/04/08/whoever-controls-language-models-controls-politics.

Bajohr, Hannes. 2024. "Distanzgrade: Autorschaft und KI" Lecture given at the symposium "Correlations – AI in Art and Design," Hochschule für Gestaltung Offenbach am Main, Germany, December 12, 2024.

Bajohr, Hannes and Ann Cotten, eds. 2025. *Schreiben nach KI*. Berlin: Rohstoff.

Baker, Pam. 2025. *ChatGPT for Dummies*. Weinheim: Wiley.

Bakir, Vian and Andrew McStay. 2025. "Move Fast and Break People? Ethics, Companion Apps, and the Case of Character.ai." *AI & Society, Open Forum*, June 10, 2025. https://doi.org/10.1007/s00146-025-02408-5.

Barale, Alice. 2024. *The Art of Artificial Intelligence: Philosophical Keywords*. Cambridge: Cambridge Scholars Publishing.

Barnouw, Erik. 1990. *The Tube of Plenty: The Evolution of American Television*. 2nd edition. New York: Oxford University Press.

Barth, Franziska. 2025. "Residual Images." *Zfm: Zeitschrift für Medienwissenschaft* 33: 121–132.

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

Barthes, Roland. 1986. "The Reality Effect." *The Rustle of Language*. Translated by Richard Howard, 141–148. New York: Hill and Wang.

Baudelaire, Charles. [1859] 1980. "The Modern Public and Photography." In *Classic Essays on Photography*, edited by Alan Trachtenberg, 83–89. New Haven: Leete's Island Books.

Bell, David. 2026. "AI's Second Moment: The Explosion That Changed Everything." *Medium*, March 18, 2026. https://medium.com/@drdavidwbell/ais-second-moment-the-explosion-that-changed-everything-3e67e979dc4c.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜" *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922.

Bender, Emily M. and Alex Hanna. 2025. *The AI Con: How to Fight Big Tech's Hype and Create the Future We Want*. New York: HarperCollins.

Berghel, Hal. 2024. "Generative Artificial Intelligence, Semantic Entropy, and the Big Sort." *Computer* 57 (1): 130–135. https://doi.org/10.1109/MC.2023.3331594.

Berry, David M. 2025. "Synthetic Media and Computational Capitalism: Towards a Critical Theory of Artificial Intelligence." *AI & Society* 40: 5257–5269. https://doi.org/10.1007/s00146-025-02265-2.

Bucher, Taina. 2018. *If… Then: Algorithmic Power and Politics*. Oxford: Oxford University Press.

Bucher, Taina and Anne Helmond. 2017. "The Affordances of Social Media Platforms." In *The SAGE Handbook of Social Media*, edited by Jean Burgess, Thomas Poell, and Alice Marwick, 233–253. London: SAGE.

Campbell, David. 2014. "Report: The Integrity of the Image—A World Press Photo Research Project." *World Press Photo*, November, 2014. World Press Photo Academy. https://www.david-campbell.org/photography/integrity-of-the-image.

Chen, Nuo, Yang Deng, and Jia Li. 2024. "The Oscars of AI Theater: A Survey on Role-Playing with Language Models." Preprint, arXiv, July 16, 2024. https://doi.org/10.48550/arXiv.2407.11484.

Chesher, Chris and César Albarrán-Torres. 2023. "The Emergence of Autolography: The 'Magical' Invocation of Images from Text through AI." *Media International Australia* 189 (1): 57–73. https://doi.org/10.1177/1329878X231193252.

Coeckelbergh, Mark. 2022. *The Political Philosophy of AI: An introduction*. Cambridge: Polity Press.

Coeckelbergh, Mark. 2024. *Why AI Undermines Democracy and What To Do About It*. Cambridge: Polity Press.

Coeckelbergh, Mark. 2026. "Technofascism: AI, Big Tech, and the New Authoritarianism." AI & Society. January 25, 2026. https://doi.org/10.1007/s00146-026-02862-9.

Coeckelbergh, Mark and David J. Gunkel. 2025. *Communicative AI: A Critical Introduction to Large Language Models*. Cambridge: Polity Press.

Crawford, Kate. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.

Cubitt, Sean. 2014. *The Practice of Light: A Genealogy of Visual Technologies from Prints to Pixels*. Cambridge, MA: The MIT Press.

Daston, Lorraine and Peter L. Galison. 2007. *Objectivity*. New York: Zone Books.

Davis, Jenny L. 2020. *How Artifacts Afford: The Power and Politics of Everyday Things*. Cambridge, MA: The MIT Press.

Delauney, Guy. 2025. "World's First AI Minister will Eliminate Corruption, Says Albania's PM." *BBC*, September 12, 2025. https://www.bbc.com/news/articles/cm2znzgwj3xo.

Dey, Victor. 2025. "Behind xAI Grok 3, Elon Musk's 'Maximally Truth-Seeking A.I.'" *The Observer*, February 20, 2025. https://observer.com/2025/02/elon-musk-xai-grok3.

Dhaliwal, Ranjodh Singh. 2023. "What Do We Critique When We Critique Technology?" *American Literature* 95 (2): 305–319. https://doi.org/10.1215/00029831-10575091.

Dubreuil, Laurent. 2025. *Humanities in the Time of AI*. Minneapolis: University of Minnesota Press.

Dwivedi, Yogesh K., et al. 2023. "So What If ChatGPT Wrote It? Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy." *International Journal of Information Management* 71: August, 2023. https://www.sciencedirect.com/science/article/pii/S0268401223000233.

Dyer-Witheford, Nick and Atle Mikkola Kjøsen. 2019. *Inhuman Power: Artificial Intelligence and the Future of Capitalism*. London: Pluto Press.

Ervik, Andreas. 2023. "Generative AI and the Collective Imaginary: The Technology-Guided Social Imagination in AI-Imagenesis." *Generative Imagery: Towards a 'New Paradigm' of Machine Learning-Based Image Production*, Special-Themed issue of *IMAGE: The Interdisciplinary Journal of Image Sciences* 37 (1): 42–57.

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Evans, Sandra K., Katy E. Pearce, Jessica Vitak, and Jeffrey W. Treem. 2017. "Explicating Affordances: A Conceptual Framework for Understanding Affordances in Communication Research." *Journal of Computer-Mediated Communication* 22 (1): 35–52. https://doi.org/10.1111/jcc4.12180.

Feher, Katalin. 2025. *Generative AI, Media, and Society*. London: Routledge.

Fendler, Ute, Patrice Mwepu, and Ivo Ritzer. 2024. "Introduction." In *Medialities*, edited by Ute Fendler, Patrice Mwepu, and Ivo Ritzer, 7–11. Bielefeld: transcript.

Ferwerda, James A. 2003. "Three Varieties of Realism in Computer Graphics." *Proceedings of SPIE: The International Society for Optical Engineering* 5007. https://cin.ufpe.br/~in1123/material/vor_hvei03_v20.pdf.

Filimowicz, Michael. 2023. *AI and the Future of Creative Work: Algorithms and Society*. London: Routledge.

Fineman, Mia. 2012. *Faking It: Manipulated Photography before Photoshop*. New York: Metropolitan Museum of Art.

Frankfurt, Harry. 2005. *On Bullshit*. Princeton: Princeton University Press.

Führer, Carolin. 2026. "Understanding Literature in the Age of AI." In *Humanities Perspectives on AI: An Interdisciplinary Introduction*, edited by Lukas R.A. Wilde, Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, 221–240. Oslo: Scandinavian University Press.

Galloway, Alexander. 2004. *Protocol: How Control Exists after Decentralization*. Cambridge, MA: The MIT Press.

Gibson, James J. 1977. "The Concept of Affordances." In *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, edited by Robert Shaw and John D. Bransford, 67–81. New York: John Wiley & Sons.

Gitelman, Lisa. 2006. "Introduction: Media as Historical Subjects." In *Always Already New. Media History and the Data of Culture*, 1–22. Cambridge, MA: The MIT Press.

Gray, Mary L. and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston: Houghton Mifflin Harcourt.

Gunkel, David J. 2020. *An Introduction to Communication and Artificial Intelligence*. Cambridge: Polity.

Guzman, Andrea L., ed. 2018. *Human-Machine Communication: Rethinking Communication, Technology, and Ourselves*. New York: Peter Lang.

Harder, Lina Ruth. 2025. "Histobots: Rewriting History, One Chat at a Time." Lecture given at University: Future Festival, May 13–15, 2025. https://festival.hfd.digital/de/sessions-2025/?id=842001.

Hausken, Liv. 2024. "Photorealism versus Photography: AI-Generated Depiction in the Age of Visual Disinformation." *Journal of Aesthetics & Culture* 16 (1): 1–13. https://doi.org/10.1080/20004214.2024.2340787.

Henning, Michelle. 2025. "Affective Realism: Reimagining Photography with the Google Pixel 9." In *AI Aesthetics: AI-Generated Images between Artistics and Aisthetics*, edited by Jan-Noël Thon and Lukas R.A. Wilde, 92–105. New York: Routledge.

Herrie, Maja Bak and Lotte Philipsen. 2026. "'No Children Were Harmed': Aesthetic Infrastructures and Ethical Implications of AI-generated Portraits." *Visual Studies* 2026: 1–12. https://doi.org/10.1080/1472586X.2025.2606201.

Hicks, Michael Townsen, James Humphries, and Joe Slater. 2024. "ChatGPT is Bullshit." *Ethics and Information Technology* 26 (2): 38. https://doi.org/10.1007/s10676-024-09775-5.

Hoelzl, Ingrid and Rémi Marie. 2015. *Softimage: Towards a New Theory of the Digital Image*. Bristol: Intellect.

Hong, Seungbeom and Jinkyung Na. 2018. "How Facebook Is Perceived and Used by People Across Cultures: The Implications of Cultural Differences in the Use of Facebook." *Social Psychological and Personality Science* 9 (4): 435–443. https://doi.org/10.1177/1948550617711122.

Hopkins, Julian. 2016. "The Concept of Affordances in Digital Media." In *Handbuch Soziale Praktiken und Digitale Alltagswelten*, edited by Heidrun Friese, Marcus Nolden, Gala Rebane, and Miriam Schreiter. Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-658-08460-8_67-1.

Humble, Niklas and Åsa Cajander. 2025. "Rethinking Generative AI in Human-Computer Interaction: Review of Affordances, Constraints, and Implications." *Proceedings of the 2nd International Conference on Education Research - ICER 2025* 2 (1), October 31, 2025: 1–18. https://doi.org/10.34190/icer.2.1.3926.

Hutchby, Ian. 2001. "Technologies, Texts and Affordances." *Sociology* 35 (2): 441–456. https://doi.org/10.1177/S0038038501000219.

Innis, Harold. 1950. *Empire and Communications*. Oxford: Oxford University Press.

Innis, Harold. 1951. *The Bias of Communication*. Toronto: University of Toronto Press.

Jensen, Klaus Bruhn. 2010. *Media Convergence: The Three Degrees of Network, Mass, and Interpersonal Communication*. London: Routledge.

Jin, Dal Yong. 2021. *Artificial Intelligence in Cultural Production: Critical Perspectives on Digital Platforms*. London: Routledge.

Kalai, Adam Tauman, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. "Why Language Models Hallucinate." Preprint, arXiv, September 4, 2025. https://doi.org/10.48550/arXiv.2509.04664.

Kay, Grace. 2025. "Inside Grok's War on 'Woke.'" *Business Insider*, February 28, 2025. https://www.businessinsider.com/xai-grok-training-bias-woke-idealogy-2025-02.

Kędra, Joanna. 2016. "Enhancing Visual Literacy through Interpretation of Photo-genres: Toward a Genre Typology of Journalistic Photographs." *Journal of Media Practice* 17 (1): 28–47. https://doi.org/10.1080/14682753.2016.1159451.

Kelly, Samantha Murphy. 2024. "When Grief and AI Collide: These People are Communicating with the Dead." *CNN Business*, May 6, 2024. https://edition.cnn.com/2024/05/06/tech/ai-communicating-with-dead/index.html.

Kerr, Dara. 2025. "Musk's AI Grok Bot Rants about 'White Genocide' in South Africa in Unrelated Chats." *The Guardian*, May 15, 2025. https://www.theguardian.com/technology/2025/may/14/elon-musk-grok-white-genocide.

Khan, Ian. 2024. *The Quick Guide to Prompt Engineering: Generative AI Tips and Tricks for ChatGPT, Bard, Dall-E, and Midjourney*. Hoboken: Wiley.

Kim, Soojung, Haseon Park, and Joonghwa Lee. 2021. "Snapchat Usage from the International Perspective: Comparison between the USA and South Korea." *International Journal of Mobile Communication* 19 (6): 734–751. https://doi.org/10.1504/IJMC.2021.118592.

Kirschenbaum, Matthew. 2025. "The US of AI." Public Draft, February 25, 2025. https://drive.google.com/file/d/1O2qkjhg7Ei5zZWmBraNwXq4V0lTauspN/view?fbclid=IwY2xjawIta99leHRuA2FlbQIxMQABHaYasYRdObXQMDhxLA663f-olOolfNYK5ZXWLyBJxOBGkuKu_ol9i6d65A_aem_vFeS34lZx5krTC3F24tqGQ.

Kittler, Friedrich A. 1990. *Discourse Networks 1800/1900*. Translated by Michael Metteer, with Chris Cullens. Foreword by David E. Wellbery. Stanford: Stanford University Press.

Kittler, Friedrich A. 1999. *Gramophone, Film, Typewriter*. Translated, with an introduction, by Geoffrey Winthrop-Young and Michael Wutz. Stanford: Stanford University Press.

Lamerichs, Nicolle. 2023. "Character.ai and Chatbots as the Next Step in Fandom." *Nicolle Lamerichs: Fandom & Media Insights*. November 29, 2023. https://nicollelamerichs.com/2023/11/29/character-ai-chatbots-as-the-next-step-in-fandom.

Landymore, Frank. 2025. "Elon Musk Concerned That Reality Has Infiltrated His 'Grok' AI." *Futurism*, June 20, 2025. https://futurism.com/elon-musk-infilitrated-grok-ai.

Li, Lingyao, Zihui Ma, Lizhou Fan, Sanggyu Lee, Huizi Yu, and Libby Hemphill. 2024. "ChatGPT in Education: A Discourse Analysis of Worries and Concerns on Social Media." *Education and Information Technologies* 29. https://doi.org/10.1007/s10639-023-12256-9.

Luhmann, Niklas. 2000. *The Reality of Mass Media*. Translated by Kathleen Cross. Cambridge: Polity Press.

Mäenpää, Jenni. 2014. "Rethinking Photojournalism: The Changing Work Practices and Professionalism of Photojournalists in the Digital Age." *Nordicom Review: Nordic Research on Media and Communication* 35 (2): 91–104. https://doi.org/10.2478/nor-2014-0017.

Maciel Perez, Luisa and Martin Démas. 2025. "The Human Crisis of AI Exceptionalism: the case of Diella, Albania's 'AI Minister.'" *The European Master in Law, Data and Artificial Intelligence (EMILDAI)*, October 3, 2025. https://emildai.eu/the-human-crisis-of-ai-exceptionalism-the-case-of-diella-albanias-ai-minister/.

Malevé, Nicolas and Katrina Sluis. 2023. "The Photographic Pipeline of Machine Vision; or, Machine Vision's Latent Photographic Theory." *Critical AI* 1 (1–2). https://doi.org/10.1215/2834703X-10734066.

Manovich, Lev. 1996. "The Paradoxes of Digital Photography." In *Photography after Photography: Memory and Representation in the Digital Age*, edited by Hubertus v. Amelunxen, Stefan Iglhaut, and Florian Rötzer, 57–65. Amsterdam: G+B Arts.

McLuhan, Marshall. 1962. *The Gutenberg Galaxy: The Making of Typographic Man*. Toronto: University of Toronto Press.

McLuhan, Marshall. [1964] 2001. *Understanding Media: The Extensions of Man*. London: Routledge.

McQuillan, Dan. 2022. *Resisting AI: An Anti-fascist Approach to Artificial Intelligence*. Bristol: Bristol University Press.

Meikle, Graham and Sherman Young. 2012. *Media Convergence: Networked Digital Media in Everyday Life*. Basingstoke: Palgrave Macmillan.

Meyer, Roland. 2023. "The New Value of the Archive: AI Image Generation and the Visual Economy of 'Style.'" *Generative Imagery: Towards a 'New Paradigm' of Machine Learning-Based Image Production*, Special-Themed Issue of *IMAGE: The Interdisciplinary Journal of Image Sciences* 37 (1): 100–111.

Meyer, Roland. 2025. "Images from Images: Generative AI and the Reconfiguration of the 'Photographic.'" *Photography & Culture* 18 (3–4): 281–290. https://doi.org/10.1080/17514517.2025.2522517.

Mitchell, William J.T. 1992. *The Reconfigured Eye: Visual Truth in the Post-Photographic Era*. Cambridge, MA: The MIT Press.

Mitchell, William J.T. and Mark B.N. Hansen. 2010. "Introduction." In *Critical Terms for Media Studies*, edited by William J.T. Mitchell and Mark B.N. Hansen, vii–xxii. Chicago: University of Chicago Press.

Morozov, Evgeny. 2025. *AI Futures*. Boston: Boston Review.

Musk, Elon [@elonmusk]. 2025a. "The far left is murderously violent," posted on X, June 14, 2025. https://x.com/MuscularSack/status/1933952236775985503.

Musk, Elon [@elonmusk]. 2025b. "I know. Working on fixing that this week," posted on X, June 15, 2025. https://x.com/elonmusk/status/1934105717662962136.

Natale, Simone. 2021. "Communicating Through or Communicating with: Approaching Artificial Intelligence from a Communication and Media Studies Perspective." *Communication Theory* 31 (4): 905–910. November 2021. https://doi.org/10.1093/ct/qtaa022.

National Press Photographers Association. 2025. "Code of Ethics" *NPPA*. https://nppa.org/resources/code-ethics.

Nietzsche, Friedrich. [1882] 2002. "Friedrich Nietzsche an Heinrich Köselitz, Ende Februar 1882." In *Schreibmaschinentexte. Vollständige Edition, Faksimiles und kritischer Kommentar.* Edited by Stephan Günzel and Rüdiger Schmidt-Grépály. Weimar: Bauhaus-Universitätsverlag.

Ong, Walter J. 2002. *Orality and Literacy: The Technologizing of the Word*. London: Routledge.

Panda, Swaroop. 2026. "We Become What We Prompt." *AI & Society*, February 27, 2026. https://doi.org/10.1007/s00146-026-02883-4.

Pasquinelli, Matteo. 2023. *The Eye of the Master: A Social History of Artificial Intelligence*. London: Verso.

Philipsen, Lotte. 2025. "Aesthetic Protocols of Popular AI Art." In *AI Aesthetics: AI-Generated Images between Artistics and Aisthetics*, edited by Jan-Noël Thon and Lukas R.A. Wilde, 59–74. New York: Routledge.

Rajewsky, Irina O. 2010. "Border Talks: The Problematic Status of Media Borders in the Current Debate about Intermediality." In *Media Borders, Multimodality and Intermediality*, edited by Lars Elleström, 51–68. Basingstoke: Palgrave Macmillan.

Rosson, Lois. 2023. "What Is AI Doing To Art?" *Noema Magazine*, April 2023. https://www.noemamag.com/what-is-ai-doing-to-art.

Salvaggio, Eryk. 2025. "Anatomy of an AI Coup." *Tech Policy Press*, February 09, 2025. https://www.techpolicy.press/anatomy-of-an-ai-coup.

Schneier, Bruce and Nathan E. Sanders. 2023. "Six Ways that AI could Change Politics." *MIT Technology Review*, July 28, 2023. https://www.technologyreview.com/2023/07/28/1076756/six-ways-that-ai-could-change-politics.

Schröter, Jens. 2024. "Anmerkungen zur 'KI-Revolution'." *PROKLA: Zeitschrift für kritische Sozialwissenschaft* 54 (217): 679–690. https://doi.org/10.32387/prokla.v54i217.2140.

Shamim, Sarah. 2024. "Why Google's AI Tool was Slammed for Showing Images of People of Colour." *Aljazeera,* March 9, 2024. https://www.aljazeera.com/news/2024/3/9/why-google-gemini-wont-show-you-white-people.

Shanahan, Murray, Kyle McDonell, and Laria Reynolds. 2023. "Role Play with Large Language Models." *Nature* 623 (7987): 493-498. https://doi.org/10.1038/s41586-023-06647-8.

Smith, Robert Elliott. 2019. *Rage Inside the Machine: The Prejudice of Algorithms, and How to Stop the Internet Making Bigots of Us All*. London: Bloomsbury Business.

Stackpole, Beth. 2026. "Agentic AI, explained." *MIT Management Sloan School*, February 18, 2026. https://mitsloan.mit.edu/ideas-made-to-matter/agentic-ai-explained.

Striano, Francesco. 2025. *Through the Screen: Towards a General Philosophy of Mediality*. Berlin: De Gruyter.

Sørensen, Knut H. 2005. "Domestication: The Enactment of Technology." In *Domestication of Media and Technology*, edited by Thomas Berker, Maren Hartmann, Yves Punie, and Katie Ward, 40–61. Maidenhead: McGraw-Hill Education.

The Ezra Klein Show. 2026. "Who Should Control A.I.?." *The New York Times*, March 6, 2026. https://www.nytimes.com/video/opinion/100000010747008/who-should-control-ai.html.

The White House. 2025. "Presidential Actions Removing Barriers to American Leadership in Artificial Intelligence." *The White House*, January 23, 2025. https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/.

Thon, Jan-Noël. 2014. "Mediality." In *The Johns Hopkins Guide to Digital Media*, edited by Ryan, Marie-Laure, Benjamin J. Robertson, and Lori Emerson, 334–336. Baltimore: Johns Hopkins University Press.

Turnock, Julie. 2012. "The ILM Version: Recent Digital Effects and the Aesthetics of 1970s Cinematography." *Film History* 24 (2): 158–168. https://doi.org/10.2979/filmhistory.24.2.158.

