AI-Driven Knowledge Production:

A Study of the Impact of Artificial Intelligence on Scientific Knowledge Production

Thesis (version 3.01)

Remco Schimmel

28th of July 2026

Abstract

This dissertation examines how generative artificial intelligence affects the production of scientific knowledge and which epistemological, methodological and institutional implications arise from it.

The research is structured around an analytical framework in which three dimensions of knowledge production — the cognitive and methodological, pedagogical and institutional, and normative and political dimensions — are intersected with five archetypes of scientific practice. This 5×3 matrix constitutes the dissertation's central analytical framework. In the first phase, the five archetypes are analysed across these three dimensions on the basis of literature reviews. This produces a provisional picture of the potential influence of generative AI on different forms of scientific research.

In the second phase, this picture is tested empirically through five practice-based studies, one within each archetype of scientific practice. The cases demonstrate how generative AI actually functions in diverse research practices, ranging from experimental and historical research to interpretative and design-oriented studies, as well as a reflexive meta-case in which the research process itself is analysed.

The synthesis of these analyses shows that generative AI does not simply automate science, but reconfigures the architecture of knowledge production. Under certain conditions, AI can accelerate research processes, enable new combinations of knowledge domains and support new forms of verification. At the same time, new epistemic vulnerabilities become visible, including narrative overcoherence, semantic drift and dependency on technological infrastructures. The quality of AI-driven knowledge production therefore proves highly dependent on how the generative process is organised and supervised.

The dissertation concludes that generative AI gives rise to a hybrid research practice in which human curation and machine-assisted analysis become structurally intertwined. The reliable application of AI in science therefore requires explicit research architectures, transparency concerning AI use and new forms of methodological control.

Executive Summary
AI-Driven Knowledge Production- A Study of the Impact of Artificial Intelligence on Scientific Knowledge Production

This dissertation forms the concluding part of a pentalogy of five academic monographs. The four preceding monographs described the use of AI in the natural sciences, history, anthropology and design-oriented research. In each case, AI was used to write a scientific dissertation in accordance with the conventions of the relevant domain. The present dissertation is a meta-dissertation: it serves as an umbrella over the other monographs. It organises science by identifying four-plus-one archetypes of scientific practice and subsequently examines how the use of AI affects these forms of scientific practice. The four aforementioned monographs therefore constitute the case material for this dissertation.

Chapter 1 – Introduction: AI and Knowledge Production

This dissertation centres on the fundamental question of how artificial intelligence (AI) is changing the foundations of scientific knowledge production. Whereas computers previously functioned primarily as calculating tools, they have now become co-producers of knowledge. This development raises urgent questions concerning the ways in which AI affects processes of data collection, analysis, explanation and presentation.

To structure this complex subject matter, the chapter introduces four archetypes of knowledge production. The first archetype concerns natural science research, which seeks to discover universal laws through measurement and experimentation. The second encompasses historical research, in which the understanding of unique events is central and is pursued through source criticism and interpretation. The third comprises anthropological research, in which culture is understood through participant observation and thick description. The fourth concerns design-oriented research, in which knowledge is produced by creating solutions.

Alongside these four archetypes, an important distinction is made between generic AI — large models trained on internet data — and domain-specific AI, trained on carefully selected datasets. In addition, the meta-domain of computer science and AI development itself receives special attention because of its reflexive character, with AI simultaneously serving as an instrument and an object of research.

The analytical framework used in this dissertation comprises six impact dimensions: cognitive, methodological, pedagogical, institutional, power and ownership, and reliability and normativity. Particular attention is paid to what is termed the AI paradox: the paradoxical situation in which AI can be used productively only by those who have first been educated without it.

Chapter 2 – Research Questions

The second chapter formulates the central research questions against the background of what is described as fragmentation within contemporary academic culture. Although AI accelerates research processes, it simultaneously risks reinforcing the regime of ‘postage-stamp insights’ — small, narrowly delimited contributions — rather than promoting genuine synthesis.

The main research question is: How does the use of artificial intelligence affect the processes, institutions and epistemology of scientific knowledge production, and what scope does this create for overcoming fragmentation and promoting multidisciplinary originality?

This main question is elaborated through three sub-questions. The first examines how AI affects the cognitive and methodological dimension, specifically the ways in which knowledge is sought, organised and verified. The second addresses the pedagogical and institutional consequences, focusing on the impact on educational practices and scientific recognition. The third examines the dimension of power and normativity, with particular attention to how AI affects the distribution of power and ownership within science.

The scientific relevance of this research lies in its provision of an integrated perspective that overcomes the fragmentation of the existing literature on AI. From a societal perspective, it is relevant because AI affects both the reliability of knowledge and the education of new generations.

Chapter 3 – Methodological Rationale

This chapter explains how the research was designed. Because AI is both the subject of study and the instrument used in conducting the research, a distinctive methodological situation arises that requires a hybrid approach.

The research combines several methods: literature research, analysis of existing AI applications across different scientific domains, and reflection on the researcher’s own experiences of using AI. This combination is necessary because AI is still so new that few established research traditions exist.

The research structure is organised as a systematic comparison. Four different types of science — the natural sciences, history, anthropology and design-oriented research — are each examined in relation to three aspects: how AI changes thinking and working practices, how it affects education and institutions, and which power relations emerge. This systematic approach prevents the research from becoming lost in disconnected observations. For each scientific domain, existing literature on AI applications is used and supplemented with practical examples. In addition, three concrete cases are examined in which the researcher worked intensively with AI, in order to gain insight into how AI-driven knowledge production operates in practice.

Quality assurance is achieved by comparing different sources and perspectives, transparently documenting every step of the research, and continuously reflecting critically on the limitations of both AI and the human researcher.

Chapter 4 – Cognitive and Methodological Impact

The fourth chapter analyses how AI reconfigures the fundamental processes of knowledge production. One important shift identified concerns the changing role of researchers: from data collectors to curators of AI-generated insights. Although AI enables large-scale replication, a problematic shift is also occurring: AI can often predict very accurately what will happen — such as which patient will become ill or which protein structure will emerge — while scientists increasingly understand less about why that prediction is correct. This undermines the traditional scientific ideal of knowing not only what happens, but also why.

In terms of opportunities, AI is found to enable considerable acceleration of literature research and data processing. It also makes feasible large-scale reanalyses and replication studies that were previously inconceivable. AI creates cognitive space for interpretation by automating repetitive tasks and enables pattern recognition across disciplinary boundaries.

These opportunities are accompanied, however, by substantial risks. These include oversimplification and loss of nuance, as well as a particularly problematic phenomenon: the fabrication of bibliographic references, which may range from accurate to entirely fictitious. There is also the reproduction of biases contained in training data, while the problem of black-box opacity remains.

The evaluation of these developments shows that the tension between acceleration and depth is crucial to the future of knowledge production. AI may reinforce fragmentation by enabling the production of more papers, but it can simultaneously promote synthesis by revealing patterns. The multiplication of AI reviewers may nevertheless create echo chambers if all models have been trained on similar data.

Chapter 5 – Pedagogical and Institutional Impact

The fifth chapter examines how AI changes both learning and the institutional structures of science. A fundamental observation is that the writing process is becoming a learning environment, while traditional rituals such as dissertations and peer review are coming under pressure.

Several pedagogical opportunities are identified. AI enables the personalisation of learning and improves accessibility for students with disabilities. It can also increase the transparency of learning processes through automatic logs that document the learning trajectory.

Institutional shifts manifest themselves at several levels. Publication culture is accelerating, but may also become more superficial. Peer review can be automated, but risks losing its human dimension. In addition, the criteria governing academic careers must be recalibrated to reflect new forms of knowledge production.

Three fundamental tensions are identified that affect the core of these changes. The first concerns the communicative function of scientific texts: they are not merely vehicles for information, but also rituals that create community. The second involves the AI paradox: students who grow up with AI lack the frame of reference needed for critical evaluation. The third concerns the tension between originality and fragmentation: AI can both reinforce and overcome postage-stamp insights.

Although the Hong Kong Principles — integrity, transparency and open science — provide frameworks for these developments, new rituals for an AI era have yet to be developed.

Chapter 6 – Power, Normativity and the Nature of AI-Driven Knowledge Production

The sixth chapter focuses on a crucial question: who actually determines what counts as reliable knowledge when AI is used? The answer proves to lie not only with scientists, but also with the companies and institutions that develop and own AI systems.