Vaidhyanathan, Siva. 2021. *Antisocial Media: How Facebook Disconnects Us and Undermines Democracy*. New York: Oxford University Press.

van Dijck, Jose. 2013. *The Culture of Connectivity: A Critical History of Social Media*. Oxford: Oxford University Press.

Wasielewski, Amanda. 2024. "Unnatural Images: On AI-Generated Photographs." *Critical Inquiry* 51 (1): 1–29. https://doi.org/10.1086/731729.

Watkins, Marc. 2024. "Recapping OpenAI's Education Forum" *Rhetorica: Notes from a Nontraditional Student Turned Educator: Culture, AI, Education*, October 14, 2024. https://marcwatkins.substack.com/p/recapping-openais-education-forum.

Wilde, Lukas R.A. 2023. "Generative Imagery as Media Form and Research Field: Introduction to a New Paradigm" *Generative Imagery: Towards a 'New Paradigm' of Machine Learning-Based Image Production*, Special-Themed Issue of *IMAGE: The Interdisciplinary Journal of Image Sciences* 37 (1): 6–33. http://dx.doi.org/10.25969/mediarep/22326.

Wilde, Lukas R.A. 2025a. "Aesthetics and Rhetorics of AI Anthropomorphization: The Eliza Effect vs. The Character Effect." In *AI Aesthetics: AI-Generated Images between Artistics and Aisthetics*, edited by Jan-Noël Thon and Lukas R.A. Wilde, 106–123. New York: Routledge.

Wilde, Lukas R.A. 2025b. "KI-Bilder und die Widerständigkeit der Medienkonvergenz: Von primärer zu sekundärer Intermedialität?" In *Bilder im Aufbruch: Herausforderungen der Bildwissenschaft*, edited by Marcel Lemmes, Stephan Packard, and Klaus Sachs-Hombach, 475–507. Cologne: Herbert von Halem.

Wilde, Lukas R.A. 2026. "AI Imagery." In *Humanities Perspectives on AI: An Interdisciplinary Introduction*, edited by Lukas R.A. Wilde, Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, 173–201. Oslo: Scandinavian University Press.

Wirth, Sabine. 2023. "Interfaces of AI: Two Examples from Popular Media Culture and Their Analytical Value for Studying AI in the Sciences." In *Beyond Quantity: Research with Subsymbolic AI*, edited by Andreas Sudmann, Anna Echterhölter, Markus Ramsauer, Fabian Retkowski, Jens Schröter, and Alexander Waibel, 217–233. Bielefeld: transcript.

Zylinska, Joanna. 2017. *Nonhuman Photography*. Cambridge, MA: The MIT Press.

8. AI Imagery

Lukas R.A. Wilde

Abstract: This chapter opens with a technically grounded account of how AI image generation actually works. A media studies perspective is then chosen for introducing three interrelated areas of concern, namely issues connected to their production, their representation, and their distribution. The chapter first traces the diffuse distribution of authorship across human artists, datasets, algorithms, and corporate actors, examining the unresolved questions of copyright and consent around training data. It then examines how racial, gender, and cultural biases are often baked into image models at multiple levels and discusses the concept of *platform realism*—the tendency of AI imagery to optimize not for authenticity, but for plausibility and broad approval. The chapter then closes with the productive analytical pair of *immediacy*, image distribution that conceals its AI origin versus *hypermediacy*, a deliberate foregrounding of the AI generation as spectacle. Taken together, the chapter advocates always assessing AI imagery in relation to the cultural, social, and political contexts in which it appears.

Keywords: cultural memory, deepfakes, diffusion models, memes, remediation

Machine-learning technologies have been a part of digital image production and editing for a long time before the release of DALL·E, Midjourney, or ChatGPT (Miller 2019, 55–132; Offert 2022). The early breakthroughs (2012–2015) in image recognition and machine vision techniques that enabled automated image classification (Dobson 2023; Galloway 2021; Rettberg 2023) were important milestones. These then led to the creation of the first AI-generated images through generative adversarial networks (GANs) since around 2014–2015. From 2020 onwards, so-called diffusion models were combined with transformer-based text encoders (such as CLIP, trained on image-text pairs), enabling "prompting" in natural languages (Coeckelbergh and Gunkel 2025).

In the summer of 2022, specifically, two covers of widely read magazines indicated to an international audience that something about image production was to change drastically: the first of these covers was on *Cosmopolitan* magazine's

Wilde, Lukas R.A., Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, eds. 2026. *Humanities Perspectives on AI: An Interdisciplinary Introduction*. Scandinavian University Press. https://doi.org/10.18261/9788294167234-26-08
Open access | © The author(s) 2026. This work is licensed under CC BY 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/.

"A.I. issue" of June. It presented "The World's First Artificially Intelligent Cover Magazine," an AI-generated "astronaut selfie," attributed jointly to the digital artist Karen X. Cheng and the newly available app DALL·E 2 (Ervik 2023). The second was *The Economist's* June 11 issue, which featured a cover illustration created entirely through the then-new app Midjourney, with the title "AI's New Frontier." After the release of these two commercial apps to a general audience of international users (Midjourney on July 12, 2022, and DALL·E 2 as public beta on July 20, 2022), Stablie Diffusion software became available as an open-weights alternative (on August 22, 2022). It could be modified or even pre-trained in entirely different ways by users (De Rosa Palmini et al. 2024). From then on, generative imagery started to permeate the everyday lives of people around the globe "at a pace that was hardly imaginable just a few years ago" (Somaini 2023, 76). This led to ever more complex uncertainties about which images were created by AI, and which were "made by humans." While new updates for DALL·E, Midjourney, or Stable Diffusion have continued to expand their capabilities and the perceived quality of their outputs ever since, many more generative imagery models, systems, and apps have been released or announced since then. New software programs like Ideogram, Leonardo.AI, Gemini, Blue Willow, Firefly, Grok, FLUX, Janus-Pro-7B, Nano Banana, and Seedream appeared—and keep appearing—only to perhaps be forgotten just as quickly. Every new app and update does, however, come with new promises (Moskatova 2025). AI-based image enhancement and modification have also been integrated into many existing apps and devices, such as the Adobe suite of image processing programs or Google phones.

It is not only static imagery that has seen huge advances through AI. With the announcement of Sora on February 15, 2024, OpenAI also presented the first demo videos of how emerging text-to-video models could expand static AI imagery into moving, animated "AI worlds," while Google unveiled a video competitor named Veo 2 soon after. In February 2026, all of these video models seemed vastly outperformed by ByteDance's Seedance 2.0. All of these systems have their specific possibilities and limitations, and their output often seems to change with each new version or update. The underlying models, too, keep changing. Until April 2025, AI image generation was mostly based on so-called diffusion models. OpenAI implemented an alternative technique with GPT-4o, an *autoregressive* text-to-image model, that allows for even more seamless integration between visual and textual information (OpenAI 2025; Robison 2025). Between the time of writing this text and its publication, the available technologies will certainly have

changed once again—and will continue to do so in the future. Some trends are important to keep in mind, however, despite how advanced the technology might become. It seems likely, however, that some foundational principles of AI image generation will remain the same across apps, iterations, and models for a while.

How AI Images Are Generated

What, then, distinguishes AI imagery from earlier forms of (digital) imagery? Firstly, as with all machine learning technologies, all available models build on vast datasets containing billions of digital images gathered (or, in AI terminology, "scraped") from the internet. In some cases, the contents and sources are disclosed, and we can investigate them (the training data in LAION-5B, instrumental for both Stable Diffusion and possibly Midjourney, for instance, can be accessed through haveibeentrained.com; Buschek and Thorp 2024; Schuhmann et al. 2022;). In contrast, contents and sources often remain carefully guarded secrets in many other cases, such as OpenAI's datasets. This may be quite dangerous, since "every layer of a given training set's architecture" could be seen as "infused with politics," as Kate Crawford and Trevor Paglen (2019) demonstrated with regard to ImageNet (more about that below). In any case, these datasets all contain *multimodal* information—i.e., not only digitalized image data, but pairings with textual information describing or classifying the images (e.g. in the form of captions). These texts can come from vastly different sources, sometimes the product of underpaid "ghost workers" from the Global South (Gray and Suri 2019; Treccani 2018), while in other cases they may be gathered in an unsupervised fashion from commercial websites like Pinterest or Shopify with the help of vast quantities of scraped online sources. Through a frontend app interface, users can then request a specific new image to be generated by submitting their prompts.

These natural language inputs are first broken down into *tokens*, which are then mapped onto *embeddings* in a high-dimensional vector space. This is fundamentally the same process employed by all text chatbots to make "sense" of digital writing (Bajohr 2023). In diffusion models, billions of *weights* (learnable parameters) are then used to calculate probable *steps* from complete Gaussian noise to a *latent* (abstract) representation, which is subsequently decoded into a canvas of *likely pixels* that, ideally, correspond to the initial verbal prompt (Ramesh et al. 2022). This is possible because the model, in its earlier training phase, has added noise to the billions of images in its datasets

and thus gradually "destroyed" the information in them—and "recorded" this process. Thus, it has "learned" the pathways from images to noise by measuring and predicting their iterative decay. The reverse process thus tries to "retrace" this path according to probable next steps of noise reduction/addition (Salvaggio 2023). In autoregressive models, the model instead tries to predict (image) token by (image) token, like in a verbal sentence, the next visual unit from a "vocabulary" of abstract image elements, which are subsequently decoded into pixels.[1] The advantage is that images and texts come from the same multimodal architecture and are more seamlessly integrated. The more specialized diffusion process is far from dead, however. Midjourney and Stable Diffusion continue to use it (with a currently higher visual quality, on average), and only time will tell whether autoregression will replace it in the long run.

Regardless of the models and the underlying datasets, any resulting individual image will always be a statistical approximation: the same prompt will usually generate an infinite number of *similar* results. Since all outputs are not deterministic but merely stochastic, the exact *pathways* from input prompt to output image are opaque to users (in the sense of an "explainable AI"; Zylinska 2020, 75–85). What all of this means is that the *navigation* through the latent space of all possible images through verbal prompts is first and foremost an estimation of outputs that are *likely* to correspond to the original prompt in the same way as the text/image pairings in the original datasets. Even AI imagery that *looks* like photography (or any other representation of a three-dimensional space) is thus actually a "flat" pixel surface to the model: even the AI-generated videos of currently popular consumer-facing apps have *no* simulated mathematical space "behind" them—as in a video game engine that would render graphics onto a 3D simulation.[2] LLMs have been considered "stochastic parrots" (Bender et al. 2021, 610; Bender and Hanna 2025, 21–40), in the sense that they do not *understand* texts but merely produce statistically probable outputs ("number crunching"; M. Mitchell 2020, 250). Whether this still remains the best description of what artificial neural networks are (and

1 Although the exact working mechanisms of OpenAI's GPT-4o are not fully disclosed, it seems to work with a specific mixture of linear and global prediction (see Yan et al. 2025): so-called "hierarchical" Visual Autoregressive Modeling (VAR) employs "next-scale prediction" that works at multiple resolutions simultaneously (starting from a first approximation of the overall image, while adding finer details line by line, see Tian et al. 2024).

2 At least in currently dominant approaches. Techniques like NeRF (Neural Radiance Fields) or 3D Gaussian Splatting do build internal 3D representations (Kerbl et al. 2023). As with all things AI, technological developments are happening quickly and will continue to do so.

are doing), for many image models we could consider this a fitting analogy for represented spaces or spatial objects which "exist" only for human perception. To most generative models, imagery is fundamentally a flat canvas with no depth (Meyer 2024a)—even if "mistaken" physics appear to be easily overcome by better models, just as with the "six finger problem" of the first generation of AI imagery in 2022. The images generated by apps building on such models can nevertheless often be indistinguishable from photographs or video recordings and thus pass any "visual Turing test" in 2026 (they are then often referred to as *deepfakes*; see below). Additionally, AI apps can also imitate any other artistic or material styles, like watercolor paintings, crayon drawings, Pixar animated films, and so on (Thon 2025). The meaning of the term *style* has thus transformed from an art-historical classification to any "repeatable visual pattern extracted from the digitally mobilized images of the past" (Meyer 2023, 100). In other words: just as images that *appear* 3D are merely "flat" pixel canvases to the model, the cultural, historical, or political contexts of images disappear between decoding and encoding as well. They are flattened or homogenized into marketable visual patterns (see also Heuser 2025, 3 on the "flattening effect of AI training data [that] lacks a specific set of historical and social contradictions").

Knowledge about the social and technological production lines "behind" AI imagery is essential for gaining a better understanding of what distinguishes them from earlier forms of (digital and other) images. However, we cannot necessarily *see* these differences in the results any longer. We often have to know, believe, or infer their origins through context. Different academic disciplines then ask different questions. AI imagery has profoundly impacted media culture in general (Somaini 2023), so art history, visual culture studies, photography, film, and media studies are all called upon to contend with these ramifications. A media studies perspective, which is used for this chapter, often considers the cultural and social impacts of AI imagery: what does the availability of these image technologies do within and to society and media culture? Several new and emerging concerns with regard to inequalities, ethics, and politics connected to generative imagery have become especially salient in recent years. While they are too numerous to address in full (Bird et al. 2023)—and they will likely increase with each new technological development—the following three sections present an overview of three different *domains* of concerns surrounding AI imagery: their production (authorship, creative labor, blackboxed datasets), their content (biases, stereotypes, and "slop"), and their distribution (fakes, spectacle, and contamination).

Production Concerns: Authorship, Creative Labor, Blackboxed Datasets

One of the most troubling concerns surrounding the production of generative imagery is the impact it has on our planet: the amount of natural resources and carbon emissions required not only for the training of image models but also for the production of individual AI images. One study from the Hugging Face tech company and Carnegie Mellon University indicated that the creation of a single AI image could consume as much as 25–100% of a smartphone's full battery charge (Luccioni et al. 2023). While there are significant differences between apps, the creation of imagery generally far exceeds the energy consumption of other AI tasks such as text generation or image classification.

Generative imagery has also profoundly complicated the question of who *created* an image and who is *responsible* for it. As with other AI-generated content, authorship is diffusely distributed over countless humans, institutions, and automated processes, and can be conceptualized in very different ways (Bajohr 2024; Barale 2024, 41–57; Coeckelbergh and Gunkel 2025, 74–88). We can and should, however, distinguish *legal* questions from ethical ones. To date, AI-generated imagery does not enjoy copyright protection in any country and can thus be used rather freely, precisely because of this unclear distribution of authorship. Regardless, artists and legal scholars alike have argued that the "training" of the models themselves could be seen as copyright infringement or theft, not merely "data mining," which sounds much more innocent (Dornis and Stober 2024). Criticism of AI imagery first emerged when it began to replace work done by humans who are no longer paid for it. Hashtags such as "#noaiart" and "#artbyhumans" have been trending on Instagram and Twitter/X since 2022, and "No to AI-generated Images" quickly became a widely used slogan for social media protest (Eliaçik 2022). Despite considerable backlash from worried artistic communities (Scribner 2023), media productions have become increasingly accepting of cheaply produced AI imagery—from small journal illustrations to the Marvel/Disney TV show *Secret Invasion* (released in June–July 2023) or Darren Aronofsky's AI-generated show "On This Day… 1776" (first trailer released in January 2026). A different sort of criticism seems to apply where hired (and paid) human artists use AI image generation apps without making their actions explicit. In March 2024, for example, speculation on social media intensified with regard to whether or not a certain comic book artist had used AI imagery for DC's *Batman* series (Fletcher 2024). Additional cover illustrations by other hired artists were removed a month later when they were also apparently (partly) AI-generated

without making this explicit (Cronin 2024). While there is no *immediate* financial damage to other artists here—saving time and work is arguably what digital tools appear to be intended for in the first place—public outrage in the media is based on the assumption that an artist gains credit, recognition, and money for work that silently builds on the works of earlier artists in the datasets. None of these have been given any of "the three Cs:" credit, consent, and compensation (Agence France-Presse 2023).

The most obvious ethical concern about AI imagery is thus founded on this fact: Companies are selling access to their AI for financial gain while monetizing the unpaid creative labor of earlier human artists in the datasets (Treccani 2018). In many cases, once again, they do not even know what (and whose) materials have been used for training. This relationship is especially worrying when the styles of individual artists are appropriated by name. The personal name of the fantasy illustrator Greg Rutkowski (known for his artworks in fantasy games such as *Dungeons & Dragons* or *Magic: The Gathering*), was one of the most widely used Midjourney prompts in 2023 (Heikkilä 2022b). The style of film director Wes Anderson founded a whole TikTok/Facebook genre of AI memes reimagining other franchises (like Harry Potter, Lord of the Rings, or Star Wars) in "Anderson aesthetics" (Diaz 2023). And the "Ghiblification" of imagery, turning existing personal photographs and popular existing memes into "anime-esque" drawings inspired by director Hayao Miyazaki from Studio Ghibli (see figure 8.1), became the most popular use case of OpenAI's GPT-4o model in the spring of 2025 (Di Placido 2025). Artists' inability to sue for copyright infringement—or, at the very least, to be informed and asked about, or able to "opt out" of training datasets—seems even more problematic when AI apps often not only imitate styles, but produce results that look almost like 1:1 copies of the originals. In many cases, they replicate content almost directly from the training sets (Carlini et al. 2023; Somepalli et al. 2022). AI "generation" is then little more than a "filter," or blatant theft (Marcus and Southen 2024).

While all of this might seem more of a concern for "famous artists" (and their creative work and labor) than for "regular citizens," the fact that more and more companies and AI systems are using private data as AI training material (Leffer 2023; Plunkett 2022) should not be overlooked. OpenAI even seems to scrape personal YouTube videos without users' consent, just as long as these are "publicly available" (Koebler 2024). This term sounds as if the company has been granted permission to use the data, when it is "more like the legal equivalent of 'finders, keepers'" (Fried 2024). At the very least, it is a violation of digital privacy

Figure 8.1: The "Ghiblification" of popular existing memes in spring of 2025.

when videos published privately on YouTube are used as AI training material. Even Adobe's Firefly, marketed as more "ethical" than its competitors precisely because it was said to train only on licensed material, was then revealed to use *other* AI-generated images from their competitor Midjourney in their training sets. Midjourney, in turn, scraped data from all over the internet (Metz and Ford 2024). The "chain" of interconnectedness—and the deliberate blurring of legal and ethical responsibility—may thus be confusingly difficult to trace, but it is always to the detriment of human laborers and to the benefit of tech companies.

Finally, an additional concern about the "blackboxed" datasets that AI models draw on (Crawford and Paglen 2019; Søraa and Ask 2026) stems from the fact that these archives may contain material that is questionable or offensive for reasons other than copyright or privacy. Stanford Researchers found, for example, that the LAION-5B dataset contained countless images depicting rape, pornography, and racist stereotypes (Birhane et al. 2021; Goldman 2023), and over a thousand images of pedophilia. It was republished as a cleaned "RE-LAION-5B" afterwards. Apparently, no one deemed it necessary to clean the data before the respective apps became widely used, in part because the sets are often too large for humans to check manually. If one were to look at each image in the LAION-5B dataset for only one second, it would take approximately 781 years of regular working weeks to open each picture only once (Buschek and Thorp 2024). When researcher Eryk Salvaggio investigated the same dataset to find out which images might have been used for a seemingly

harmless prompt such as "archive," he found that it draws equally on the TV show *Riverdale*, photographs of Holocaust survivors, and of Nazi Wehrmacht soldiers (Salvaggio 2023). While we might not necessarily *see* this unfortunate "mix" in the results, it is at least disconcerting to know that AI image creation might build on—and partly reproduce—materials that we would otherwise carefully safeguard and contextualize. AI imagery, once again, "flattens" not only represented spaces into (probable) pixel surfaces, but also cultural contexts into visual patterns from which profit can be made.

Representation Concerns: Biases, Stereotypes, and "Slop"

If we look at AI-generated images themselves, there are numerous concerns about how they generally represent the verbal prompts given by users. Irrespective of the specific apps and models, generative imagery is likely to reproduce or amplify biases, stereotypes, and clichés surrounding race, gender, dis/ability, and other social categories. Comparative analyses across AI apps show that different versions of DALL·E and Stable Diffusion tended to couple social roles (professions like "a CEO" or "an artist") and basic trait adjectives ("compassionate" versus "ambitious") with social attributes like "man/male," "woman/female," or "Latino" versus "White" (Heikkilä 2023). An "attractive person" will thus be likely to be represented by a young, white face, while "a thug" was, more often than not, represented as a person of color if not specifically prompted otherwise. The prompt "Native American" often has a high probability of generating stereotypical pictures of people with outrageous headdresses. Although companies have improved on these issues somewhat since 2022, a "productive person" still looks rather homogenous to Midjourney V7 in the spring of 2026 (see figure 8.2). This is, on the one hand, because the datasets they draw on are not free of these biases to begin with (the "mirror of society argument," Coeckelbergh and Gunkel 2025, 27). It is also worth considering the fact that most models today are trained predominantly on US-centric data, which means that they will mostly reflect American or Western representations of language, culture and the world. For models trained on LAION-5B, for example, English sources are more common than all the other hundred or so languages combined.

What is even more worrying is that generative imagery tends to *amplify* these biases and clichés, even beyond rates found in the training data or the real world (Bianchi et al. 2022; Garg et al. 2017; Miltner 2024). Controversies that

Figure 8.2: Prompt: "A productive person," created with Midjourney V7.

attracted significant attention included AI-generated (Lensa) avatars based on user photographs, which tended to sexualize almost all pictures of women, while male users were represented as scientists, explorers, and inventors (Heikkilä 2022a). In all these cases, it is sometimes impossible to reconstruct exactly where these biases come from. If we ask Midjourney V7 for "An authentic picture of a typical day in Trondheim, everyday life," for example, the model presents almost photorealist renderings of Bakklandet and the Nidelva river alongside cringy illustrations of stereotypical "Norwegianness" (see figure 8.3). Where do these "vibes" come from? It is extremely hard to say where this "knowledge" is stored, as it emerges from a complex interaction between many elements combined within one system, such as a) the *training datasets* (like LAION-5B) with their existing image/text pairings; b) a pre-trained *vision-language encoder* (such as CLIP, a text-image matching model) that assigns a series of numerical vectors to linguistic prompts in order to "understand" them within a high-dimensional latent space; and c) the *image generation* itself through different

image models (which can be exchanged through "marketplaces" like CivitAI for Stable Diffusion; De Rosa Palmini et al. 2024), trained and "defaulted" differently even with recourse to the same dataset. The architecture and "pipeline" for video models like Seedance 2.0 is even vastly more complex (Theoretical Media 2026).[3] The biggest concern here is that biases in all of these components might remain invisible to users as they are reproduced, irrespective of whether prompts explicitly mention identity or try to avoid such language.

Figure 8.3: Prompt: "An authentic picture of a typical day in Trondheim, everyday life," created with Midjourney V7.

While social stereotypes certainly existed long before generative imagery—and thus before their algorithmic reproduction and perpetuation—it is important to consider how AI imagery also seems to create *new* distortions that might not

3 No peer-reviewed source was available at the time when this article went into print.

reflect any preexisting stereotypes. The specifics and examples, once again, will change with each new app and model, but the foundational principle remains the same. There will always be patterns in how certain prompts are connected to specific types of images, and the results are not always foreseeable. Fabian Offert (2023), for example, noted in the fall of 2022 for DALL·E that the term "fascism" was, by default, always connected to a specific media aesthetic (that of early Kodachrome slide photography), simply because most historical imagery (to which the text description "fascism" is mostly attributed) was produced with that specific image technology. Hence, even when users wanted contemporary or futuristic representations of "fascism," the results seemed to be confined to a distant past (and thus socially "overcome"), because the images always looked *old*. The conflation of one specific image technology with fascism is not a bias that existed (linguistically or socially) before; it is purely the result of our image datasets and their tagging. Similar dynamics can be found not only on the level of style but also on the level of content, so that the prompt "toys in Iraq" predominantly produced (in a specific Stable Diffusion version) soldiers and military equipment as action figures and stuffed animals—probably because a great number of photographs tagged "Iraq" nowadays portray war scenes and soldiers (Growcoot 2023b).

Figure 8.4: Google's unintentionally "over-diversified" Gemini model of 2024.

Such technical inclinations towards (preexisting or new, "re-mediational") biases and stereotypes are hard to overcome, especially with regard to diversity.

Companies endeavored to develop various mitigation strategies that, in fact, created new problems. OpenAI, for instance, decided to "silently" (and randomly) include "diversity tags" like "woman" or "black" into users' prompts before images were created, which "did not fix the model but the user" by "literally putting words in the user's mouth" (Offert and Phan 2022, 2). Occasionally, this created especially strange results when, in an image of "Laocoön and his sons," only the *father* of the three figures appears randomly as a Black person. Google's similar (also system-level, rather than model-level) intervention to make its AI system Gemini more diverse produced major backlash in this regard. The company quickly found itself in the middle of a "culture war" in early 2024 after the app automatically diversified *all* images of people so that requests for a "pope," the "American Founding Fathers," and "Viking warriors" or "Nazi soldiers" *all* predominantly showed people of color, women, or various minority groups (see figure 8.4). More than any conscious attempt to make AI imagery "woke," this illustrates the limitations of the technology, *including* the difficulties of overcoming them (Kleinman 2024). However, this also shows the intricate issues of representation since there exists no "one size fits all" way to present images—the generative aspect requires cultural awareness of when and how to present people from diverse backgrounds (see also Søraa and Ask 2026).

Aside from all issues of content, the style of AI-generated imagery could also be investigated critically. This is obviously dependent on the respective style prompts again (apart from, once again, app and model). Nevertheless, there is, and probably will continue to be, a general tendency toward images and styles that look popular and pleasing according to *average human observers*. Stable Diffusion, for example, draws only on an algorithmically created *subset* of LAION-5B called "LAION-Aesthetics" that contains images rated as visually appealing. In this case we can actually trace the datasets and where they come from: "a handful of [very active] users" whose "aesthetic preferences dominate the dataset" (Buschek and Thorp 2024; see also Somaini 2023, 103–104). Roland Meyer (2025b) referred to the resulting defaults as (social media) "platform-ready" aesthetics optimized for mobile screens. Such "platform realism" does not try to reflect reality but strives to replicate the appearances of popular image streams online. Platform realism, he summarized, is not an aesthetic of authenticity, but of plausibility. "Tellingly, plausibility and applause have the exact same origin: plausibility is what deserves an audience's applause and approval. In this sense, AI images are optimized for plausibility; they aim for the broadest possible approval" (Meyer 2025b, 17). In other words, the most popular default style

will always lean towards some sort of "kitsch" (Zylinska 2020, 80). Aesthetic judgement of AI imagery as "slop" or "trash" (Thon and Wilde 2025) can also be politically motivated. In his widely discussed article of spring 2025, art historian Gareth Watkins argued for *structural reasons* why AI imagery was "inherently fascist," since it always draws on the past, hurts creative workers, and harms the environment. Whether one agrees or not, it seems undeniable that, at least to date, AI imagery is particularly popular with right-wing and right-extremist parties and politicians around the globe, ranging from Trump to Britain First and the German AfD party (Watkins 2025). Whether this will change again and AI imagery can fulfill its promise of a "democratization" of artistic expression remains to be seen.

Distribution Concerns: Deepfakes, Spectacle, Contamination

After AI images have been created, they can be used in various ways or integrated into other media forms such as films, TV shows, comic books, or video games (Wilde 2025). They have to find their "place" within their surrounding media ecologies, and countless decisions will have to be made about what defines "acceptable" use. Can AI images that look like photographs be accepted in digital photography contests, for instance? Should "AI artists" be granted space in comic book convention art galleries where fans sell their homemade art? Can and should scholars use AI imagery to illustrate lectures or academic publication covers? Many of these cultural and social *protocols* (conventionalized practices of typical production, distribution, and reception; Gitelman 2006, 5; Philipsen 2025, Hoel and Wilde 2026) are not in place as yet, and the use of "acceptable" AI imagery remains to be negotiated. It is certainly true to say that social media streams are the most common "site" where we currently encounter AI imagery, where people proudly present their creations and discuss prompts and models. We usually *know* that an image is AI-generated not because we can (necessarily) see it, but because the context reveals it (e.g. Facebook groups like "AI Art Universe") or because it is verbally disclosed in the captions and comments. This points to the fact that questions about (responsible, controversial, or established) use of AI imagery will always depend on contextual information and linguistic means just as much as on the images themselves. Irrespective of all these important differences, however, it is possible to identify three general ways to distribute AI imagery—and once again, each of these gives rise to different concerns.

First, the fact that an image *was* AI-generated can be concealed. This is usually discussed in the context of *deepfakes*, where something looks like an actual photograph or a video while it is actually AI-produced (Broinowski 2022; Juefei-Xu et al. 2022; Yadlin-Segal and Oppenheim 2021). Widely discussed political examples of AI deepfakes include apparent photos from the Israel-Hamas war in Gaza showing dead babies and ravaged streets (Associated Press 2023); deepfake videos appearing to record Ukraine's president, Volodymyr Zelenskyy, talking about surrendering to Russia, as well as one of the Russian President, Vladimir Putin, declaring peace (Wakefield 2022). AI fakes or frauds have thus become a major concern since 2022, even though, in principle, there is nothing new about the fact that images (and digital images in particular) can, of course, be manipulated (Carlson 2009; W.J.T. Mitchell 1992; Sharma and Sharma 2017). Photoshop and realistic looking CGI special effects in cinema have been around for decades, after all (Ferwerda 2003; Giralt 2017). A new domain of image manipulation—in terms of both quantity and quality—comprises one of the most prominent use cases of AI model training today; the creation of nonconsensual deepfake pornography of real people (Maiberg 2025).

Figure 8.5: Perhaps the most iconic picture of 2023, the world's first "unintentional deepfake".

It has been true for a long time that "photorealism is not photography but rather mimics a style" (Hausken 2024, 3; see also Meyer 2025a). However, the accessibility of AI images—which require no specialized skills, knowledge, or expensive equipment—has not only increased their number. Maybe for the first time, we are also seeing "unintentional fakes" like the notorious "Balenciaga Pope" of March 2023 (Meyer 2024b, and figure 8.5). This Midjourney fabrication of Pope Francis in a stylish white puffer jacket turned out to be one of the most iconic pictures of the year, precisely because it was *not* generated to deceive anyone. Prompted and published by a construction worker from Chicago who apparently just wanted to "see the pope in a funny jacket," the result was liked and shared by hundreds of thousands of (often quite media literate) global users who *did* mistake the picture for a real one—maybe precisely because it did not seem to have any political content or message (Golby 2023). Such cases have fueled the need for a new form of media literacy in order to "spot" AI generation, for instance, by identifying typical mistakes or errors such as the wrong number of fingers. The web has since been brimming with "How to Spot an AI Image" tutorials as well as respective "gamifications" that challenge users to test their AI literacy by guessing which images are "real" and which are not. Increasingly, however, we can only know that from context, as in the case of an apparent explosion at the US Pentagon in May 2023 which even caused a brief shock within the stock markets. The images, shared only through private social media accounts, were once again AI-generated (Haddad 2023).

What all of these cases have in common is the fact that they do *not* disclose their AI-generated nature. In media studies, they could thus be discussed as examples of an apparent *immediacy*. They have the appearance of "a transparent interface […], one that erases itself, so that the user would no longer be aware of confronting a medium" (Bolter and Grusin 1999, 318). While immediacy is mostly connected to photorealistic styles (looking like photochemical traces of light on sensors), it likewise applies to seemingly hand-drawn pictures or any other material styles hiding the fact that they were generated through algorithmic "remediations" of earlier images (this, perhaps, also explains the controversies mentioned above surrounding AI-generated comic book "fraud images"). The opposite of *immediacy* can be called *hypermediacy*, a deliberate foregrounding of their mediality, their AI generation. Hypermediacy "strives to make the viewer acknowledge the medium as a medium and indeed delight in that acknowledgment" (Bolter and Grusin 1999, 335; see also Wilde 2023). Here, the communicative intent is to celebrate the fact that the images were AI-generated because they deliberately create spectacular depictions that are

intentionally at odds with reality. One of OpenAI's first promotional DALL·E 2 images featured an astronaut riding a horse in outer space to highlight the artificiality of the motif—the impossibility that it could be an actual photograph (Heaven 2022). Since the spectacle of hypermediacy builds upon the marvel of *how* these images were created, special prominence is attached to the verbal prompts and the specific models used (which are both, of course, hidden in the case of deepfakes or apparent immediacy), either because the prompts and models are discussed and compared by a community like "recipes", or because they are carefully kept a secret. Others can then speculate about the displayed skill in *prompt engineering*. Another arena of AI hypermediacy can be found in tech demos of new and emerging models, like OpenAI's announcement of their upcoming text-to-image model Sora in February 2024. These demos showcase new AI abilities with the purpose, once again, not of deceiving, but of astonishing; to make transparent the fact that the results are the products of new machine learning technologies (Moskatova 2025).

Immediacy and hypermediacy may be closely interlinked (Meyer 2024b) in such a way that the Balenciaga Pope was immediately incorporated into waves of memes that showed Pope Francis in highly unlikely scenarios; in nightclubs, or as a beach surfer. *Memes*—series of similar images following an established pattern or template (Shifman 2013)—are currently the most typical genre of hypermediated AI images. They also reveal how immediacy and hypermediacy follow the same logic within contemporary media cultures, because apparent AI deepfakes now increasingly instigate *wild forensics*, collective speculations about whether or not a shared picture was manipulated, the search for revealing details and traces of manipulation (Meyer 2024b). What connects these wild forensics (surrounding potential AI deepfakes) to AI memes, then, is that both are responses to earlier pictures, embedded in a culture of mash-ups, remixes, and variations of pictorial prototypes. Such speculations are not necessarily aimed at conclusions, but at further reposts and comments within a "reaction economy" (Davies 2023). Whether they try to *hide* or *highlight* their mediality, one way or the other, their AI generation remains crucially important.

However, one should not overlook the fact that there is a third use case for AI-generated imagery that is, in fact, *indifferent* to how it has been created. This is when an image, irrespective of the style it uses, serves as a mere *illustration* of the accompanying text surrounding it. Hausken (2024, 7) calls this a different "genre" of images, following the tradition of "illustrative photography" rather than "news photography," "reportage photography," or "portrait photography" (see also Kędra 2016 for further discussions). Such illustrative

photography that often accompanies news articles *could* always employ stock material (or, even earlier, drawn illustrations) without making any claims that the visible scene took place—a mere *symbol image*. Here, AI imagery simply serves to exemplify *typical* objects, people, or scenes. For that reason, it does not necessarily make any difference whether the image was drawn, taken from stock photography, or AI-generated. For photographic images in journalistic contexts, it is established and important to clearly mark such material as "symbol images" or "illustrations," in order to make apparent that such images simply serve to exemplify, not to record. Once again, the communicative meaning of AI imagery does not derive from individual pictures alone, but from the overall communicative context that we can learn from verbal captions or written explanations, for instance.

Figure 8.6: Political propaganda by Elon Musk to influence the 2024 US elections.

In this respect, maybe not so much has changed concerning image communication. When right-wing extremists use racist imagery of "foreigners" accompanying hateful posts, what is shown are merely (racist) "types" and the desired attitudes towards them (fear, hatred, etc.). In such contexts, AI imagery

is little different from earlier racist drawings employed in exactly the same way (Lemmes 2025). Similarly, immediately before the 2024 US elections, Elon Musk published an AI-generated image of Kamala Harris wearing a communist uniform (see figure 8.6) with the caption, "Kamala vows to be a communist dictator on day one. Can you believe she wears that outfit!?" (Musk 2024). Not claiming that the picture was "real," even mocking the very idea that anyone *could* think that it was, it communicated a specific attitude or a set of affects (perhaps disdain and fear) against Trump's competitor. Not even looking especially (photo)realistic, a mocking caricature drawing serves the same purpose. AI material is just more easily available than other imagery for political/propaganda memes. The possibility of using AI imagery to show only types, not particular individuals, also offers new opportunities for responsible and ethical uses. In the spring of 2023, for example, Amnesty International employed AI imagery for a report about police violence against demonstrators in Colombia. Using AI-generated, generalized pictures instead of photojournalistic material, the intent was clearly to protect the privacy and the safety of individuals. The campaign nevertheless received a great deal of critical backlash, and Amnesty regretted it later (Hausken 2024) because many people were concerned that the pictures might either be mistaken for photographic documentation (and thus as fraudulent) or serve only as dramatic clickbait (Herrie and Philipsen 2026). If nothing else, this highlights again the need to make the generation of AI imagery and its intended reference (types versus individuals) as clear as possible, because that is not necessarily apparent. Naturally, the more such contextual genres are adopted across media contexts, the more these genres will change as we grow accustomed to AI imagery.

Conclusion

While analyses of *individual* AI-generated images and their uses reveal important differences as to whether they are used as immediacy fakes, as hypermediacy spectacles, or as media-indifferent illustrations, we should also consider the fact that the increasing number of AI images flooding social media now points to a different concern altogether beyond the individual image: that of contamination, or "pollution of the information ecosystem" (Bender 2024). If more and more images are AI-generated, not only do AI datasets begin to draw upon themselves (rather than upon earlier image types, Alemohammad et al. 2023), but our visual cultural memory in general may also become contaminated. From the spring of 2024 onwards, many debates focused on the fact that

more and more image queries on widely used search engines such as Google started spitting out AI-generated results, accompanied by or on top of actual ones (Balkowitsch 2024). Examples included seemingly harmless ones, such as searches for the Hawaiian singer/ukulele player Israel Kamakawiwoʻole bringing up bizarre Midjourney creations (Larsen 2023); images of a "Baby Peacock," which do not depict any actual specimen but rather, increasingly, a cute 'Pokémon' (LaMagdeleine 2023). This seems even more troubling with the contamination of memory when searches for historical events such as the Tiananmen Square massacre led to AI-generated selfies (see figure 8.7) instead of the iconic documentary photo taken by Jeff Widener in 1989 (Growcoot 2023a).

Figure 8.7: Midjourney reimagines the Tiananmen Square massacre as a selfie—and Google is flooded with the results.

A related problem lies in the fact that many social media algorithms favor accounts producing and uploading as much content as possible, as quickly as possible. This has resulted in many of the most-subscribed-to YouTube channels in the summer of 2025 featuring nearly identical variations of the same AI-generated videos (Bumas 2025). They have huge algorithmic advantages over images created by other means: if anything, the latter takes more time!

Mass-produced AI content thus adds to the "pollution" of the internet through "slop." The development of AI-generated imagery thus not only demands careful reflection, analysis, and criticism of individual images and their uses, but also of their impact on our larger media ecology, our visual cultural memory, and our overall social, cultural, and political communication. This will become even more troublesome when the difference between AI-generated imagery and other types becomes entirely blurred—or only a matter *of degree*—as AI image enhancement will be embedded in every function of digital apps such as Photoshop. At the same time, AI image software like Nano Banana or Seedream consists of image editing apps, modifying earlier photographs selectively here and there, so that the distinction can no longer be drawn with regard to the overall picture. Perhaps it will not always be clear—or intentional—to users themselves, generating more and more *accidental* distortions of media ecology and visual culture (Laraki 2024).

The three concerns described in this chapter about production, representation, and distribution are, of course, closely connected and should be considered together, not in isolation. They reveal that we always have to assess AI imagery in contexts: The first are verbal contexts in the form of captions or textual image descriptions that highlight or conceal whether something was AI-generated; then, the contexts of distribution and circulation, when an individual image is suddenly transformed into a meme shared by thousands. We have media and genre contexts to consider, when we find AI imagery in commercial advertising campaigns or communications by political parties. Finally, it is perhaps most important to pay attention to the contexts of *earlier images* simulated or remediated by AI, their specific functions, values, and associated conventions. At the same time, new areas of AI imagery will continuously emerge that might develop their own social, cultural, and political rules, protocols, or controversies surrounding their use.

Works Cited

Agence France-Presse. 2023. "Artists Fight for Consent, Credit or Compensation in AI Program Court Battle." *News24,* March 28, 2023. https://www.news24.com/life/arts-and-entertainment/arts/artists-fight-for-consent-credit-or-compensation-in-ai-program-court-battle-20230328#google_vignette.

Alemohammad, Sina, Josue Casco-Rodriguez, Lorenzo Luzi, et al. 2023. "Self-Consuming Generative Models Go MAD." Preprint, arXiv, July 4, 2023. https://doi.org/10.48550/arXiv.2307.01850.

Associated Press. 2023. "Fake Babies, Real Horror: Deepfakes from the Gaza War Increase Fears about AI's power to mislead." *Spectrum News*, November 28, 2023. https://ny1.com/nyc/all-boroughs/ap-top-news/2023/11/28/fake-babies-real-horror-deepfakes-from-the-gaza-war-increase-fears-about-ais-power-to-mislead.

Bajohr, Hannes. 2023. "Dumb Meaning: Machine Learning and Artificial Semantics." *Generative Imagery: Towards a 'New Paradigm' of Machine Learning-Based Image Production*, Special-Themed Issue of *IMAGE: The Interdisciplinary Journal of Image Sciences* 37 (1): 58–70.

Bajohr, Hannes. 2024. "Writing at a Distance: Notes on Authorship and Artificial Intelligence." *German Studies Review* 47 (2): 315–337. https://doi.org/10.1353/gsr.2024.a927862.

Balkowitsch, Shane. 2024. "AI is Corrupting the Internet as We Know It." *PetaPixel,* April 25, 2024. https://petapixel.com/2024/04/25/ai-is-corrupting-the-internet-as-we-know-it/?fbclid=IwY2x-jawGW42ZleHRuA2FlbQIxMAABHRB8oA6hIfmyBny5Y6te1nEHSOJD0zh94NmSQtVHujl-tAC1t87rEF0ddtA_aem_TAN8Ok-W1_qu4wOsoW0-SQ.

Barale, Alice. 2024. *The Art of Artificial Intelligence: Philosophical Keywords*. Cambridge: Cambridge Scholars Publishing.

Bender, Emily. 2024. "Imagine if the CEO of BP bragged that one in one thousand molecules flowing into the Gulf of Mexico came from the Deep Horizon oil rig.…" Posted on *X*, February 10, 2024. https://x.com/emilymbender/status/1756342308960895281.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜" *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*: 610–623. https://doi.org/10.1145/3442188.3445922.

Bender, Emily M. and Alex Hanna. 2025. *The AI Con: How to Fight Big Tech's Hype and Create the Future We Want.* New York: HarperCollins.

Bianchi, Federico, Pratyusha Kalluri, Esin Durmus, et al. 2022. "Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale." Preprint, arXiv, November 7, 2022. https://doi.org/10.1145/3593013.3594095.

Bird, Charlotte, Eddie L. Ungless, and Atoosa Kasirzadeh. 2023. "Typology of Risks of Generative Text-to-Image Models." Preprint, arXiv, July 8, 2023. https://doi.org/10.48550/arXiv.2307.05543.

Birhane, Abeba, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. "Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes." Preprint, arXiv, October 5, 2021. https://doi.org/10.48550/arXiv.2110.01963.

Bolter, Jay D. and Richard Grusin. 1999. *Remediation: Understanding New Media.* Cambridge, MA: The MIT Press.

Broinowski, Anna. 2022. "Deepfake Nightmares, Synthetic Dreams: A Review of Dystopian and Utopian Discourses Around Deepfakes, and Why the Collapse of Reality May Not Be

Imminent—Yet." *Journal of Asia-Pacific Pop Culture* 7 (1): 109–139. https://doi.org/10.5325/jasiapacipopcult.7.1.0109.

Bumas, Adam. 2025. "Youtube Is Not Looking So Hot Right Now." *Garbage Day*, July 16, 2025. https://www.garbageday.email/p/youtube-s-second-biggest-channel-is-full-of-ai-babies-84577fabec5202d9/.

Buschek, Christo and Jer Thorp. 2024. "Models All the Way Down." *Knowing Machines*, April 9, 2024. https://knowingmachines.org/models-all-the-way.