Power and access play a central role because training AI is extremely expensive. Only large technology companies can afford systems such as GPT or AlphaFold. This means that smaller universities and researchers become dependent on what these companies make available. As a result, researchers do not have equal opportunities to produce knowledge.

Embedded choices within AI systems help to determine what counts as knowledge. When an AI model is trained, decisions are made about which data are included and which are excluded. A model trained predominantly on English-language texts will identify different patterns from one trained on a diversity of languages. These choices are made by developers rather than by the scientists who later use the model.

Data function as an instrument of power because high-quality datasets are scarce. Medical records, historical archives and social-media data are not equally accessible to everyone. Those who control these data partly determine which research is possible and which questions can be asked.

The black-box problem means that researchers often do not know how an AI system reaches its conclusions. Even in specialised systems such as AlphaFold, it is not always clear why the model produces a particular prediction. This makes it difficult to verify results or detect errors.

The chapter concludes that the use of AI in science is never neutral, but is always influenced by who has access to technology and data, and by the choices embedded within the systems.

Interim Assessment 1 – Taxonomies of Opportunities, Risks and Challenges

Following the treatment of the three dimension chapters, the insights are brought together in systematic taxonomies that serve as the analytical framework for examining the archetypes.

The taxonomy of opportunities comprises three main categories. The first is acceleration and scale, allowing feedback loops to be radically shortened and processes to become cyclical. The second concerns discovery and synthesis, through which relationships across disciplinary boundaries become visible. The third comprises pedagogical and institutional innovation through personalised learning and new forms of transparency.

The taxonomy of risks identifies five core problems. The first concerns black boxes and loss of transparency, which undermine reproducibility. The second comprises decontextualisation and loss of nuance, creating an appearance of universality around contextually contingent patterns. The third concerns ritualisation and loss of meaning, resulting in the erosion of science's community-forming function. The fourth involves the concentration of power, with epistemic agendas being shaped by commercial interests. The fifth concerns the fabrication of references, which threatens source attribution.

The taxonomy of challenges identifies three central problems: how can we determine whether AI-supported research is reliable when we cannot inspect the underlying processes? How can the meaningful rituals of science, such as dissertations and peer review, be preserved while AI is integrated? And how should different disciplines establish their own standards for the degree of AI-related uncertainty that is acceptable?

Chapter 7 – Natural Science Research and AI

The seventh chapter analyses how AI affects the first archetype. The natural sciences, traditionally centred on reproducible measurements and falsifiable hypotheses, exhibit the most harmonious relationship with AI because their value systems converge around empirical verification.

The opportunities AI offers natural science research are substantial. It can accelerate discovery processes in domains such as molecular biology and astronomy. The analysis of large datasets creates new possibilities in climatology and other data-intensive fields. The democratisation of knowledge through public databases improves accessibility, while advanced simulations and modelling increase methodological precision.

These opportunities are nevertheless accompanied by risks. The problem of black-box explainability persists even when predictions are accurate. Dependency on closed infrastructures is a growing concern. Bias in training data may create a misleading appearance of objectivity. High energy and resource costs also affect accessibility.

The challenges manifest themselves at three levels. Methodologically, the question arises whether accurate but inexplicable predictions are scientifically valid when the underlying mechanism is not understood. Institutionally, smaller research groups must also be given access to expensive infrastructure. Normatively, each domain must determine how much uncertainty is acceptable.

The conclusion is that AI functions primarily as an accelerator in the natural sciences, reinforcing existing values while also introducing new dependencies and normative choices that must be addressed explicitly.

Chapter 8 – Historical Research and AI

In historical scholarship, where authenticity and context are central, AI creates specific tensions concerning anachronism and canon formation. This tension forms the central thread of the analysis.

The opportunities offered by AI are considerable. Access to digital archives is greatly improved, while pattern recognition can reveal unexpected relationships. OCR and translation tools substantially increase the accessibility of sources. Education and institutional innovation become possible through the more active role archives can assume in knowledge production.

These opportunities are accompanied by significant risks. Anachronistic interpretations may arise when contemporary categories are applied to historical data. The datafication of sources threatens authenticity and context. Black-box patterns may be adopted uncritically. Canon formation through selective digitisation presents a particular problem.