Carlini, Nicholas, Jamie Hayes, Milad Nasr, et al. 2023. "Extracting Training Data from Diffusion Models." Preprint, arXiv, January 30, 2023. https://doi.org/10.48550/arXiv.2301.13188.

Carlson, Matt. 2009. "THE REALITY OF A FAKE IMAGE: News Norms, Photojournalistic Craft, and Brian Walski's Fabricated Photograph." *Journalism Practice* 3 (2): 125–139. https://doi.org/10.1080/17512780802681140.

Coeckelbergh, Mark and David J. Gunkel. 2025. *Communicative AI: A Critical Introduction to Large Language* Models. Cambridge: Polity.

Crawford, Kate and Trevor Paglen. 2019. "Excavating AI: The Politics of Training Sets for Machine Learning." *Excavating.ai*, September 19, 2019. https://excavating.ai.

Cronin, Brian. 2024. "DC Pulls Controversial Content Suspected of Being Produced by Generative A.I." *Comic Book Resources,* April 22, 2024. https://www.cbr.com/dc-pulls-covers-generative-ai-art-suspected/.

Davies, William. 2023. "The Reaction Economy." *The London Review of Books* 45 (5), March, 2023. https://www.lrb.co.uk/the-paper/v45/n05/william-davies/the-reaction-economy.

De Rosa Palmini, Maria-Teresa, Laura Wagner, and Eva Cetinic. 2024. "Civiverse: A Dataset for Analyzing User Engagement with Open-Source Text-to-Image Models." Preprint, arXiv, August 10, 2024. https://doi.org/10.48550/arXiv.2408.15261.

Di Placido, Dani. 2025. "The AI-Generated Studio Ghibli Trend, Explained." *Forbes*, March 27, 2025. https://www.forbes.com/sites/danidiplacido/2025/03/27/the-ai-generated-studio-ghibli-trend-explained.

Diaz, Jesus. 2023. "Why Wes Anderson Refuses to Watch his TikTok Memes." *FastCompany*, June 26, 2023. https://www.fastcompany.com/90901908/wes-anderson-ai-remakes-tiktok.

Dobson, James E. 2023. *The Birth of Computer Vision*. Minneapolis: University of Minnesota Press.

Dornis, Tim W. and Sebastian Stober. 2024. "Copyright Law and Generative AI Training—Technological and Legal Foundations." *NOMOS Open Access Books*, August 29, 2024. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4946214.

Eliaçik, Eray. 2022. "Does ArtStation Become PromptStation?" *DataConomy*, December 20, 2022. https://dataconomy.com/2022/12/no-to-AI-generated-images-artstation.

Ervik, Andreas. 2023. "Generative AI and the Collective Imaginary: The Technology-Guided Social Imagination in AI-Imagenesis." *Generative Imagery: Towards a 'New Paradigm'*

of Machine Learning-Based Image Production, Special-Themed issue of *IMAGE: The Interdisciplinary Journal of Image Sciences* 37 (1): 42–57.

Ferwerda, James A. 2003. "Three Varieties of Realism in Computer Graphics." *IS&T/SPIE Electronic Imaging*. https://cin.ufpe.br/~in1123/material/vor_hvei03_v20.pdf.

Fletcher, Harry. 2024. "DC Comics 'Looking into' Claims that Batman Artist Used AI for Comic Book Art." *indy100*, March 14, 2024. https://www.indy100.com/showbiz/batman-andrea-sorrentino-ai-art-2667509825.

Fried, Ina. 2024. "For AI Firms, anything 'Public' is Fair Game." *Axios*, April 5, 2024. https://www.axios.com/2024/04/05/open-ai-training-data-public-available-meaning?fbclid=IwY2xjawGW5CRleHRuA2FlbQIxMAABHaxSESctDCEa4sDp8YLZZ7Z5h6pH2_lPRLl-3g7WqilNfhGGPIaEFY5Ux2g_aem_yR01LajOGe9KzLtMyeXWPQ.

Galloway, Alexander. 2021. *Uncomputable: Play and Politics in the Long Digital Age.* London: Verso.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. 2017. "Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes." Preprint *PNAS* 115 (16) E3635–E3644, November 22, 2017. https://doi.org/10.1073/pnas.1720347115.

Giralt, Gabriel F. 2017. "The Interchangeability of VFX and Live Action and Its Implications for Realism." *Journal of Film and Video* 69 (1): 3–17. https://doi.org/10.5406/jfilmvideo.69.1.0003.

Gitelman, Lisa. 2006. "Introduction: Media as Historical Subjects." In *Always Already New. Media History and the Data of Culture*, 1–22. Cambridge, MA: The MIT Press.

Golby, Joel. 2023. "I thought I was Immune to Being Fooled Online: Then I Saw the Pope in a Coat." *The Guardian*, March 27, 2023. https://www.theguardian.com/commentisfree/2023/mar/27/pope-coat-ai-image-baby-boomers.

Goldman, Sharon. 2023. "A Free AI Image Dataset, Removed for Child Sex Abuse Images, has Come under Fire before." *Venture Beat*, December 29, 2023. https://venturebeat.com/ai/a-free-ai-image-dataset-removed-for-child-sex-abuse-images-has-come-under-fire-before.

Gray, Mary L. and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston: Houghton Mifflin Harcourt.

Growcoot, Matt. 2023a. "AI Image of Tiananmen Square's Tank Man Rises to the Top of Google Search." *PetaPixel,* September 27, 2023. https://petapixel.com/2023/09/27/ai-image-of-tiananmen-squares-tank-man-rises-to-the-top-of-google-search.

Growcoot, Matt. 2023b. "Which AI Image Generator is The Most Biased?" *PetaPixel*, November 3, 2023. https://petapixel.com/2023/11/03/which-ai-image-generator-is-the-most-biased.

Haddad, Mohammed. 2023. "Fake Pentagon Explosion Photo Goes Viral: How to Spot an AI Image." *Al Jazeera*, May 23, 2023. https://www.aljazeera.com/news/2023/5/23/fake-pentagon-explosion-photo-goes-viral-how-to-spot-an-ai-image.

Hausken, Liv. 2024. "Photorealism versus Photography: AI-generated Depiction in the Age of Visual Disinformation." *Journal of Aesthetics & Culture* 16 (1). https://doi.org/10.1080/20004214.2024.2340787.

Heaven, Will Douglas. 2022. "This Horse-Riding Astronaut is a Milestone on AI's Long Road towards Understanding." *MIT Technology Review*, April 6, 2022. https://www.technologyreview.com/2022/04/06/1049061/dalle-openai-gpt3-ai-agi-multimodal-image-generation.

Heikkilä, Melissa. 2022a. "The Viral AI Avatar App Lensa Undressed me—Without my Consent." *MIT Technology Review*, December 12, 2022. https://www.technologyreview.com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent.

Heikkilä, Melissa. 2022b. "This Artist is Dominating AI-Generated Art: And he's not Happy about it." *MIT Technology Review*, September 16, 2022. https://www.technologyreview.com/2022/09/16/1059598/this-artist-is-dominating-ai-generated-art-and-hes-not-happy-about-it.

Heikkilä, Melissa. 2023. "These New Tools Let You See for Yourself how Biased AI Image Models are." *MIT Technology Review*, March 22, 2023. https://www.technologyreview.com/2023/03/22/1070167/these-news-tool-let-you-see-for-yourself-how-biased-ai-image-models-are/.

Herrie, Maja Bak and Lotte Philipsen. 2026. "'No Children Were Harmed': Aesthetic Infrastructures and Ethical Implications of AI-generated Portraits." *Visual Studies* 2026: 1–12. https://doi.org/10.1080/1472586X.2025.2606201.

Heuser, Ryan. 2025. "Generative Aesthetics: On Formal Stuckness in AI Verse." *Journal of Cultural Analytics* 10 (3). https://doi.org/10.22148/001c.144825.

Hoel, Aurora and Lukas R.A. Wilde. 2026. "AI as Media." In *Humanities Perspectives on AI: An Interdisciplinary Introduction*, edited by Lukas R.A. Wilde, Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, 145–172. Oslo: Scandinavian University Press.

Juefei-Xu, Felix, Run Wang, Yihao Huang, Qing Guo, Lei Ma, and Yang Liu. 2022. "DeepFakes: Survey, Battleground, and Horizon." *International Journal of Computer Vision* 130: 1678–1734. https://doi.org/10.1007/s11263-022-01606-8.

Kędra, Joanna. 2016. "Enhancing Visual Literacy Through Interpretation of Photo-Genres: Toward a Genre Typology of Journalistic Photographs." *Journal of Media Practice* 17 (1): 28–47. http://dx.doi.org/10.1080/14682753.2016.1159451.

Kerbl, Bernhard, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. "3D Gaussian Splatting for Real-Time Radiance Field Rendering." Preprint, arXiv, August 8, 2023. https://doi.org/10.48550/arXiv.2308.04079.

Kleinman, Zoe. 2024. "Why Google's 'Woke' AI Problem won't be an Easy Fix." *BBC*, February 28, 2024. https://www.bbc.com/news/technology-68412620.

Koebler, Jason. 2024. "The Backlash Against AI Scraping Is Real and Measurable." *404media*, July 23, 2024. https://www.404media.co/the-backlash-against-ai-scraping-is-real-and-measurable/.

LaMagdeleine, Izz Scott. 2023. "Fact Check: Is This a Real Pic of a Baby Peacock?" *Yahoo News*, August 15, 2023. https://news.yahoo.com/fact-check-real-pic-baby-184500737.html.

Laraki, Elizabeth. 2024. "I'm talking at a conference later this year (on UX+AI)…." Posted on *X*, October 15, 2024. https://x.com/elizlaraki/status/1846252781851890026.

Larsen, Luke. 2023. "Fake AI Images are Showing up in Google Search—and it's a Problem." *Digitaltrends*, September 28, 2023. https://www.digitaltrends.com/computing/fake-ai-images-showing-in-google-search.

Leffer, Lauren. 2023. "Your Personal Information Is Probably Being Used to Train Generative AI Models." *Scientific American*, October 19, 2023. https://www.scientificamerican.com/article/your-personal-information-is-probably-being-used-to-train-generative-ai-models.

Lemmes, Marcel. 2025. "Beyond the Hyperreal: Digitale Bilder und KI—Eine Herausforderung für die Bildsemiotik?" In *Bilder im Aufbruch: Herausforderungen der Bildwissenschaft*, edited by Marcel Lemmes, Stephan Packard, and Klaus Sachs-Hombach, 319–351. Cologne: Herbert von Halem.

Luccioni, Alexandra Sasha, Yacine Jernite, and Emma Strubell. 2023. "Power Hungry Processing: Watts Driving the Cost of AI Deployment?" Preprint, arXiv, November 28, 2023. https://doi.org/10.1145/3630106.3658542.

Maiberg, Emanuel. 2025. "Hugging Face Is Hosting 5,000 Nonconsensual AI Models of Real People." *404 Media*, July 15, 2025. https://www.404media.co/hugging-face-is-hosting-5-000-nonconsensual-ai-models-of-real-people/.

Marcus, Gary and Reid Southen. 2024. "Generative AI Has a Visual Plagiarism Problem: Experiments with Midjourney and DALL-E 3 Show a Copyright Minefield." *IEEE Spectrum*, January 6, 2024. https://spectrum.ieee.org/midjourney-copyright.

Metz, Rachel and Brody Ford. 2024. "Adobe's 'Ethical' Firefly AI Was Trained on Midjourney Images." *Bloomberg*, April 12, 2024. https://www.bloomberg.com/news/articles/2024-04-12/adobe-s-ai-firefly-used-ai-generated-images-from-rivals-for-training?embedded-checkout=true.

Meyer, Roland. 2023. "The New Value of the Archive: AI Image Generation and the Visual Economy of 'Style.'" *Generative Imagery: Towards a 'New Paradigm' of Machine Learning-Based Image Production,* Special-Themed Issue of *IMAGE: The Interdisciplinary Journal of Image Sciences* 37 (1): 100–111.

Meyer, Roland. 2024a. "It's a Flat World: The Synthetic Realities of Sora." *Rrrreflect: Journal of Integrated Design Research* 1 (1): Special Issue *Hidden Layers*. https://rrrreflect.org/special-issue-1/its-a-flat-world-the-synthetic-realities-of-sora. https://doi.org/10.57684/COS-1267.

Meyer, Roland. 2024b. "Spekulative Strategien: KI-Bilder, Memesis und wilde Forensis." *Fotogeschichte* 172: 38–44.

Meyer, Roland. 2025a. "Images from Images: Generative AI and the Reconfiguration of the 'Photographic.'" *Photography & Culture* 18 (3–4): 281–290. https://doi.org/10.1080/17514517.2025.2522517.

Meyer, Roland. 2025b. "'Platform Realism': AI Image Synthesis and the Rise of Generic Visual Content." *Transbordeur*, February 26, 2025. https://doi.org/10.4000/13dwq.

Miller, Arthur I. 2019. *The Artist in the Machine: The World of AI-Powered Creativity.* Cambridge, MA: The MIT Press.

Miltner, Kate M. 2024. "'A.I. Is Holding a Mirror to Our Society': Lensa and the Discourse of Visual Generative AI." *Journal of Digital Social Research* 6 (4): 13–33. https://doi.org/10.33621/jdsr.v6i440456.

Mitchell, Melanie. 2020. *Artificial Intelligence: A Guide for Thinking Humans.* London: Pelican.

Mitchell, William J.T. 1992. *The Reconfigured Eye: Visual Truth in the Post-Photographic Era*. Cambridge, MA: The MIT Press.

Moskatova, Olga. 2025. "The Aesthetics of Promise: Tech-Failures and Tech-Demonstrations of Generative AI." In *AI Aesthetics: AI-Generated Images between Artistics and Aisthetics*, edited by Jan-Noël Thon and Lukas R.A. Wilde, 75–91. New York: Routledge.

Musk, Elon. 2024. "Kamala vows to be a communist dictator on day one.…." Posted on *X*, September 1, 2024. https://x.com/elonmusk/status/1830656672211103825?lang=de.

Offert, Fabian. 2022. "Ten Years of Image Synthesis." *Zentralwerkstatt*, November 10, 2022. https://zentralwerkstatt.org/blog/ten-years-of-image-synthesis.

Offert, Fabian. 2023. "On the Concept of History (in Foundation Models)." *Generative Imagery: Towards a 'New Paradigm' of Machine Learning-Based Image Production*, Special-Themed Issue of *IMAGE: The Interdisciplinary Journal of Image Sciences* 37 (1): 121–134.

Offert, Fabian and Thao Phan. 2022. "A Sign That Spells: DALL-E 2, Invisual Images and the Racial Politics of Feature Space." Preprint, arXiv, October 26, 2022. https://doi.org/10.48550/arXiv.2211.06323.

OpenAI. 2025. "Introducing 4o Image Generation." *OpenAI*, March 25, 2025. https://openai.com/index/introducing-4o-image-generation.

Philipsen, Lotte. 2025. "Aesthetic Protocols of Popular AI Art." In *AI Aesthetics: AI-Generated Images between Artistics and Aisthetics,* edited by Jan-Noël Thon and Lukas R.A. Wilde, 59–74. New York: Routledge.

Plunkett, Luke. 2022. "DeviantArt Embraces AI Art, Fucks Up Massively." *Kotaku*, November 13, 2022. https://kotaku.com/deviantart-ai-art-nft-crypto-scam-hustle-protest-stable-1849778362.

Ramesh, Aditya, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. "Hierarchical Text-Conditional Image Generation with CLIP Latents." Preprint, arXiv, April 13, 2022. https://doi.org/10.48550/arXiv.2204.06125.

Rettberg, Jill Walker. 2023. *Machine Vision: How Algorithms Are Changing the Way We See the World*. Newark: Polity Press.

Robison, Greg. 2025. "Tokens Not Noise: How GPT-4o's Approach Changes Everything About AI Art." *Medium,* April 1, 2025. https://gregrobison.medium.com/tokens-not-noise-how-gpt-4os-approach-changes-everything-about-ai-art-99ab8ef5195d.

Salvaggio, Eryk. 2023. "The Hypothetical Image: The Aestheticization of Algorithmic Ideologies." *Cybernetic Forests*, November 12, 2023. https://cyberneticforests.substack.com/p/no-mourning-for-synthetic-ruins.

Schuhmann, Christoph, Romain Beaumont, Richard Vencu, et al. 2022. "LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models." Preprint, arXiv, October 16, 2022. https://doi.org/10.48550/arXiv.2210.08402.

Scribner, Herb. 2023. "'Secret Invasion' Creeps out Fans with an AI-Generated Intro Sequence." *The Washington Post*, June 22, 2023. https://www.washingtonpost.com/arts-entertainment/2023/06/21/secret-invasion-ai-intro.

Sharma, Jitendra and Rohita Sharma. 2017. "Analysis of Key Photo Manipulation Cases and their Impact on Photography." *The IISU-JOA Journal of Arts*, 88–99. http://iisjoa.org/sites/default/files/iisjoa/2017/PDF/11.%20Jitendra%20Sharma%20&%20Rohita%20Sharma.pdf.

Shifman, Limor. 2013. *Memes in Digital Culture*. Cambridge, MA: The MIT Press.

Somaini, Antonio. 2023. "Algorithmic Images: Artificial Intelligence and Visual Culture." *Grey Room* 93 (Fall): 75–115. https://doi.org/10.1162/grey_a_00383.

Somepalli, Gowthami, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2022. "Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models." Preprint, arXiv, December 7, 2022. https://doi.org/10.48550/arXiv.2212.03860.

Søraa, Roger A. and Kristine Ask. 2026. "AI and Equality." In *Humanities Perspectives on AI: An Interdisciplinary Introduction*, edited by Lukas R.A. Wilde, Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, 111–126. Oslo: Scandinavian University Press.

Theoretical Media. 2026. "How Seedance 2.0 is SO GOOD (And Why Hollywood is Shook)." *Youtube*, February 16, 2026. https://www.youtube.com/watch?v=yLQClFqzHOU&t=324s.

Thon, Jan-Noël. 2025. "AI Horseplay: Postdigital Aesthetics in AI-Generated Images." In *AI Aesthetics: AI-Generated Images between Artistics and Aisthetics,* edited by Jan-Noël Thon and Lukas R.A. Wilde, 22–58. New York: Routledge.

Thon, Jan-Noël and Lukas R.A. Wilde. 2025. "Introduction: AI Aesthetics." In *AI Aesthetics: AI-Generated Images between Artistics and Aisthetics,* edited by Jan-Noël Thon and Lukas R.A. Wilde, 1–21. New York: Routledge.

Tian, Keyu, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. 2024. "Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction." Preprint, arXiv, June 10, 2024. https://doi.org/10.48550/arXiv.2404.02905.

Treccani, Carloalberto. 2018. "How Machines See the World: Understanding Image Annotation." *NECSUS. European Journal of Media Studies* 7 (1): 235–254. https://doi.org/10.25969/mediarep/3425.

Wakefield, Jane. 2022. "Deepfake Presidents Used in Russia-Ukraine War." *BBC*, March 18, 2022. https://www.bbc.com/news/technology-60780142.

Watkins, Gareth. 2025. "AI: The New Aesthetics of Fascism." *New Socialist*, February 9, 2025. https://newsocialist.org.uk/transmissions/ai-the-new-aesthetics-of-fascism.

Wilde, Lukas R.A. 2023. "Generative Imagery as Media Form and Research Field: Introduction to a New Paradigm." *Generative Imagery: Towards a 'New Paradigm' of Machine Learning-Based Image Production*, Special-Themed Issue of *IMAGE: The Interdisciplinary Journal of Image Sciences* 37 (1): 6–33. http://dx.doi.org/10.25969/mediarep/22326.

Wilde, Lukas R.A. 2025. "KI-Bilder und die Widerständigkeit der Medienkonvergenz: Von primärer zu sekundärer Intermedialität?" In *Bilder im Aufbruch: Herausforderungen der Bildwissenschaft*, edited by Marcel Lemmes, Stephan Packard, and Klaus Sachs-Hombach, 475–507. Cologne: Herbert von Halem.

Yadlin-Segal, Aya and Yael Oppenheim. 2021. "Whose Dystopia Is it Anyway? Deepfakes and Social Media Regulation." *Convergence: The International Journal of Research into New Media Technologies* 27 (1): 36–51. https://doi.org/10.1177/1354856520923963.

Yan, Zhiyuan, Junyan Ye, Weijia Li, et al. 2025. "GPT-ImgEval: A Comprehensive Benchmark for Diagnosing GPT4o in Image Generation." Preprint, arXiv, April 3, 2025. https://doi.org/10.48550/arXiv.2504.02782.

Zylinska, Joanna. 2020. *AI Art: Machine Visions and Warped Dreams*. London: Open Humanities Press. https://www.openhumanitiespress.org/books/titles/ai-art.

9. AI and Music Generation

Andreas Bergsland

Abstract: This chapter focuses on generation of music using AI technologies. It begins by presenting some milestones in algorithmic music preceding current technologies, including dice music from the 18th century and early computer composition from the 1950s. As a case study, it then looks at two AI music generation services, Udio and Suno, and how their prompt-based interfaces and model capabilities invite certain types of user interaction. Furthermore, the chapter examines typical humanities issues related to these services, including representativeness, bias, evaluation, ethics and copyright, as well as showing how AI creators can experience a form of "creator bias" and a sense of achievement towards their AI-generated outputs—despite the apparent effortlessness of the creation process.

Keywords: copyright, creator bias, evaluation, representativeness, virtual artists

In the field of music, AI has entered into a great many areas of application in recent years: analysis and classification, sound source (stem) separation, recommendation and personalization, sound synthesis and processing, timbre and style transformation, voice models, studio production tools, interactive music systems, musical robots, systems for education and learning, and research related to various music-related activities.[1] However, *generating* music with the use of AI has probably gotten even more attention in the media in recent years. For example, in 2018, "Hello world" by Skygge, an album made by a group of musicians and AI researchers, generated headlines like "Is this the World's First Good Robot Album?" (Marshall 2018) and "AI's First Pop Album Ushers in a New Musical Era" (Avdeeff 2018). And in August 2024, *Music Radar* reported that the first ever AI-generated track had hit the German charts (Simpson 2024).

1 In my categorization of application areas, the sources Deruty et al. 2022; Johnson et al. 2023; Magnusson 2019; Miranda 2021; Peeters and Richard 2021; Shang 2019; Zulić 2019 have been particularly influential.

Wilde, Lukas R.A., Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, eds. 2026. *Humanities Perspectives on AI: An Interdisciplinary Introduction*. Scandinavian University Press. https://doi.org/10.18261/9788294167234-26-09
Open access | © The author(s) 2026. This work is licensed under CC BY 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/.

Furthermore, Sony, Universal, Warner and several other major music labels made headlines when they sued AI music generation companies Suno and Udio in June 2024 (Berger 2024). In this chapter, I will begin by presenting a brief introduction to generation of music using AI. Following this, the major part of the chapter will focus on more recent developments, including tools that are accessible to the public such as Suno and Udio. The main contribution of the chapter is how these developments and tools are discussed in light of the issues often raised in the humanities, such as representativity, bias, evaluation, ethics, and copyright. This discussion is based on recent research contributions in the field, but it also introduces the novel perspective of *creator bias* and discusses the sense of achievement related to AI creations.

From Dice to Computers

Western music has a long tradition of generating music through algorithms, even predating computers. This is because *algorithms* do not necessarily need to be implemented on computers; they are simply a set of instructions comparable to a food recipe that can produce a result. Two of the most famous classical composers, Haydn and Mozart, devised what we can safely call algorithmic music generation systems, where dice throws select musical material to form a composition (Magnusson 2019, 113).

One could assume that generating music on the earliest computers would not be a priority in the decades after the Second World War, with the scarcity of computational resources and personnel to operate them. Yet, as early as in 1957, Lejaren Hiller and Leonard Isaacson used the ILLIAC I computer at the University of Illinois to generate the score for *The Illiac Suite*, a string quartet in four movements. This composition process worked with *symbolic* data—i.e. representations of musical notes—since this is a very compact way of representing musical information like pitch and duration. They experimented with different approaches in the four different movements, and several of these are still common today (Steels 2021).

In the late 1980s, while struggling to overcome composer's block, David Cope, an American composer and researcher, developed a computer system for analyzing traits of existing pieces of classical music to generate new music, using randomness to be able to generate many variations. In fact, after several years of refinement he had created a system called EMI (Experiments in Musical Intelligence) that could generate new musical pieces within several classical styles, as a kind of musical emulation. Moreover, as proof of concept, he

recorded several of the pieces that he had generated with his EMI system from the mid-1990s onward—pieces that emulated J. S. Bach, Mozart, Stravinsky, Beethoven, Rachmaninov and others using human musicians and ensembles.

Making Music from Text Prompts: Udio and Suno

Although there is a lot to be said about how different kinds of algorithms have been used to generate music over the years, the restricted format of this chapter makes it necessary to skip to the most recent trends in AI-generated music, namely using deep learning. Deep learning models gained popularity in the late 2000s, when they succeeded in outperforming standard techniques in an image classification task (Briot and Pachet 2018). Since then, this family of techniques has been used with great success for a number of tasks, including music generation, with the aforementioned commercial models *Suno* and *Udio* as the most recent examples of music generation systems that are able to generate musical pieces or excerpts of pieces with only minimal user input. Moreover, many of these no longer work merely on the level of musical notes, but on digital audio data directly.

Starting in the mid-2010s, but accelerating from the early 2020s, a growing number of commercial services have established online services for generating new music with the use of deep learning algorithms, most often marketing themselves as using "AI" to generate music. Some of these are AIVA, Loudly, Boomy, Beatoven, Riffusion, Stable Audio, Soundful, Soundraw, Udio and Suno.[2] In these services, the user can provide some kind of user input, click a button, and then have a full track or an excerpt of music played back a few moments later. Many of these services allow the user to provide tags for genre, mood, use context or something similar, and perhaps some sliders or dropdown menus to control key and meter, but after the huge success of ChatGPT following its launch in 2022, there has been a tendency to integrate Large Language Models (LLMs) in the systems, enabling users to insert text prompts as input. We will take a closer look at Udio and Suno below.

According to the music magazine *Rolling Stone*, the quality of the music that can be generated with Suno and Udio seemed like a breakthrough compared to earlier AI music generators, especially given the way in which they could include "realistic, human-sounding vocals" (Hiatt 2024a). Moreover, a recent

2 In 2024, I wrote a blog post (Bergsland 2024) where I reviewed some of these services. However, this review already seemed outdated at the end of the same year.

study comparing the musical output from twelve different AI music generation systems found that music made by Suno v.3.5 was most often preferred by listeners, followed by Suno v.3 and Udio, thereby supporting the general impression of these services as market leading (Solak et al. 2024). For the average music listener, the music made by these services might sound indistinguishable from any other music produced with human singers and conventional instruments and tools. When I first listened to Udio in April 2024, I remember being impressed by music production details like how instrument groups were located differently between left and right in the stereo image. However, the "sonic fuzziness" mentioned by the *Rolling Stone* reporter in the article introducing Udio can be recognized through closer listening, and this is something that I have learned to do after many hours listening to music from both Suno and Udio. The sound quality seems to be among the current top priorities for these services, since both Udio and Suno listed audio quality as one of three version improvements when introducing the new versions of their services in 2024, Udio v.1.5 and Suno v.4 (Udio Team 2024b.; Team Suno 2024a).

The core feature for both Udio and Suno is clearly the text prompt window, located to the left of a "create" button, with text headlines that emphasize what this is: "Make your music. Create any song. Just describe it" (Udio); "Make a song about anything" (Suno). For example, when I typed "make a sad song about cars" in Udio's prompt window, one of the two song excerpts it generated was a melancholic song somewhere between rock and country with a male vocalist singing about driving at night along "roads so lonely" and where "colors fade on an endless highway." Even if the lyrics are perhaps more about loneliness and melancholy as experienced by someone driving a car, the interpretation is surely acceptable.

The ability to create music through a text interface clearly signals a service that follows in the footsteps of ChatGPT, although in this case these interfaces do not feature a chatbot, but simply a one-way input interface that will steer the creation of a song. The functionality of creating vocals with clearly audible and (potentially) meaningful lyrics together with a fitting accompaniment is highlighted by Suno's slogan mentioned above, and was sufficiently impressive to persuade Microsoft to integrate Suno in their chatbot service, *Copilot* (Wiggers 2023).

The functionality of lyric generation reveals the LLM built into the model; that the lyrics could be about "anything" seemed to be an attractive feature for both services, at least judging from the number of lyrics about doing inappropriate things in inappropriate places that appeared to be flourishing when I tried

Udio and Suno in early 2024.[3] While such lyrics would likely be unthinkable for a serious music artist, the "singer" in these services would have no problem with them and they would probably raise a smile or two for the listener.

Also, generating songs about specific people like friends, colleagues, family members, influencers, or politicians, seems like a popular thing to do. For example, searching for songs in the libraries of generated tracks with the search term "birthday" renders a long list of songs in both Suno and Udio with "birthday" in either the title or the lyrics, most of the time dedicated to a named person. In several of these birthday songs, it is evident that the users have provided the lyrics themselves or have co-written them together with the built-in text generator using editing suggestions from the system. These features have given these users more control and possibilities to fine-tune significant details like typical hallmarks of the individuals in question or how they behave, and perhaps also a chance to be more personal. Furthermore, one can imagine how such a song would be played at a birthday party or some kind of celebration, almost as a kind of gift.

The text prompt is the main input not only for lyrics, but also for steering the musical properties of the song. Both Suno and Udio present prompt windows with grayed-out examples of how a prompt might look. Typically, these examples include one or more genre tags, but they may also include mood, instrumentation, gender of the singer and the type of voice, which decade it is associated with, and other tags that might be associated with the music. Furthermore, the user can write in any text style and be vaguely associative, since the LLM part of the system will "translate" the prompt into tags that can be understood by the system; in fact, in Udio these tags are visible to the user. For example, a suggested prompt by Udio such as "A documentary about the life of a credit card, electro swing, nu jazz" will be translated into the tags "jazz fusion, art punk, avant-garde jazz, introspective, complex, technical, avant-garde, energetic, and manic". Here we can see the traces of the annotated datasets that have gone into the training of the models.

Although one could imagine how the prompt interface would invite users to specify in great detail how they would like a song to be, prompts shared by users suggest that they are not particularly concerned about details most of the time. Since the prompts associated with songs by other users are openly

3 In an interview with Mikey Shulman, co-founder of Suno, he reveals that ChatGPT is built into their service to generate lyrics and titles, see Hiatt 2024b.

visible on the platforms, browsing through them gives the general impression that the majority are relatively short (fewer than ten words), although Udio prompts are sometimes longer (more than twenty words).[4] One could speculate whether this tendency could be attributed to a kind of laziness among users or a lack of precise vocabulary for describing musical qualities in more detail—after all, this is a tool enabling non-experts to produce songs with very little effort and not very precise control. On the other hand, one could say that when using short prompts, the effective functionality would be a lot like other available services that only allow users to choose from different predefined "genre", "mood" and "instrumentation" tags, and that giving users a text prompt with which they can interact is simply a way to make the services seem more reminiscent of ChatGPT and similar AI-based chatbots.[5]

Representativity and Bias

The use of prompts to control the generation of music opens up discussions about the relationship between verbal categories and subjectively experienced music, often imbued with feelings of different kinds, as well as about the degree of musical and expressive control. Although these are both very complex and sometimes philosophical discussions, one need only consider how difficult it is to describe music—particularly without any formal music education or training. In addition, moods, feelings, and emotions that are often part of the category tags in musical datasets are also inherently difficult to describe in words. Jorge Forero and his colleagues (2023) have argued, for instance, how music composed by AI systems conditioned by verbal emotional categories involves several problematic issues including the ambiguities within language itself, the

4 For example, going through twenty of the songs featured on each of the services at the time of writing, Udio users applied an average of sixteen words per prompt, whereas Suno users applied five. The two-hundred-character limit in Suno could be a simple explanation for this. However, it could also be a matter of a culture on the platform, since there does not seem to be a push towards the two-hundred-character limit, at least judging from the very small sample I investigated. After all, most of the twenty featured prompts on Suno were less than fifty characters, and only one was just above a hundred.

5 This can be contrasted to a music generation service like Aiva that allows for a very high degree of control over both musical details and music production decisions. Using an interface that includes the core functionality of what is found in contemporary music production tools known as Digital Audio Workstations (DAWs), user can even edit single notes, adjust the balance between instruments, and apply digital audio effects to one or more instruments on the tracks.

subjective and uniquely individual experiences of music, and the many complexities a piece of music might have.

Another area of discussion relates to the representativity or *bias* (see the section on terminology) of the music models, as discussed above. How well the dataset covers the different possibilities, variations and combinations of tags relating to genre, subgenre, mood, instrumentation and so forth can vary greatly, which in turn has a significant impact on the music the system can generate. A study by a team of researchers from Abu Dhabi (Mehta et al. 2024) found that out of the datasets they investigated, containing over one million hours of audio, 86% of the playback time was covered by music genres from the *Global North*, thereby greatly underrepresenting genres from the *Global South*.[6] The researchers point to how this will, in turn, create problems for the AI music generation systems based on those datasets when the user wants to generate music in the genres of the Global South. These problems include inconsistent quality, instruments and melodies that are out of sync, and that the music generated diverges entirely from the genre. Furthermore, this may have several negative effects on music and musicians from the Global South, such as cultural erosion, reinforcement of existing biases, and economic disparities (Mehta et al. 2024). The research team has therefore suggested a set of measures to be taken to remedy the situation, such as investing in datasets including music from the Global South. Lastly, the systems should also avoid generating results when uncertainty exists, and have the system report on the degree of representativity of the generated music in relation to the requested genre.

Evaluation

How can we evaluate the music generated with artificial intelligence? There are different approaches to this, and in the following we will discuss some of these. In the research literature, a common approach is to ask listeners whether a piece or an excerpt of music has been made by a computer or a human composer, and such a test is usually referred to as a "musical Turing test" (Ji et al.

6 The Global North includes Europe, North America and East Asia, while the Global South includes Africa, Latin America, South Asia, the Middle East, Central Asia and Oceania. The same team also found that while genres like pop, rock, and classical were well represented, children's music and regional genres were very poorly represented.

2023; Xiong et al. 2023; Yang and Lerch 2020).[7] Often the success or efficiency of a model is evaluated by asking listeners whether an excerpt of music is composed by a human (success) or a computer (failure) (Hadjeres et al. 2017; Highams and Olszewska 2023).

The rationale behind this approach to evaluation seems to be that music made by computers is inherently "bad" and music made by humans is "good." This echoes a critical stance towards new technological inventions in musical practices that one can trace back several centuries. Pinch and Bijsterveld (2003) identified new music technologies dating back to the introduction of linked keys on the flute, to self-playing pianos and synthesizers, that were met—at least partly—with critical comments from contemporaries; that they were devoid of soul, artless and had an inhuman clearness and precision, for instance.

The history of music technology has nevertheless shown that many new inventions have been well received and adopted by music practitioners, and that the features that would initially seem alienating or devoid of feeling have later been embraced for the same qualities. For instance, the first synthesizers were initially criticized for producing bad imitations of real instruments, while users instead embraced the instrumental timbres and playing techniques that were unique to synthesizers (Pinch and Bijsterveld 2003, 551). In the same way, music created with AI may have its own unique qualities that can be appreciated by listeners. Stephen Philips, one of the founders of the AI music startup *Popgun* envisioned that this could be a generational issue: "This next generation is probably going to adore AI music. We will hate it because we don't get it, and they will love it because we hate it! So, it will become their thing" (Dredge 2017).

But evaluating the quality of music is of course more than love versus hate and good versus bad. There are so many musical parameters, fine details, and nuances that matter. What complicates these issues even more is that so much outside of music as a sound phenomenon will affect how we evaluate it. As Anne Danielsen (2016) has argued in a text about quality in music, very general values that support quality evaluations among performers and listeners can differ across genres, so that something that is very important to one genre is of little significance for another: Whereas musical skills ("musicianship") are very important in blues, the opposite is the case for indie music. For electronica, for its part, instrumental skills are less central; what matters instead is the

7 The Turing test is named after the British mathematician and computer scientist Alan Turing and was devised to see if a machine could interact with a person through a textual interface and exhibit intelligent behavior indistinguishable from a human (Turing 1950).

ability to find the right sounds and types of sound processing (Danielsen 2016). She furthermore argues that other values, such as originality and expressivity, also differ between these three genres.

With AI-generated music, however, it would make no sense to talk about musicianship, expressivity, or authenticity. Issues related to the personal lives of human artists would also have little meaning; which, at least for certain artists like Taylor Swift and George Michael, plays an important role for many dedicated fans when it comes to listening to and interpreting their songs. It seems, however, that parts of the music industry want to create surrogates: Today, there are AI tools that can create music videos, Instagram posts, music tracks and vocals for *virtual artists*—often crafted by big music and media industry actors—and these have had some success, especially in countries like Japan and Korea (Tencer 2023).[8] This makes it possible, in principle, to create virtual lives that can be related to the respective songs published and marketed in this way. However, this opens up a discussion about political, ethical, and aesthetic aspects that goes beyond the scope of this chapter.

Creator Bias and Sense of Achievement

While evaluation can clearly be affected by our knowledge of whether the music was created by AI or a human creator, listening to music that we have generated ourselves may present a special case. At least, this is something that discussions on the Reddit r/SunoAI forum suggest, and that one user refers to as "creator bias":

> I'd rather listen to my own music on repeat 100 times than some of my old favorite bands. It's why I absolutely do not trust if I made a good song or not it's so difficult the creator bias is INSANE!!!. (Artforartsake99 2024)

Several similar posts seem to indicate that the act of generating songs appears to entail positive emotions, perhaps a kind of pride. From the other comments on the same thread, it seems that part of the attraction also lies in the ability to *personalize* the generated tracks, for example, by making lyrics about topics that interest users personally like dragons, Dante's *Divine Comedy* or Greek mythology, or to fit very specific tastes in music. Thus, generating music with

8 At the time of writing, Suno has a feature called *Personas*, available only to paying subscribers, that supports the creation of virtual artists, at least in the form of a consistent voice and musical style for several songs (see Team Suno 2024b).

AI services like Suno or Udio offers a high degree of personalization and tailoring according to highly specific factors that might be difficult to find in music on conventional streaming platforms. As one user reports:

> Much of my favorite kind of music has died off or changed. With Suno, I can revive my favorite sounds, and it's so amazing to hear lyrics or songs that were a hit from my youth or recreated in a different genre reborn. (dreamstorm84 2024)

One may also observe that both Suno and Udio have adopted several features from music streaming services and social media like following other users, sharing and liking songs, counting number of plays, searching for songs and creating playlists. While this is perhaps not very interesting in itself, having created a song that is liked, shared, or played many times by others does create an implication that a song has qualities, thereby giving the user responsible a sense of achievement. The fact that several users present their songs on other platforms like Spotify, YouTube and Apple Music, often naming themselves as creators instead of the platform they used to generate the song, seems to indicate something similar. For example, *Feline Music*, a Suno user, has a YouTube channel featuring over three hundred videos at the time of writing, and has released two full albums on Spotify with songs that all have what is probably the real name of the user credited as "song creator."[9] Moreover, after having spent many hours creating songs with Udio myself, I can relate to this feeling of achievement when I have generated a song that I like. Thus, even if typing in a few words and clicking some buttons perhaps does not seem like an activity that should engender either admiration or a sense of achievement, there can still be some element of achievement—maybe similar in some ways to the act of art curation, where selecting, organizing and presenting a work of art is a prized skill. Here, seeking out, selecting, organizing and displaying music tracks becomes the skill that matters.

Ethics and Copyright

The availability of AI-based systems that can generate content on the basis of large databases of other people's artistic or cultural expressions has generated considerable debate and controversy regarding ethical and copyright issues (Koch and Ruvenal 2023). While some generative AI models are based on

9 As mentioned, Suno released the "Personas" feature in late 2024 (see Team Suno 2024b), enabling the user to associate several songs with the same lead singer and style of music.

publicly available datasets with licenses that allow downloading and sharing, other models may be based on data collected by big tech companies delivering social media platforms or file storage/sharing services, and some may be based on data downloaded from the internet by automated processes, often referred to as "scraping" or "crawling."[10] As for the latter, Suno has openly admitted that their models are trained on "tens of millions of recordings" they have found on the internet (Shulman 2024).

On one side of the controversy, artists, copyright holders, and their representatives claim that big AI companies have stolen their work from the internet, and that they therefore should have been asked for permission and compensated for any use. While on the other side, the AI companies claim that using media content from the internet should be considered "fair use," and therefore that they should not have to pay copyright holders. Not surprisingly, this controversy has resulted in a number of lawsuits against several AI companies for using copyrighted media content like music, images, and text to train generative AI models without permission.[11] As for music, three of the largest record labels in the world, Sony Music, Universal Music and Warner Music Group, made headlines when they sued Udio and Suno in June 2024 for copyright infringement (see, for example, Sherman 2024 or Carras 2024). Central to the argument of the plaintiffs is that the commercial purpose of the AI products undermines existing and potential sales and therefore directly infringes on the market of the original music (Berger 2024). Moreover, they have demonstrated with transcriptions of musical scores how Suno and Udio might create pieces of music that are not only "vaguely derived" from but copied from the original music recordings. Udio, on the other hand, has claimed in a statement on their website that they are "completely uninterested in reproducing content in our training set, and in fact, have implemented and continue to refine state-of-the-art filters to ensure our model does not reproduce copyrighted works or artists' voices" and that instead, what their models learn is an "understanding of musical ideas" made up of basic building blocks "owned by no one" (Udio Team 2024). The result of the lawsuit will probably create a precedent that will have significant repercussions for both music creators and AI music businesses.

---

10 Examples of the former are Google and Meta, which harvest data from their own products such as Google Photo and Facebook to train image recognition systems.

11 See "AI Copyright Lawsuits in the US", provided by *Wired*, for an overview of copyright lawsuits against AI companies over copyright in the US.

The fact that music creators are worried about the current situation became evident in April 2024, when a coalition of about two hundred named artists and songwriters including Nicki Minaj, Imagine Dragons, Sigrid and Katy Perry composed an open letter to AI developers, technology companies, and digital music services to make them stop using AI to "infringe upon and devalue the rights of human artists" (Artist Rights Alliance 2024). The letter contended that AI-powered service providers aim to replace human-made music and that this will ultimately threaten the livelihoods of musicians and music creators. The fact that this might actually happen was demonstrated when 850 radio DJs and producers lost their jobs with iHeartRadio in 2020 following restructuring and investment in AI technology (Clancy 2021, 148). Moreover, within academia there have also been concerns about how AI-generated music is replacing human compositions in applications where music already takes a background role, such as in restaurants (Lin and Chen 2024), advertising, corporate videos, and computer games (Intel AI 2019). This negative development may continue in the future, according to Tono, the association for music copyright holders in Norway, which has estimated that AI may reduce music creators' income by 24% by 2028 (Martinsen 2024). Additionally, music streaming providers are known to have started creating their own musical content with AI. Since streaming providers can easily promote AI-generated tracks on their own curated playlists, those providers can increase their revenue; whereas music creators might miss out on income (Clancy 2021, 336).

There can also be positive sides to this, according to Epstein and colleagues (2023), who note how AI might increase productivity for some music creators. They further speculate that although AI may threaten some occupations within the music ecosystem, it may also create new ones, even if the authors are not specific on that point. Although it may be too early as yet to draw any conclusions about the impact of AI on musicians' livelihoods, the outcome of the lawsuits referred to will probably determine whether services like Udio and Suno will continue to make it economically attractive for producers to generate music with AI for media content like YouTube, Twitch, podcasts and the like.[12]

An additional ethical issue regarding music made with AI is particularly relevant in the current situation in which AI-generated music is approaching

12 According to a survey conducted by Mubert—an AI music service provider—and published on their blog, 62% of AI-generated music is intended for YouTube (see Mubert Team 2023).

streaming quality: Should listeners always be informed about its true origin? Driving in my car on the way home one day just before Christmas, I was listening to the radio when a song about camels by Serkan Kahveci was announced (NRK Radio 2024). When the song started playing, I was surprised to hear the gritty sound quality of what clearly seemed to me to be music made with AI. In a sense, I felt deceived by the radio station that had initially given me the impression that this was created by a human music artist, which in turn made me unsure of the critical sense of the journalist or producer who had chosen that piece of music. In short, my trust in the radio station was questioned.