The challenges centre on three key issues. Methodologically, historians must always translate AI-generated patterns back into the historical contexts in which the sources originated. Institutionally, decisions made during digitisation must be made explicit because they determine which narratives become visible. Normatively, the historian remains responsible for judging whether AI-supported analyses are historically valid.

The conclusion is that historical scholarship can use AI to improve access and pattern recognition, but that human interpretation remains essential to prevent anachronism and safeguard authenticity.

Chapter 9 – Anthropological Research and AI

The analysis of the third archetype shows that anthropology, founded on relational proximity and contextual richness, experiences the greatest friction with AI's tendency towards datafication and pattern recognition. This fundamental tension shapes the entire treatment of this archetype.

The opportunities offered by AI for anthropological research include several innovative applications. Digital ethnography enables analysis on a larger scale, while AI-assisted transcription and thematic coding can accelerate the research process. The documentation of endangered languages and rituals becomes technically feasible, while remote sensing creates new possibilities for landscape analysis and the study of mobility patterns.

The associated risks are substantial, however. The reduction of communities to data points threatens the core of anthropological work, while the loss of proximity and empathy undermines its methodological foundations. AI may impose classifications that do not correspond to how communities understand themselves — for example, by labelling ritual gestures as everyday behaviour or identifying universal patterns where local meanings differ considerably.

The challenges facing anthropological research affect the foundations of the discipline. Methodologically, AI must function as a supplement to, rather than a replacement for, interpretative work. Institutionally, it is essential that data be returned to research participants in order to preserve relational legitimacy. Normatively, consent must be guaranteed relationally and continuously.

The conclusion is that anthropology must explicitly embed AI within the values of proximity, empathy and shared responsibility in order to preserve its disciplinary integrity.

Chapter 10 – Design-Oriented Research and AI

Within this fourth archetype, AI acts as a co-designer, creating new possibilities while simultaneously raising fundamental questions concerning the explainability of artefacts and responsibility for them.

The opportunities offered by design-oriented research are impressive. AI as a co-designer can generate thousands of design variants, thereby broadening the creative horizon. Faster and less expensive simulation and prototyping make iterative development more accessible. Iterative cycles are greatly shortened, increasing efficiency. New forms of pedagogy become possible through direct collaboration between students and AI tools.

These promises are nevertheless accompanied by risks. Black-box designs are difficult to explain, which undermines the legitimacy of design choices. Accountability for design decisions may be lost when AI makes autonomous decisions. Ownership and authorship become diffuse within hybrid human–machine processes. Moreover, AI-supported optimisation may disregard moral values that have not been explicitly programmed.

The challenges facing design-oriented research centre on three core areas. Methodologically, standards of explainability must be developed to safeguard transparency. Institutionally, new forms of publication and recognition are needed that do justice to hybrid authorship. Normatively, values and ethics must be explicitly embedded in design processes.

The analysis concludes that AI can strengthen design-oriented research, provided that explainability, institutional recognition and collective accountability are systematically incorporated.

Chapter 11 – The Meta-Domain of AI & Computer Science

The meta-domain is distinctive because AI is both the instrument and the object of research — research into AI-driven knowledge development conducted using AI — which creates reflexivity and circular validation. This distinctive position makes the domain an instructive test case for the broader analysis.

The utopian scenario presents impressive possibilities. AutoML and meta-learning create self-improving systems capable of exponential development. Open platforms such as arXiv and Wikipedia promote collective ownership and democratise knowledge production. New pedagogical forms become possible in which learning about AI takes place through AI, thereby developing reflexive competencies.

The dystopian scenario, however, warns of serious dangers. Circular validation against self-defined benchmarks may render apparent progress illusory. Extreme concentrations of power among a small number of companies threaten to undermine research autonomy. The black-box character becomes so pronounced that even designers no longer understand their own systems. The fragmentation of legitimacy across different validation practices undermines scientific consensus.

The challenges facing the meta-domain are wide-ranging. Methodologically, criteria must be developed that break circular validation and safeguard external validity. Institutionally, a balance must be found between open and closed systems. Normatively, responsibility must be redefined within a recursive ecosystem in which traditional boundaries become blurred.