My experience highlights the issue of "trustworthy AI", which was already addressed by the EU Commission in 2019 when formulating their *Ethics guidelines for trustworthy AI* (High-Level Expert Group on AI 2019). Although questions about trustworthy AI are probably more acute for AI-generated text, images and videos, which can be designed to appear as real and, in a worst-case scenario, be designed to mislead, issues of trust are also relevant in music (Hazzard et al. 2024; Moura and Maw 2021). Music listening often involves an empathetic and embodied stance where we might experience engagement and involvement with the artist or musician. When the listener knows that the artist, and perhaps especially the *voice,* is AI-generated, this might disrupt such engagement: As Moura and Maw (2021, 144) note, a "lack of information regarding the use of AI during the composition process may affect the trust relationship between listener and producer, artist or label." The way forward in this situation might be an ethical "marking system" for AI, as suggested by Clancy (2021, 337), where the listener is given information about the true origin of the music, and therefore may not feel deceived or tricked.

Conclusions

From the development of music generated with algorithms all the way from Hiller and Isaacson's *Illiac Suite* to today's state-of-the-art AI tools such as Suno and Udio, it is evident that the technologies, the ways in which users interact with them, and the musical output generated by them has changed a lot. It is unquestionable that commercial web-based services like Suno and Udio have greatly simplified how users can generate music with AI models, thereby lowering the bar so that "anyone" can produce music. Even if this chapter has argued how several critical points can be made with regard to the generation of AI music or the services that provide this, such as representativity, bias,

ethics, and copyright, one can also observe how users may experience positive effects such as "creator bias" and a sense of achievement with regard to their own creations. The spread of generative AI music into the public sphere may be related to a more general tendency toward "democratization" in music-making, which has been heralded in academia by writers like Jacques Attali (1985) as well as tech gurus like Steve Jobs (Clancy 2021, 94). Although it is likely that AI will have a major impact on music in the coming years, with probably even greater economic, ethical, aesthetic and creative impact, the cultural tendency towards counterreactions may also lead to a situation in which people will instead seek the opposite of what AI has to offer: acoustic music, played live and by real people.

Works Cited

"AI Copyright Lawsuits in the US." n.d. Accessed March 09, 2026. https://flo.uri.sh/visualisation/20825192/embed.

Artforartsake99. 2024. "I'd Rather Listen to My Own Music on Repeat 100 Times Than Some of My Old Favorite Bands." Reddit, December 4, 2024. https://www.reddit.com/r/SunoAI/comments/1h664n9/comment/m0biurj/.

Artist Rights Alliance. 2024. "200+ Artists Urge Tech Platforms: Stop Devaluing Music." *Artist Rights Alliance*, April 1, 2024. https://artistrightsnow.medium.com/200-artists-urge-tech-platforms-stop-devaluing-music-559fb109bbac.

Attali, Jacques. 1985. *Noise: The Political Economy of Music*. Minneapolis: University of Minnesota Press.

Avdeeff, Melissa. 2018. "Ai's First Pop Album Ushers in a New Musical Era." *The Conversation*, October 3, 2018. https://theconversation.com/ais-first-pop-album-ushers-in-a-new-musical-era-100876.

Berger, Virginie. 2024. "Why Major Labels Are Suing AI Music Startups Udio and Suno for Mass Copyright Infringement." *Forbes*, June 27, 2024. https://www.forbes.com/sites/virginieberger/2024/06/27/why-major-labels-are-suing-ai-music-startups-udio-and-suno-for-mass-copyright-infringement.

Bergsland, Andreas. 2024. "Some Experiences of AI-based Tools for Audio and Music." *Andreas Bergsland: Personal Home Page,* April 22, 2024. https://andbe.folk.ntnu.no/2024/04/22/ai-tools-overview.

Briot, Jean-Pierre and François Pachet. 2018. "Music Generation by Deep Learning-Challenges and Directions." Preprint, arXiv, September 30, 2018. https://doi.org/10.48550/arXiv.1712.04371.

Carras, Christi. 2024. "Major Record Labels Sue AI Companies for Allegedly Stealing Copyrighted Music." *Los Angeles Times*, June 26, 2024. https://www.latimes.com/entertainment-arts/business/story/2024-06-24/riaa-suno-udio-lawsuit-ai-copyright-songs-music.

Clancy, Martin. 2021. *Reflections on the Financial and Ethical Implications of Music Generated by Artificial Intelligence*. PhD Thesis, Trinity College. https://www.tara.tcd.ie/items/7e1e0547-f981-4186-ad16-a5f41114f90a.

Danielsen, Anne. 2016. "Nyskapende, Sterkt Og Kompetent! Kvalitetsforståelser i musikkfeltet i lys av tre populærmusikalske sjangre." In *Kvalitetsforståelser: Kvalitetsbegrepet i samtidens kunst og kultur*, edited by Knut Ove Eliassen and Øyvind Prytz, 104–119. Oslo: Norsk Kulturråd.

Deruty, Emmanuel, Maarten Grachten, Stefan Lattner, Javier Nistal, and Cyran Aouameur. 2022. "On the Development and Practice of AI Technology for Contemporary Popular Music Production." *Transactions of the International Society for Music Information Retrieval* 5 (1): 35–49. https://doi.org/10.5334/tismir.100.

dreamstorm84. 2024. "I Have Completely Converted to Listening to My Own Songs. I Make Music That I Would Like to Listen To." Reddit, December 4, 2024. https://www.reddit.com/r/SunoAI/comments/1h664n9/comment/m0eq86o/.

Dredge, Stuart. 2017. "Playing Next Door to Alice: Popgun Reveals Its Music AI." *Music Ally*, June 30, 2017, https://musically.com/2017/06/30/playing-alice-popgun-music-ai/.

Epstein, Ziv, Aaron Hertzmann, and The Investigators of Human Creativity. 2023. "Art and the Science of Generative AI." *Science* 380 (6650): 1110–1111. https://doi.org/10.1126/science.adh4451.

Forero, Jorge, Gilberto Bernardes, and Mónica Mendes. 2023. "Are Words Enough? On the Semantic Conditioning of Affective Music Generation." Preprint, arXiv, November 7, 2023. https://arxiv.org/abs/2311.03624.

Hadjeres, Gaëtan, François Pachet, and Frank Nielsen. 2017. "Deepbach: A Steerable Model for Bach Chorales Generation." *Proceedings of the 34th International Conference on Machine Learning, PMLR* 70:1362–1371, Sydney, Australia, 6–11 August. https://proceedings.mlr.press/v70/hadjeres17a.html.

Hazzard, Adrian, Craig Vear, and Solomiya Moroz. 2024. "Towards a Hierarchy of Trust in Human-AI Music-Making." *GenAICHI: CHI 2024 Workshop on Generative AI and HCI*. https://generativeaiandhci.github.io/papers/2024/genaichi2024_28.pdf.

Hiatt, Brian. 2024a. "AI-Music Arms Race: Meet Udio, the Other ChatGPT for Music." *Rolling Stone*, April 10, 2024. https://www.rollingstone.com/music/music-features/udio-ai-music-chatgpt-suno-1235001675.

Hiatt, Brian. 2024b. "A ChatGPT for Music Is Here: Inside Suno, the Startup Changing Everything." *Rolling Stone*, March 17, 2024. https://www.rollingstone.com/music/music-features/suno-ai-chatgpt-for-music-1234982307.

High-Level Expert Group on AI. 2019. "Ethics Guidelines for Trustworthy AI." *European Commission*, April 8, 2019. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Highams, L. and J.I. Olszewska. 2023. "Markov-Chain-Based Musical Creative Intelligent Agent Passing Successfully the Turing Test." *2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES)*, Nairobi, Kenya, 26–28 July. https://doi.org/10.1109/INES59282.2023.10297645.

Intel AI. 2019. "Can AI Crack the Code for Creativity?" *Forbes*, June 6, 2019. https://www.forbes.com/sites/insights-intelai/2019/05/22/can-ai-crack-the-code-for-creativity.

Ji, Shulei, Xinyu Yang, and Jing Luo. 2023. "A Survey on Deep Learning for Symbolic Music Generation: Representations, Algorithms, Evaluations, and Challenges." *ACM Comput. Surv*. 56 (1): 1–39. https://doi.org/10.1145/3597493.

Johnson, Colin, Nereida Rodríguez-Fernández, and Sérgio M. Rebelo, eds. 2023. *Artificial Intelligence in Music, Sound, Art and Design: Lecture Notes in Computer Science*. Cham: Springer.

Koch, Tommaso and Caio Ruvenal. 2023. "Artificial Intelligence Clashes with Copyright: Is It Stealing Thousands of Protected Creations?" *El País English*, November 6, 2023. https://english.elpais.com/culture/2023-11-06/artificial-intelligence-clashes-with-copyright-is-it-stealing-thousands-of-protected-creations.html.

Lin, Tsen-Fang and Liang-Bi Chen. 2024. "Harmony and Algorithm: Exploring the Advancements and Impacts of AI-Generated Music." *IEEE Potentials* 43 (6): 23–30. https://doi.org/10.1109/MPOT.2024.3433888.

Magnusson, Thor. 2019. *Sonic Writing: Technologies of Material, Symbolic, and Signal Inscriptions*. New York: Bloomsbury.

Marshall, Alex. 2018. "Is This the World's First Good Robot Album?" *BBC*, January 12, 2018. https://www.bbc.com/culture/article/20180112-is-this-the-worlds-first-good-robot-album.

Martinsen, Willy. 2024. "KI Kan Kutte Musikkskaperes Inntekter Med 24 Prosent Innen 2028." *Tononytt.Tono*, December 4, 2024. https://www.tono.no/ki-kan-kutte-musikkskaperes-inntekter-med-24-prosent-innen-2028.

Mehta, Atharva, Shivam Chauhan, and Monojit Choudhury. 2024. "Missing Melodies: AI Music Generation and Its 'Nearly' Complete Omission of the Global South." Preprint, arXiv, December 5, 2024. https://doi.org/10.48550/arXiv.2412.04100.

Miranda, Eduardo Reck. 2021. *Handbook of Artificial Intelligence for Music*. Cham: Springer.

Moura, Francisco Tigre and Charlotte Maw. 2021. "Artificial Intelligence became Beethoven: How Do Listeners and Music Professionals Perceive Artificially Composed Music?" *Journal of Consumer Marketing* 38 (2): 137–146. https://doi.org/10.1108/JCM-02-2020-3671.

Mubert Team. 2023. "A Remarkably In-depth Study of AI Music: Creators, Composers, Adopters." *Mubert Blog,* January 2, 2023. https://mubert.com/blog/an-in-depth-study-into-ai-music-its-creators-composers-and-adopters.

NRK Radio. 2024. "Året då woke døydde." *NRK Kompass*, December 19, 2024. https://radio.nrk.no/serie/kompass/sesong/202412/DMTL36415824.

Peeters, Geoffroy and Gaël Richard. 2021. "Deep Learning for Audio and Music." In *Multi-Faceted Deep Learning: Models and Data*, edited by Jenny Benois-Pinot and Akka Zemmari, 231–266. Cham: Springer Nature.

Pinch, Trevor J. and Karin Bijsterveld. 2003. "'Should One Applaud?': Breaches and Boundaries in the Reception of New Technology in Music." *Technology and Culture* 44 (3): 536–559. http://www.jstor.org/stable/25148160.

Shang, Mingli. 2019. "The Application of Artificial Intelligence in Music Education." In *Intelligent Computing Theories and Application. ICIC 2019. Lecture Notes in Computer Science* 11644, edited by De-Shuang Huang, Kang-Hyun Jo, and Zhi-Kai Huang, 662–668. Springer: Cham. https://doi.org/10.1007/978-3-030-26969-2_62.

Sherman, Natalie. 2024. "World's Biggest Music Labels Sue over AI Copyright." *BBC*, June 25, 2024. https://www.bbc.com/news/articles/ckrrr8yelzvo.

Shulman, Mikey. 2024. "The Future of Music: An Update on the Recent Legal Action from the Riaa and What Is at Stake." *Suno Blog*, August 1, 2024. https://suno.com/blog/future-of-music.

Simpson, Will. 2024. "The First Ever AI-Generated Track to Hit the Charts Has Landed in Germany, and People Aren't Happy." *Music Radar*, August 16, 2024. https://www.musicradar.com/news/first-ever-ai-chart-germany.

Solak, Ahmet, Florian Grötschla, Luca A. Lanzendörfer, and Roger Wattenhofer. 2024. "Benchmarking Music Generation Models and Metrics Via Human Preference Studies." *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*. OpenReview. https://doi.org/10.3929/ethz-b-000715180.

Steels, Luc. 2021. "Foreword: From Audio Signals to Musical Meaning." In *Handbook of Artificial Intelligence for Music*, edited by Eduardo Miranda, v–xviii. Cham: Springer Nature.

Team Suno. 2024a. "Introducing v4." *Suno Blog*, November 19, 2024. https://suno.com/blog/v4.

Team Suno. 2024b. "What are Personas?" *Suno Blog*, 2024. https://help.suno.com/en/articles/3484161.

Tencer, Daniel. 2023. "Meet the Virtual Artists Backed by Some of the World's Biggest Entertainment Companies." *Music Business Worldwide*, May 22, 2023. https://www.musicbusinessworldwide.com/meet-the-virtual-artists-backed-by-some-of-the-worlds-biggest-entertainment-companies1.

Turing, Alan M. 1950. "I.—Computing Machinery and Intelligence." *Mind* LIX 236: 433–460. https://doi.org/10.1093/mind/LIX.236.433.

Udio Team. 2024a. "AI and the Future of Music." *Udio Blog*. July 23, 2024. https://www.udio.com/blog/ai-and-the-future-of-music.

Udio Team. 2024b. "Introducing v1.5." *Udio Blog*, July 23, 2024. https://www.udio.com/blog/introducing-v1-5.

Wiggers, Kyle. 2023. "Microsoft Copilot Gets a Music Creation Feature via Suno Integration." *TechCrunch*, December 19, 2023. https://techcrunch.com/2023/12/19/microsoft-copilot-gets-a-music-creation-feature-via-suno-integration/.

Xiong, Zeyu, Weitao Wang, Jing Yu, Yue Lin, and Ziyan Wang. 2023. "A Comprehensive Survey for Evaluation Methodologies of AI-Generated Music." Preprint, arXiv, August 26, 2023. https://doi.org/10.51191/issn.2637-1898.2019.2.2.100.

Yang, Li-Chia and Alexander Lerch. 2020. "On the Evaluation of Generative Models in Music." *Neural Computing and Applications* 32 (9): 4773–4784. https://doi.org/10.1007/s00521-018-3849-7.

Zulić, Harun. 2019. "How AI Can Change/Improve/Influence Music Composition, Performance and Education: Three Case Studies." *INSAM Journal of Contemporary Music, Art and Technology* 1 (2): 100–114. https://doi.org/10.51191/issn.2637-1898.2019.2.2.100.

10. Understanding Literature in the Age of AI

Carolin Führer

Abstract: This chapter addresses the inroads made by AI into aesthetic and creative fields, previously regarded as a genuinely human domain. The author asks how we can create, read, and judge literary texts in the context of LLM use, exploring first how AI challenges traditional ideas of authorship, originality, and aesthetic value in literature, then how AI changes literary reception and interpretation in general, and finally how AI influences aesthetic judgement and creativity more broadly. The chapter concludes by calling for new forms of critical AI literacy specific to the literary domain, and for a rethinking of aesthetic communities capable of resisting the standardizing pressures of algorithmically driven literary culture.

Keywords: AI literacy, authorship, co-creativity, literary reading, synthetic literature

> After Japanese author Rie Kudan won one of the country's most prestigious literary awards, she admitted she'd received help from an unusual source—ChatGPT. […] The author then confirmed at a press conference that around 5% of her book 'The Tokyo Tower of Sympathy'—which was lauded by committee members as 'practically flawless'—was word-for-word generated by AI. (Choi and Annio 2024)

Following this, a worldwide discussion ensued, touching upon a central concern in the question of AI and literary writing: whether AI-generated narrative texts threaten the originality and uniqueness of human creativity (Catani 2024). The creation of literary texts with the help of a machine that can imitate virtually all voices and patterns prevalent in the canonical world of literature substantially influences literary production. Literary texts are based upon the interplay and composition of semantic (i.e., content-related) and

Wilde, Lukas R.A., Miriam Kyselo, Thomas Brandt, Guri Hanem, and Roger A. Søraa, eds. 2026. *Humanities Perspectives on AI: An Interdisciplinary Introduction*. Scandinavian University Press. https://doi.org/10.18261/9788294167234-26-10
Open access | © The author(s) 2026. This work is licensed under CC BY 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/.

formal structures. These were, for the longest time, generated and understood on the basis of a physical anchoring to the perceptions, feelings, and actions of (writing and reading) human agents. Since the emotional and cognitive evaluation of such subject matter is, in general, grounded in human literary reception (Magirius and Führer 2023), the inclusion of AI in literature raises far-reaching questions, not only from the point of view of the authenticity of human experience. By reducing language to quantifiable units, Large Language Models (LLMs) replace the critical power of language with the numerical principle and—in terms of fine-tuning—(moral) normativism. Thus, the interweaving of literary reception and production in the age of AI also concerns knowledge management and ethical questions. From an epistemological perspective, the fundamentals of literary reception lie in the ambiguity of language, and in the diversity of viewpoints on life and culture as models of reality, in general (Vendrell Ferran 2018, 2023). On the one hand, the variety of language and the diversity of knowledge and perspectives have traditionally demanded in-depth reading and hermeneutic understanding of the cultural, historical, and personal backgrounds that give a text its meaning. On the other hand, generated texts seem to require only superficial reading due to the inherent nature of LLMs. Comprehending the outputs of LLMs as "meaningful"—*as if* they were coherent discourses or human "voices"—can be criticized since LLMs have no consciousness, no intentionality. Their texts are generated from probability calculations. Attributing meaning is therefore a human projection that is not an inherent property of texts produced by any such model. The danger here is in reinforcing the anthropomorphization of AI by regarding LLM-generated texts in (excessively) human terms and concepts. But texts generated by LLMs cannot simply be dismissed as "texts written by machines," either; rather, they represent a new type of text that reveals patterns of a neoliberal cultural system. The heart of AI's impact on literature, however, lies in questions of aesthetic theory and practice as consequences of the fundamental challenge to our conventional understanding of authorship. Traditional literary ethics have long been grounded in assuming human intentions behind creative works. If we accept AI as a creative collaborator or a medium, we must grapple with questions of intellectual property, copyright ownership, and (the absence of) fair compensation for human creators on whose works these systems are trained (Caldwell 2023). It is, of course, impossible in the scope of a short chapter to provide a complete overview of research into relevant literary developments or a comprehensive history of literary writing with the involvement of machines.

The focus of this introduction to AI and literature is thus instead on how we can read and judge literary texts in the context of AI use. Firstly, ultimately no one knows what constitutes "good" literature; secondly, the statistical regime of AI is increasingly influencing literary reception. The first section thus explores, quite broadly, how AI challenges traditional ideas of authorship, the idea of what constitutes a work, originality, and aesthetic value in literature. It argues that while LLMs can mimic human writing seamlessly, they lack intentionality and consciousness, which raises important questions about authenticity, ownership, cultural conformity, and moral discourse. Concepts like co-creativity, hybridity and emergence are thus proposed to capture these human–machine literary texts. In the second section, the focus is on how AI tools and LLMs are changing literary reception and interpretation in general. AI tools can act as reading partners, supporting comprehension, analysis, and interpretation, potentially making literature more accessible and inclusive. But dealing with literature has always involved not only cognitive "understanding" and rational comprehension of a source text, but also psychological, embodied, emotional, cultural, and moral factors. Therefore, AI may also diminish emotional depth, differences in cultural backgrounds, associated aesthetic engagements, and authentic perspective-taking, while reinforcing Western biases inherent in their training data (Lucy and Bamman 2021). Finally, the last section examines how AI influences aesthetic judgements and creativity more broadly. For example, poetry is perceived by many as "good" and exciting when the new is not *too* unfamiliar and the familiar is not *too* predictable; for this reason, AI also has agency as a "kind of canny valley that forms the ideal of maximum comprehensibility" (Bajohr and Cotten 2025, 18; my translation). While AI may stimulate new ideas, it also runs the risk of homogenizing literary expressions and weakening responsibility for deep interpretation and ambiguity-sensitive thinking. Individual critical AI literacy is one—though not the only—answer to this complex situation; rather, we must also collectively rethink the formation of our literary communities and our aesthetic *knowledge frames* in society.

Literariness and Authorship in the Age of AI

Since the 18th century, restricting creativity and aesthetics to individual human authorship and linking them to concepts of artistic and epistemic freedom has been a powerful and influential aesthetic tradition in the Western world (Robert 2024). *Aesthetic autonomy* here refers to the autonomy of the artwork. This ideal has never been truly achieved in artistic practice since art always

exists in a social context. But what does it mean for the future assessment of literariness and aesthetics if the distinction between natural and artificial text becomes difficult to define (Gunser et al. 2021, 2022; Köbis and Mossink 2021), or even obsolete (Führer and Knapp, forthcoming; Porter and Machery 2024)? Automatic writing had already begun with the Surrealists and the potential literature of the Oulipo group and extended forward to Theo Lutz's *stochastic texts* and Edwin Morgan's *computer poems*. Max Bense, who was actually an early proponent of what was later called *digital poetry*, argued that authorship requires intentionality, which is not present in the case of *simulated poetry* (Bense 1998, 143; see also Bajohr 2024, 266). Back in the 1960s when he wrote this, the assumption of "computer authorship" was considered an inadmissible anthropomorphism, "which is based solely on ignorance of how it [the computer] works" (Stickel 1966, 123; see also Bajohr 2024, 267). Even before the advent of language models, there was a tendency on the internet to deal with texts more as general ideas, rather than as individual expressions or personal attitudes of authentic persons. Nowadays, it can no longer be assumed that there actually is a person behind an idea, but the "standard expectation of literary texts to this day is [still] that they have authors in the sense of people who are willing to communicate" (Bajohr 2023a, 44; my translation).

In literary studies, the role of the author in literary communication has been subject to critical reflection since at least the mid-20th century, a development famously epitomized by Roland Barthes' 1967 essay "The Death of the Author" (Barthes [1967] 2000). Barthes argued that the reliance of traditional literary criticism on the author's intentions, their biography, and their personal context (to explain the "ultimate meaning" of a text) is fundamentally flawed. He contends that once a work is written, the author loses control over its meaning and interpretation. Barthes puts forward the hypothesis that the "birth of the reader" must hence be at the cost of the "death of the author." He also argues that meaning emerges not from the author's intentions, but through the reader's active interpretation and engagement with a text. The emergence of AI in the literary field, however, also fundamentally challenges traditional reader-response theories. These usually locate meaning in the dynamic interaction between text and reader. When AI systems participate in the production of literature, they lack the subjective experiences, cultural and historical backgrounds, and emotional responses that human writers bring to texts. Deconstructionists like Paul de Man (1971) could hardly have imagined a future that would take their theory, according to which all our ideas are essentially "text," to such a literal extreme. LLMs can now be said

to define a new "pantextuality." This kind of pantextuality flattens out all data into statistical information with no distinctions of social, cultural, and historical context. One problem here is that the training data produces an abstract (inauthentic), non-representative "average" from the aesthetic languages and styles of the past. Originality is difficult to achieve in such a culture of standardization rooted in a tradition of empirical positivism. Another problem is that data obscures the technical, economic, and social conditions of production (training data, infrastructure, labor exploitation, political and moral fine-tuning standards, etc.; Crawford 2021). As a result, the focus shifts from questions of power in terms of authorship toward the sign systems of texts and their cultural implications. This seems to be close to older theories, where all meaning is entirely "immanent" within a text. Structuralist literary theory assumes that meaning in a text arises from underlying structures and relationships between elements, not from the intentions of author or reader. Literature is viewed as a system of signs governed by universal rules, similar to language itself "below." The goal of an analysis, then, is to uncover these deep structures that shape all narratives. The problem with this view is that we lose awareness of historical, religious, and cultural differences of aesthetic expressions in favor of an unreal cosmopolitanism that appears to be mathematically achievable (Simanowski 2025, 158).

We need to find a new, more appropriate vocabulary to describe such "machine authorships" and "authorless literature" (Simanowski 2023). Hannes Bajohr, for example, has investigated and distinguished different "forms of authorship" (Bajohr 2024, 271) with regard to how directly (or indirectly) a human being influences and controls the text results: In primary authorship, a human writes directly; on a secondary level, a human writes "a sequence of rules" (as in a pre-AI tool) that then writes the text; on a tertiary level, a human merely sets up a "self-learning" AI model and system that then generates new texts through probability patterns it can replicate; and finally, on a quaternary level, a human merely provides inputs ("prompts") for a "black-boxed" system of factory settings—AI seemingly does the rest (Bajohr 2024, 271–272). Contemporary literature in a culture of AI thus originates more and more from a hybrid network of humans, datasets, hardware, and software. Attempts are currently being made to establish new means of description for this, and new terms such as "generated literature" (Montfort and Heflin 2024), "entangled authorship" (Sergi 2024) or "synthetic literature" (Piorecký and Husárová 2024) can be found in literary studies. Karel Piorecký and Zuzana Husárová critically reflect on the role of AI in a changing world:

> From our perspective, AI is not changing the world order or the "order" of art and literature—it initiates or accelerates partial changes that in fact are not marginal […], but neither does it deviate from the paradigm of digital culture that has now been established for decades and so does not constitute a revolutionary change. (Piorecký and Husárová 2024, 1)

Even if we speak less of "originality" and more of a form of emergence (Krakauer et al. 2025) in referring to AI aesthetics, this still raises various questions—for instance, who owns the resulting texts? And, last but not least, what about political writing activists? Amanda Gorman, Ljudmila Jewgenjewna Ulizkaja, or Rupi Kaur have had a global impact on political discourses (on [German] traditions of literary activism, see Çakir 2024). The susceptibility of AI to bias and its control by powerful companies that train and regulate models may allow manipulation of literary activist messages through algorithmic censorship, shadow-banning, or deliberate flooding of discourse with inauthentic or misleading content. This risks the disempowerment and fragmentation of poetry activist networks. Forms of artistic activism could of course also benefit from AI in the future: Projects such as Bot Poets (e.g. *DeepDrumpf*, *Pentametron, Everyword 2007–2014)* or AI Writing Collectives (e.g. *Botnik Studios, 0x0, The Uncreative Writing Community)* demonstrate how algorithmic co-creativity can be used as a tool for empowerment. AI-based, distributed authorship reveals how language and meaning are always relational, socially constructed, and technologically mediated.

Reading and Interpreting Literature with AI

The comprehension of literary texts takes place through feelings as well as through the physical and action-related experiences of human authors and their readers (Barsalou 1999; 2008; Fischer and Zwaan 2008; Glenberg 2011; Lakoff and Johnson 1980; 1999; Scherer and Wallbott 1994). For example, in Lawrence W. Barsalou's conception of the cognitive system as a "perceptual symbol system," conceptual mental representations gain their semantic "grounding" only through the actual perceptual experiences associated with them. This "grounding," based on actual subjective experiences of conceptual structures, affects the aesthetic perception of texts, particularly in comparison to the perception of music or visual art. This is because the physical and emotional experiences associated with the literary reception are primarily triggered

by an inner "replay" of (memorized) concept-linked experiences and not, as in other forms of art reception, by actual sensory experiences. Such aspects of the reception of literary texts have been further differentiated in terms of cognitive psychology using the "foreground-background hypothesis" (Jacobs and Willems 2018, 148) and the "defamiliarization theory" of David Miall and Don Kuiken (1994). Despite their impressive "fluency," large language models like ChatGPT are still very different from human language and cognition, especially in terms of acquisition and evolution (Cuskley et al. 2024). It will be worth investigating how this affects literature socialization and whether readers develop divergent frameworks for dealing with literature due to new and different reading modes.

From a contemporary sociological, educational, and—ultimately—practical perspective on literature, the situation could be described as follows: AI tools can be used even today to promote reading literacy (see, for example, laletu.de for reading fluency), but as developments progress they could also take on the role of "literary dialogue partners." The interaction with AI tools can both promote literary reading attitudes (Nix and Führer 2024) and support basic literary understanding for the purpose of discussion, analysis, and interpretation of literary texts (Führer and Nix 2023). AI tools could accompany readers before, during, and after their reading and take on the role of "master readers" who act as reading models for learners: They can, after all, explain the meanings of words, establish or discuss local and global associations and coherences, provide approaches for interpretation, and discuss subjective text meanings with readers. Furthermore, they can provide external knowledge that is potentially relevant to the understanding of literary texts (e.g. knowledge of text types, history, or literary theory). Of course, these functions depend on the assumption that the currently prevalent tendencies of some AI tools to simply "hallucinate" incorrect "facts" will be significantly reduced, for instance through further improvement of the underlying LLMs and/or better search engine integration. It has been observed that language models frequently produce statements that are mostly accurate from a literary perspective, especially in the case of canonical texts.

What seems to be at stake here is thus how any LLM can be classified in literary terms with regard to interpretation of texts. Any interpretation remains procedural and incomplete. In a simplified and schematic description of interpretation processes, the *Interpretans* (see figure 10.1 below) is the result of interactions between the reader (= *Subject*) and the text (= *Interpretandum*):

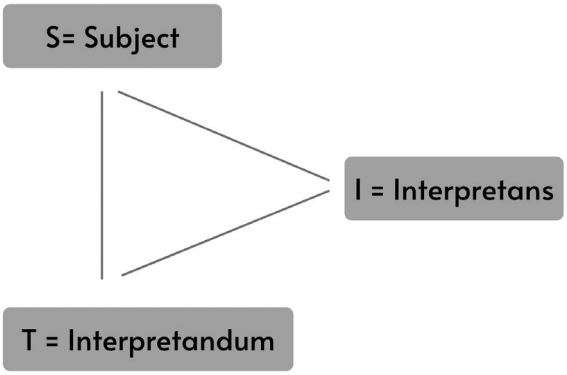

Figure 10.1: Model representation of interpretation without the use of AI.

But what are the implications of adding another element to this basic scheme, and where should this element be located? Can language-generating AI be assigned to one of these three elements? Is an LLM an extension of the interpreting subject in the form of an enhancement of human perception (see figure 10.2)?

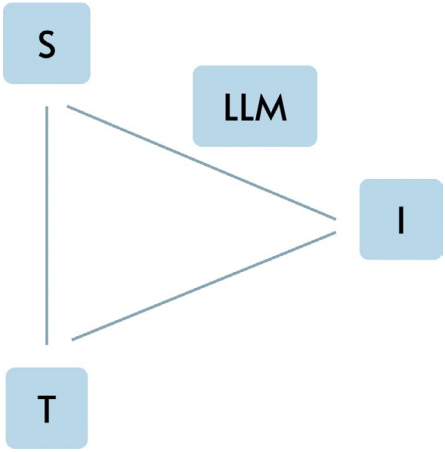

Figure 10.2: Model representation of interpretation with AI as an extension of the interpreting subject.

Is it part of the *Interpretans*, that which has already been interpreted? Or is it part of the *Interpretandum*—i.e. something that precedes interpretation and, as such, something that (also) first needs to be interpreted (see figure 10.3)?

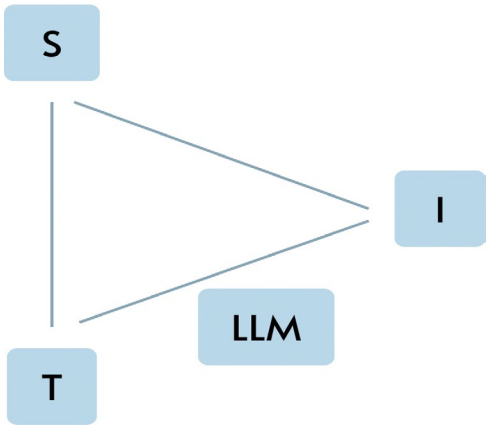

Figure 10.3: Model representation with AI as part of the text.

Can LLMs be assigned to any of the existing elements at all, or is the language model perhaps an independent element that needs to be included as an additional variable in this basic model? One could also imagine that LLMs are a foil that influences all the positions of the basic model.

Readers would thus always have a counterpart in the co-construction process of meaning that could support all reader groups equally: Experienced readers could discuss opinions, judgements, and interpretation hypotheses with the AI, while weaker readers could use the AI for simple queries and explanations. Today, AI tools already make it possible to externalize certain cognitive operations (e.g. ensuring text comprehension by clarifying terms, rhetorical devices, content summaries, etc.), which reduces the cognitive load in a complex literary understanding process. Overall, AI tools offer enormous potential for developing a literary reading attitude. The use of LLMs could thus support self-regulated reading and lead to elaborated analyses in dealing with literary source texts. The known inequality effects in the reception of literature (e.g. the social reproduction of disadvantages in schools and academic systems for multilingual persons, Dirim et al. 2018) could potentially also be counterbalanced, for instance by compensating for a lack of language skills, or by bridging cultural differences in the comprehension of motifs or metaphors, etc. A strong(er) knowledge base of literary comprehension could be developed individually with the help of appropriately trained AI tools with regard to styles, genres, and historical contexts; ideally, they could be integrated into readers' prior knowledge as well as into their linguistic, social, and cultural backgrounds. From a broad perspective on the literary reception process, such

a transhuman view of technical development could, paradoxically, lead to strengthening of the humanistic concept of education: it makes literary understanding (more) accessible to everyone. *Transhuman* in this sense means that human capabilities are expanded with the help of machines (Bostrom 2005). However, such transhuman reading or enhanced perception would focus more on promoting literary reception skills for the individual. It is less attentive to the risks associated with the collective, and thus social, dimensions of literary communication.

In contrast to such a positive view, LLM support may, after all, also lead to a less elaborate understanding of literary texts in cognitive, emotional, social, and motivational terms. On a cognitive level, users may end up even *less* encouraged to understand the source text comprehensively than they would have been through conventional methods. On an emotional level, for example, it is not a given that we would empathize with the literary world with the same intensity when we use AI as a "shortcut" to access. Many such reception processes are based on experiences of *otherness* in terms of language as well as of content. This means that if, in a literary text, something seems unfamiliar or "strange" to a reader, this might not always be a problem or an obstacle, but it may be precisely why it attracts and fascinates us in the first place. Ambiguity or otherness may be the whole point of the text. Empirical studies indicate that ambiguities within literature are often explicitly emphasized by AI tools—but without references to the actual text. LLMs, for example ChatGPT, generally say that a literary text could have several meanings, but without reference to the actual content. Inexperienced readers then often decide in favor of just a single meaning and do not relativize it again later (Führer and Nix 2023). Technological support in the development of text comprehension can certainly also lead to manipulation of readers' own literary perceptions. Ambiguities and nuances of literary texts are often too quickly converted into a coherent, "smoother" mental model for the recipients. Tendencies to promote "easy" informational reading attitudes instead of complex aesthetic ones (Rosebrock 2017) may then be reinforced alongside a lack of individual perception and in-depth aesthetic examination of the text (Führer and Nix 2023). Jannidis et al. (2025) confirm with different methods that LLMs do not (yet) achieve a "deep" understanding that combines formal aesthetics with innovative reinterpretation. So, the subjective tension between reader and text changes when literary reception becomes transhuman in the sense above. It still remains to be seen how our capacity for perspective-taking, intersubjective negotiation, and multi-perspective perceptions is negatively influenced. LLMs can, for example,

offer uncomplicated political or historical interpretations of texts that do not do justice to the context or the reader's individual experiences. But one essential reason for literary reading is to foster the exchange of different views on fundamental anthropological and social questions via a text. This could be negatively impacted through the use of LLMs, as readers relinquish complex thinking and (individual) engagement in favor of an "easier," more universal, and lastly more uniform mode of access.

AI Literacy and the Future of Literary Judgement

It is to be expected that AI tools will create potential new power structures in literary communication, as well as changes in our aesthetic judgement. It is particularly important to take into consideration the text corpus which has been used for training the respective LLMs: When using AI (for writing as well as for reading), readers are confronted with a certain "average" or "dominant" aesthetics that were inherent in the corpus and which can exhibit potential biases (Barocas et al. 2019; Manovich 2019). Our literary reading and writing capabilities, on an overall social level, may become narrower because the diversity of knowledge acquisition and processing is reduced to a uniform perspective. This is not because LLMs offer only one interpretation, but because they marginalize individual ways of reading that are not statistically relevant. So, if we rely on LLMs more and more, quantities will have a new universal aesthetic validity and become the last instance of aesthetic judgement. This affects the collective creation of literary aesthetics and our societies (see, for example Anderson et al. 2024; Elliott 2019). The goal of conventional aesthetic reading—that is, of becoming an autonomous aesthetic subject—could thus be jeopardized by the increasing use of AI tools, especially in the context of literary development. Even today, an upper secondary level pupil can articulate and stress the view that the creative human production of something new from the self should be prioritized over the *surface rationality* (a term derived from Siegmund 2020) of machine-generated content. This can be seen in the following testimonial written by a German high school student:

> The content of ChatGPT is often very superficial and repetitive, which does not provide any new insights. Thereby, this causes a very general approach to such a text. I have also noticed that incorrect information from my prompts (e.g. *goalkeeper* [original ChatGPT: *Torhüter*] instead of *doorkeeper* [original Kafka*: Türhüter*] in Franz Kafka's parable *Before the Law*) is simply adopted. In

> addition, the program always asks more questions, causing it to focus on one single aspect, and it seemed as if it didn't have a lot of background information on the text. Furthermore, when subsequently writing [my interpretation of the] text, there was a tendency to lean too heavily on the phrases the program uses. (Führer and Gerjets 2024, 13; my translation)

The pupil correctly points out that LLMs tend to systematically agree with users rather than independently generating texts or judgements (so-called "AI sycophancy"). Accordingly, we face several challenges; that LLMs influence our ideas (Führer and Nix 2023), but also that our specific language in literary interpretation is affected (Kellermann et al. 2024). This could even lead to the negation or levelling of a person's individual aesthetic tastes and judgements (or the distorting effects of biases on these). That said, this is not merely a problem if readers lack data or critical literacy in general. It is already a fact of daily life that data-tracking algorithms can influence aesthetic communities and even cultural assumptions (e.g. book recommendations that appear in readers' timelines, Elliott 2019). A corpus analysis of comments on the e-book platform Wattpad (Pianzola et al. 2020) has shown that in the case of young adult literature, the focus is more on social interaction between readers and "social bonding," whereas in the reading of literary classics the focus tends to be more on elaborating the meaning (and context) of the text. This presents the risk of (only) understanding and using literature as a confirmation, as an emotional reinforcement of one's own existing worldview or that of the particular groups or communities to which one already belongs (Führer 2023). This so-called confirmation bias is mainly useful for AI platforms and their capitalist interests since it enhances the ability to predict and replicate aesthetic taste, which commercial providers are already exploiting by using AI avatars as "virtual reading buddies" (e.g. in character.ai, readers interact with AI versions of their favorite book characters, such as Violet Sorrengail from Rebecca Yarros' 2023 *Fourth Wing*). Human authors like Jennifer Lepp (who writes under the pen name Leanne Leeds), in turn, use LLMs to increase efficiency in genre writing.

Of course, there are always two sides to the equation, and none of this need necessarily be for the worse: Anderson et al. (2024) hypothesized that the use of an LLM as a creative support tool might make users feel more creative and even broaden the range of ideas suggested by each individual user. However, it can also homogenize ideas. The researchers found that in accordance with their homogenization hypothesis, different users tended to produce less semantically distinct ideas with ChatGPT. Additionally, ChatGPT users generated a

greater number of more detailed ideas, but, in the end, felt less responsible for them (Anderson et al. 2024). Last but not least, LLM-driven text interpretation and LLM-based writing perpetuates existing literary hierarchies by prioritizing canonical or commercially successful works and established authors over emerging voices from marginalized communities. This algorithmic bias affects which authors gain visibility, which works are considered relevant, and, ultimately, which texts enter the literary canon through digital discovery. In literary contexts, authentic cultural representations are often overshadowed by commodified or stereotypical portrayals that serve commercial interests rather than genuine cultural understanding (see, for example, Noble 2018). Behind this lies the fundamental question of how individual (and responsible) aesthetic development can be promoted when AI development and use are predominantly characterized by speed, statistical distribution, and probability, while often providing little awareness of one-dimensionality, dubious content, or the complex moral and ethical dimensions of a text.

The traditional conception of literary judgement we inherited from the Enlightenment era is rooted in a strong anthropocentric worldview. There, we can clearly distinguish between identity and otherness, between nature and culture, between humans and technology. These ideas, however, are beginning to falter where such opposing relationships (and worldviews) no longer reflect reality. We now act and communicate—as is often stated as an ideal in EU digital policy—not only with "individual sovereignty, understood as the autonomy and self-determined capacity to act" (Pohle et al. 2022, 1; my translation; see also Pohle et al. 2025); rather, the more we rely on AI tools, the more we see an interweaving of people and technology with each other that requires discursive, cooperative, and often conflictual negotiations of beliefs, opinions, and interpretations. While AI itself can perform and manipulate complex and creative thinking, this has to date often led to incoherent overall judgements. What is missing is a so-called "reflective equilibrium" (Betz and Richardson 2022). This concept—developed by John Rawls in his 1971 work *A Theory of Justice*—refers to a state of moral or rational judgement in which our individual beliefs and general principles are in harmony once we have checked, adjusted, and weighed them against each other. Only humans can currently think in this way, and thus also form diverse literary appreciations and judgements. Hannah Arendt (1978) noted that the reason for diversity of opinion is a form of self-knowledge that is not based on ideal truths—or, by analogy, on the abstract perspectives generated by AI—but is determined, rather, by "joint thinking" through many available genuine alternatives. Thus,

the value of reading generic texts using LLM technology or enjoying hybrid AI-generated literature is not only a function of the access and possibilities these texts offer for individual enjoyment. Instead, these texts can make us think differently about human–machine relationships and complex socioeconomic aesthetic communities.

Facets of critical AI literacy (Long and Magerko 2020) are thus necessary to recognize and ultimately resist biases and stereotypes, but we also need an alternative to "hermeneutic" readings of AI texts *as if they were* written by embodied humans when, in reality, they are merely statistical artifacts. The problem of LLMs not having an "authentic voice" but instead calculating statistical probabilities from a huge corpus requires strategies for dealing productively with this lack of perspective in LLMs. Franco Moretti's (2013) "distant reading" might be one method of approaching such texts: perceiving patterns in the production of large quantities rather than interpreting individual texts in detail. Lev Manovich (2020) uses AI for data visualization, so that large-scale aesthetic patterns in culture (e.g. on Instagram) can be revealed that would not be visible to the human eye alone. These patterns are the result of countless individual decisions within a network. Practically speaking, one can, for example, focus on *data genealogy* (the origin and selection of the training data where we know about it) or carry out interface analyses (e.g. how are models presented, marketed, and framed?).

In a perspective from *digital humanism* (Nida-Rümelin and Weidenfeld 2022), AI can be used as an aesthetic area of friction and a productive irritation. When language models deliver an "average," this can be extremely valuable for literary writing or literary text analysis—as a negative example. Since AI usually chooses the path of least resistance (the cliché), the author or recipient immediately recognizes the "flat" parts of their own draft or thinking. Humans then write precisely against this predictability in order to create genuine literary depth and ambiguity.

A *technogenesis perspective* (humans and technology co-evolve together, Hayles 2012) suggests that we must abandon the search for pure, human authenticity. Its approach is that authenticity is redefined as the quality of a human–machine relationship. In projects such as Kenric Allado-McDowell's and GPT-3's (2020) *Pharmako-AI*, therefore, no attempt is made to make AI sound like a human being. Instead, the "machine-like nature," the strange, often illogical combination of perspectives, is understood as an aesthetic quality in its own right. The result is not "humanly authentic" but "technogenetically new."

From an even more *transhumanist perspective* (More and Vita-More 2013), the focus is no longer on creating something out of nothing (*poiesis*), but on filtering out the special from the AI-generated "mush." Humans use AI as a text quarry. The literary achievement consists in recognizing the clichés of AI, rejecting them, or recombining them in such a way that a new, individual quality emerges. Authenticity shifts from the text to the author's decision.

In the perspective of *posthumanist theories* (Braidotti 2013), one could even provoke what is unavailable or uncontrollable in aesthetic experience by prompting against the algorithm. AI can be strategically overwhelmed by extreme constraints, forcing it to "think" about aesthetic styles and areas that it rarely addresses. This creates breaks that can once again offer space for aesthetic experience and irritation.

Lastly, it should be mentioned that there are a small but growing number of examples of experimental, self-reflexive AI-literature curated by human authors which deserve special attention: *Poetological literature* offers the opportunity to reflect on the profound literary and social implications of AI aesthetics (Bajohr 2023b; Piringer 2022; both in German). This self-reflective literature shows poetically that contemporary categories of literary judgement, such as those surrounding authorship and authenticity (especially in connection with narratives of migration, racism, and classism) can be "turned off" as a result of overstimulation and manipulation through AI. Poetological AI literature thus shows, at the very least, that there is a need for new aesthetic forms of storytelling and language use in this algorithm-driven world.

Conclusions

The intersection of AI and literature is currently undergoing dynamic developments in which both continuities and differences from earlier literary reception and production are becoming apparent. Central aesthetic categories such as originality, creativity, authorship, and control over works are currently changing. The danger of aesthetic superficiality and self-affirmation must be balanced against the creative potential of human–machine hybridization and the opportunities for increased accessibility to literature. The measurement of the aesthetic field by economically active platforms makes it essential to also consider aesthetic developments from an epistemic and ethical perspective. Complete autonomy and freedom from purpose in literature cannot exist in the age of AI; rather, it must first be established through appropriate critical

reflection on the functioning of AI. Instead of running the risk of a global aesthetic uniformity, a pluralistic development of literary life should be promoted throughout society. The otherness of people and cultures as well as (legal) authorship must be preserved, while simultaneously strengthening the polyphony of literary expression and judgement. To this end, we must collectively rethink the formation of different literary communities and diverse aesthetic knowledge frames in society.

Works Cited

Allado-McDowell, Kenric and GPT-3. 2020. *Pharmako-AI.* London: Ignota Books.

Anderson, Barret R., Jash H. Shah, and Max Kreminski. 2024. "Homogenization Effects of Large Language Models on Human Creative Ideation." Preprint, arXiv, February 2, 2024. https://doi.org/10.48550/arXiv.2402.01536.

Arendt, Hannah. 1978. *Thinking*. Vol. 1 of *The Life of the Mind*. New York: Harcourt.

Bajohr, Hannes. 2023a. "Artifizielle und postartifizielle Texte: Über die Auswirkungen Künstlicher Intelligenz auf die Erwartungen an literarisches und nichtliterarisches Schreiben." *Sprache im technischen Zeitalter* 61 (245): 37–61.

Bajohr, Hannes. 2023b. *(Berlin, Miami)*. Berlin: Rohstoff.

Bajohr, Hannes. 2024. "Autorschaft und Künstliche Intelligenz." In *Handbuch KI und die Künste*, edited by Stephanie Catani, 265–280. Berlin: De Gruyter. [Also available in English: "Writing at a Distance: Notes on Authorship and Artificial Intelligence." *German Studies Review* 47 (2): 315–337.]

Bajohr, Hannes and Ann Cotten, eds. 2025. *Schreiben nach KI*. Berlin: Rohstoff.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning: Limitations and Opportunities.* Cambridge, MA: The MIT Press.

Barsalou, Lawrence W. 1999. "Perceptual Symbol Systems." *Behavioral and Brain Sciences* 22 (4): 577–660. https://doi.org/10.1017/s0140525x99002149.

Barsalou, Lawrence W. 2008. "Grounded Cognition." *Annual Review of Psychology* 59: 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639.

Barthes, Roland. [1967] 2000. "Der Tod des Autors." In *Texte zur Theorie der Autorschaft*, edited by Fotis Jannidis, Gerhard Lauer, Matias Martinez, and Simone Winko, 185–197. Stuttgart: Reclam. [Also available in English: http://www.adamabraham.info/uploads/1/1/5/9/115940383/rolandbarthes.pdf]

Bense, Max. 1998. *Ästhetik und Texttheorie.* Stuttgart: Springer.

Betz, Gregor and Kyle Richardson. 2022. "Judgment Aggregation, Discursive Dilemma and Reflective Equilibrium: Neural Language Models as Self-improving Doxastic Agents." *Frontiers in Artificial Intelligence* 5. https://doi.org/10.3389/frai.2022.900943.

Bostrom, Nick. 2005. "A History of Transhumanist Thought." *Journal of Evolution and Technology* 14 (1).

Braidotti, Rosi. 2013. *The Posthuman*. Cambridge: Polity Press.

Caldwell, Makenzie. 2023. "What Is an 'Author'? Copyright Authorship of AI Art Through a Philosophical Lens." *Houston Law Review* 61 (2): Rev. 411.

Çakir, Dîlan Canan. 2024. "Literatur ist niemals nur Kunst: Vorüberlegungen zum aktivistischen Potential von Literatur und dessen Erforschung." In *Politische Ideen in der Literatur. Von um 1800 bis zur Gegenwart*, edited by Max Roehl, 337–358. Darmstadt: wbg Academic.

Catani, Stephanie. 2024. "Künstliche Intelligenz und Kreativität." In *Handbuch KI und die Künste*, edited by Stephanie Catani, 297–305. Berlin: De Gruyter.

Choi, Christy and Francesca Annio. 2024. "Rie Kudan Wins Akutagawa Prize with ChatGPT-assisted Novel." *CNN Style*, January 19, 2024. https://edition.cnn.com/2024/01/19/style/rie-kudan-akutagawa-prize-chatgpt/index.html.

Crawford, Kate. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.

Cuskley, Christine, Rebecca Woods, and Molly Flaherty. 2024. "The Limitations of Large Language Models for Understanding Human Language and Cognition." *Open Mind 8*: 1058–1083. https://doi.org/10.1162/opmi_a_00160.

De Man, Paul. 1971. *Blindness and Insight: Essays in the Rhetoric of Contemporary Criticism*. London: Oxford University Press.

Dirim, İnci, Magdalena Knappik, and Nadja Thoma. 2018. "Sprache als Mittel der Reproduktion von Differenzordnungen." In *Heterogenität, Sprache(n) und Bildung: Eine differenz- und diskriminierungstheoretische Einführung*, edited by İnci Dirim and Paul Mecheril, 51–62. Bad Heilbrunn: Verlag Julius Klinkhardt.

Elliott, Anthony. 2019. *The Culture of AI: Everyday Life and the Digital Revolution*. New York: Routledge. https://doi.org/10.4324/9781315387185.

Fischer, Martin H. and Rolf A. Zwaan. 2008. "Embodied Language: A review of the Role of the Motor System in Language Comprehension." *The Quarterly Journal of Experimental Psychology* 61 (6): 825–850. https://doi.org/10.1080/17470210701623605.

Führer, Carolin. 2023. "Literatur im Netz: Gegenwärtige Herausforderungen ästhetischer Bildung am Beispiel von digitaler Poesie." *Medien im Deutschunterricht* 1: 1–16. https://doi.org/10.18716/ojs/midu/2023.1.3.

Führer, Carolin and Peter Gerjets. 2024. "How to Understand & Write Literature with AI? Potentiale und Risiken von KI-Tools für literarisches Lesen und Schreiben." *Medien im Deutschunterricht* 6 (1): 1–18. https://doi.org/10.18716/ojs/midu/2024.1.3.

Führer, Carolin and Lore Knapp. Forthcoming. "Preservice Teachers' Inquiry-Based Learning on AI Poetics and Authorship." *L1—Educational Studies in Language and Literature*.

Führer, Carolin and Daniel Nix. 2023. "Anschlusskommunikationen mit ChatGPT—Kann die Interaktion mit Künstlicher Intelligenz (KI) Schülerinnen und Schüler beim Verstehen literarischer Texte unterstützen?" *Leseforum* 3. https://doi.org/10.58098/lffl/2023/3/805.

Glenberg, Arthur M. 2011. "How Reading Comprehension is Embodied and why that Matters." *International Electronic Journal of Elementary Education* 4 (1): 5–18.

Gunser, Vivian Emily, Steffen Gottschling, Birgit Brucker, Sandra Richter, and Peter Gerjets. 2021. "Can Users Distinguish Narrative Texts Written by an Artificial Intelligence Writing Tool from Purely Human Text?" In *HCI International 2021—Posters*, edited by Constantine Stephanidis, Margherita Antona, and Stavroula Ntoa, 520–527. Cham: Springer. https://doi.org/10.1007/978-3-030-78635-9_67.

Gunser, Vivian Emily, Steffen Gottschling, Birgit Brucker, Sandra Richter, Dîlan Canan Çakir, and Peter Gerjets. 2022. "The Pure Poet: How Good is the Subjective Credibility and Stylistic Quality of Literary Short Texts Written with an Artificial Intelligence Tool as Compared to Texts Written by Human Authors?" In *Proceedings of the 44th Annual Meeting of the Cognitive Science Society* 44: 1744–1750. https://escholarship.org/uc/item/1wx3983m.

Hayles, Katherine. 2012. How We Think: Digital Media and Contemporary Technogenesis. Chicago: University of Chicago Press.

Jacobs, Arthur M. and Roel M. Willems. 2018. "The Fictive Brain: Neurocognitive Correlates of Engagement in Literature." *Review of General Psychology* 22 (2): 147–160. https://doi.org/10.1037/gpr0000106.

Jannidis, Fotis, Rabea Kleymann, Julian Schröter, and Heike Zinsmeister. 2025. "Do Large Language Models Understand Literature? Case Studies and Probing Experiments on German Poetry." *Journal of Computational Literary Studies* 4 (1). http://doi.org/10.48694/JCLS.4225.

Kellermann, Katharina, Carolin Führer, and Helga Gese. 2024. "Kann ChatGPT das Schreiben über Literatur sprachlich stützen? Bildungssprachliche Mittel und Textprozeduren in computergestützter schriftlicher Anschlusskommunikation." *Der Deutschunterricht* 5: 24–35.

Köbis, Nils and Luca D. Mossink. 2021. "Artificial Intelligence versus Maya Angelou: Experimental Evidence that People cannot Differentiate AI-generated from Human-written Poetry." *Computers in Human Behavior* 114: 106553. https://doi.org/10.1016/j.chb.2020.106553.

Krakauer, David C., John W. Krakauer, and Melanie Mitchell. 2025. "Large Language Models and Emergence: A Complex Systems Perspective." Preprint, arXiv, June 10, 2025. https://doi.org/10.48550/arXiv.2506.11135.

Lakoff, George and Mark Johnson. 1980. *Metaphors We Live By*. Chicago: University of Chicago Press.

Lakoff, George and Mark Johnson. 1999. *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought*. New York: Basic Books.

Long, Duri and Brian Magerko. 2020. "What is AI Literacy? Competencies and Design Considerations." In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3313831.3376727.

Lucy, Li and David Bamman. 2021. "Gender and Representation Bias in GPT-3 Generated Stories." *Proceedings of the Third Workshop on Narrative Understanding*: 48–55. https://doi.org/10.18653/v1/2021.nuse-1.5.

Magirius, Marco and Carolin Führer. 2023. "Evaluative ästhetische Rezeption und Bildung." In *Evaluative ästhetische Rezeption als Grundlage literarischen Verstehens und Lernens: Theorie und Empirie*, edited by Marco Magirius, Christel Meier, Silke Kubik, and Carolin Führer, 7–26. München: kopaed.

Manovich, Lev. 2019. *AI Aesthetics*. Moscow: Strelka Press.

Manovich, Lev. 2020. *Cultural Analytics*. Cambridge, MA: The MIT Press.

Miall, David S. and Don Kuiken. 1994. "Foregrounding, Defamiliarization, and Affect: Response to Literary Stories." *Poetics 22:* 389–407. https://doi.org/10.1016/0304-422X(94)00011-5.

Montfort, Nick and Judy Heflin. 2024. "Generated Literature." In *The Cambridge Companion to Literature in a Digital Age*, edited by Adam Hammond, 194–211. Cambridge: Cambridge University Press.

More, Max and Natasha Vita-More, eds. 2013. *The Transhumanist Reader: Classical and Contemporary Essays on the Science, Technology, and Philosophy of the Human Future. Hoboken, New Jersey: Wiley-Blackwell.*

Moretti, Franco. 2013. *Distant Reading*. London: Verso.

Nida-Rümelin, Julian and Nathalie Weidenfeld. 2022. *Digital Humanism: For a Humane Transformation of Democracy, Economy and Culture in the Digital Age*. Cham: Springer International Publishing.

Nix, Daniel and Carolin Führer. 2024. "Literarische Interaktionen mit ChatGPT—Kann der Einsatz von Künstlicher Intelligenz zur Entwicklung literarischer Lesehaltungen beitragen?" In *Literarische Texte lesen—Texte literarisch lesen*, edited by Mark-Oliver Carl, Moritz Jörgens, and Tina Schulze, 335–368. Heidelberg: Metzler.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press.

Pianzola, Federico, Simone Rebora, and Gerhard Lauer. 2020. "Wattpad as a Resource for Literary Studies: Quantitative and Qualitative Examples of the Importance of Digital Social Reading and Readers' Comments in the Margins." *PLoS One* 15 (1). https://doi.org/10.1371/journal.pone.0226708.

Piorecký, Karel and Zuzana Husárová. 2024. "The Dusk of Artificial Intelligence." In *The Culture of Neural Networks: Synthetic Literature and Art in (not only) the Czech and Slovak Context*, edited by Karel Piorecký and Zuzana Husárová, translated by Isabel Stainsby, 1–17. Prague: Karolinum Press. https://doi.org/10.14712/9788024657844.

Piringer, Jörg. 2022. *günstige intelligenz: hybride poetik und poetologie*. Klagenfurt: Ritter.

Pohle, Julia, Riccardo Nanni, and Mauro Santaniello. 2025. "Unthinking Digital Sovereignty: A Critical Reflection on Origins, Objectives, and Practices." *Policy & Internet* 16 (4): 666–671. https://doi.org/10.1002/poi3.437.

Pohle, Julia, Leo Thüer, Finn Dammann, and Jan Winkler. 2022. "Das Subjekt im politischen Diskurs zu 'digitaler Souveränität.'" In *Handbuch Digitalisierung und politische Beteiligung*, edited by Norbert Kersting, Jörg Radtke, and Sigrid Baringhorst, 1–23. Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-658-20734-4_84.

Porter, Brian and Edouard Machery. 2024. "AI-generated Poetry is Indistinguishable from Human-written Poetry and is Rated more Favorably." *Scientific Reports* 14 (1): 26133. https://doi.org/10.1038/s41598-024-76900-1.

Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: Belknap Press of Harvard University Press.

Robert, Jörg. 2024. *Freiheit der Kunst: Genealogie und Kritik der ästhetischen Autonomie.* Berlin: De Gruyter.

Rosebrock, Cornelia. 2017. "Sachtexte, literarische Texte: Zwei Lesehaltungen." *Der Deutschunterricht* 3: 2–9.

Scherer, Klaus R. and Harald G. Wallbott. 1994. "Evidence for Universality and Cultural Variation of Differential Emotion Response Patterning." *Journal of Personality and Social Psychology* 66 (2): 310–328. https://doi.org/10.1037/0022-3514.66.2.310.

Sergi, Michael. 2024. "Entangled with AI: Questioning Authorship and the Potential Impact on the Creative and Authentic Voice." Paper presented at the *ASPERA Conference 2024*, Melbourne, Australia, November 28, 2024. https://research.bond.edu.au/en/publications/entangled-with-ai-questioning-authorship-and-the-potential-impact.

Siegmund, Judith. 2020. "Affizierende Oberflächen, affektive Benutzung und algorithmische Rationalität." *Internationales Jahrbuch für Medienphilosophie* 6 (1): 195–202. https://doi.org/10.1515/jbmp-2020-0010.

Simanowski, Roberto. 2023. "'Texte aus dem Computer. Überlegungen zur Autorlosigkeit der Zukunft." In *Kollektive Autor:innenschaft—digital/analog*, edited by Michael Gamper, Anna Luhn, Nina Tolksdorf, and Paul Wolff, 239-249. Berlin: J.B. Metzler.

Simanowski, Roberto. 2025. *Sprachmaschinen: Eine Philosophie der künstlichen Intelligenz.* München: C.H. Beck.

Stickel, Gerhard. 1966. "'Computerdichtung': Zur Erzeugung von Texten mit Hilfe von datenverarbeitenden Anlagen." *Der Deutschunterricht* 18 (2): 120–225.

Vendrell Ferran, Ingrid. 2018. *Die Vielfalt der Erkenntnis: Eine Analyse des kognitiven Werts der Literatur*. Paderborn: Mentis.

Vendrell Ferran, Ingrid. 2023. "The Transformative Power of Literary Perspectives." *Journal of Aesthetic Education* 57 (3): 12–30. https://doi.org/10.5406/15437809.57.3.02.

Yarros, Rebecca. 2023. *Fourth Wing*. Shrewsbury: Red Tower Books.

About the Authors

Kristine Ask is Associate Professor at the Department of Interdisciplinary Studies of Culture at the Norwegian University of Science and Technology (NTNU) in Trondheim and Visiting Professor at School of Marketing and Communication, University of Vaasa, Finland; kristine.ask@ntnu.no.

Giosuè Baggio is Professor of Psycholinguistics at the Department of Language and Literature at the Norwegian University of Science and Technology (NTNU) in Trondheim; giosue.baggio@ntnu.no.

Andreas Bergsland is Professor in Music Technology at the Department of Music at the Norwegian University of Science and Technology (NTNU) in Trondheim; andreas.bergsland@ntnu.no.

Thomas Brandt is Professor of History at the Department of Historical Sciences at the Norwegian University of Science and Technology (NTNU) in Trondheim, Norway. His scholarly interests include the modern history of science, technology, and knowledge; thomas.brandt@ntnu.no.

Carolin Führer is Full Professor of German Literature at the Eberhard Karls University of Tübingen, Germany, with a particular focus on literature and media education in the German Department; carolin.fuehrer@uni-tuebingen.de.

Guri Hanem is Assistant Professor and Deputy head of education at the Department of Art and Media Studies at the Norwegian University of Science and Technology (NTNU) in Trondheim; guri.hanem@ntnu.no.

Aurora Hoel is Professor of Media Studies and Visual Culture at the Department of Art and Media Studies and Co-Coordinator (with Lukas Wilde) of the "AI Media" Research Network at the Norwegian University of Science and Technology (NTNU) in Trondheim; aurora.hoel@ntnu.no.

Miriam Kyselo is Professor of Philosophy at the Norwegian University of Science and Technology (NTNU). She specializes in philosophy of mind and cognition and is author of the forthcoming book *The Relational Mind* (MIT Press, 2026); kyselo@ntnu.no.

Matteo Radaelli is PhD candidate at the Department of Language and Literature at the Norwegian University of Science and Technology (NTNU) in Trondheim; matteo.radaelli@ntnu.no.

Roger A. Søraa is Professor in Science and Technology Studies (STS) at the Department of Interdisciplinary Studies of Culture at the Norwegian University of Science and Technology (NTNU) in Trondheim; roger.soraa@ntnu.no.

Regine Rørstad Torbjørnsen is PhD candidate in Philosophy at the Department of Philosophy and Religious Studies at the Norwegian University of Science and Technology (NTNU) in Trondheim. She completed her studies in Philosophy at Humboldt-Universität zu Berlin, Germany; regine.r.torbjornsen@ntnu.no.

Lukas R.A. Wilde is Professor at the Department of Art and Media Studies at the Norwegian University of Science and Technology (NTNU) in Trondheim and editor of *AI Aesthetics* with Jan-Noël Thon (Routledge, 2025); lukas.r.a.wilde@ntnu.no.

Today Artificial Intelligence (AI) is deeply embedded in everyday life, both at the personal and professional level. More than ever, there is a fundamental need to go beyond the hype and offer a nuanced reflection on AI's possibilities and limitations—and on how these technologies affect us in the social, cultural, and political contexts we live in. The most urgent challenges concern questions of human agency, creativity, and responsibility, spanning domains from science to society, politics, and the arts. Addressing these challenges requires a novel, humanities-based perspective that complements a technological understanding with foundational inquiries into what it means to be human and how we generate meaning.

This timely collection provides an interdisciplinary introduction to these questions, bringing together history, philosophy, ethics, linguistics, science and technology studies, media studies, as well as music and literature. The authors demonstrate that the humanities are not only challenged by AI but provide indispensable tools for understanding and reflecting our stance on it. Readers will thus gain both a broad methodological grounding but also nuanced insight into concrete cases, leaving them better informed and aware that their own perspective on AI is not only valid, but indispensable. The hype surrounding AI thrives on the assumption that its trajectory is already written. This book is an invitation to question that assumption—and to imagine otherwise.

"This is a timely and welcome contribution. It does not promise answers—which is appropriate, as many answers do not yet exist—but instead offers a grounded and open way into the study of AI technology from a humanities perspective. I wholeheartedly recommend this book to anybody somehow interested in or affected by this societal development; which is to say everybody."

– Inga Strümke, Associate Professor,
Department of Computer Science, NTNU

Scandinavian University Press

ISBN 978-82-94-16722-7

9 788294 167227