Important qualifications are provided by counterexamples such as Wikipedia and arXiv, which demonstrate that alternative forms of validation are possible. Older research traditions such as hermeneutics, autoethnography and Actor–Network Theory offer conceptual resources for reflexive legitimation that remain relevant in the AI era. These traditions teach, respectively, that AI results always require interpretation through hermeneutics, that self-study can be legitimate provided it is accompanied by transparent reflection through autoethnography, and that knowledge always emerges within networks of people and technologies through Actor–Network Theory.

Chapter 12 – Interim Assessment 2: Overarching Patterns and Comparison of Archetypes

Chapter 12 presents a second interim assessment. It brings together the insights from the three impact dimensions and the four archetypes of science. Comparison of these domains shows that AI is applied differently across disciplines, while several recurring tensions simultaneously become apparent. The analysis demonstrates that these tensions can be understood as three fundamental paradoxes of AI-driven knowledge production.

The first paradox concerns the relationship between acceleration and legitimacy. In virtually all the domains examined, AI increases the speed and scale of analysis. At the same time, this acceleration places under pressure precisely those mechanisms that have traditionally safeguarded scientific legitimacy. In the natural sciences, for example, AI may generate hypotheses that can be validated experimentally only at a later stage; in historical research, large-scale textual analyses may suggest patterns that are historically anachronistic; and in design-oriented research, generative AI may accelerate the creative process without the underlying design choices remaining fully transparent.

The second paradox concerns access and inequality. In many cases, AI lowers barriers to research because digital archives, analytical tools and generative systems are widely available. At the same time, new forms of dependency emerge in relation to technological infrastructure, specialised expertise and representative datasets. Existing inequalities therefore do not disappear, but are reproduced in new forms.

The third paradox concerns knowledge formation and dependency. Effective use of AI presupposes precisely the analytical skills and methodological discipline that AI systems partly take over. When researchers do not develop these skills independently, there is a risk that AI systems will be used productively while their outputs cannot be assessed with sufficient critical rigour.

This interim assessment therefore shows that AI not only creates new possibilities for knowledge production, but also introduces structural tensions. These paradoxes do not appear to be temporary transitional problems, but structural characteristics of a research practice in which human and artificial intelligence are becoming increasingly intertwined. Chapters 13 to 17 examine whether empirical experience leads to the revision or further development of these findings.

Chapter 13 – Case Study: ‘Avian Intelligence & Homing Pigeons’

This chapter analyses a natural science-style study of navigational behaviour in homing pigeons. The manuscript follows the classical architecture of an empirical dissertation, comprising a literature review, a methodological rationale, multiple datasets and a synthesis chapter in which the results are interpreted. The study concludes that pigeon navigation cannot be explained by a single mechanism, but instead involves a combination of overlapping orientation strategies.

The case shows that generative AI can play a role in different phases of the research process. Its greatest added value appears to lie in the early stages of research, such as literature exploration, hypothesis formation and the systematic examination of possible explanatory models. AI can rapidly synthesise large bodies of literature, generate alternative hypotheses and

assist in structuring research designs. This can considerably accelerate the processes of problem exploration and conceptual delimitation.

At the same time, the case demonstrates that epistemic responsibility for the ultimate research architecture remains with the human researcher. The selection of hypotheses, the interpretation of results and the integration of different explanatory strands continue to require curatorial oversight and methodological discipline. In this configuration, AI functions primarily as an instrument for analysis and synthesis, rather than as an autonomous producer of scientific knowledge.

Scientific illusionism. The final section of the chapter, however, makes one crucial fact explicit: the experiment on which the dissertation is based never actually took place. The datasets underpinning the analysis were generated synthetically using generative AI. The case therefore functions as a controlled experiment in scientific illusionism. It demonstrates that generative AI is capable of producing an apparently empirical study that appears methodologically plausible and stylistically conforms to academic conventions. The manuscript's persuasiveness rests not on spectacular claims, but precisely on methodological restraint, plausible variation within the datasets and an interpretative style that explicitly acknowledges uncertainties and alternative explanations.

This exercise reveals that the form of scientific argumentation can itself be an important source of persuasiveness. When datasets appear plausible and the argumentation conforms to established disciplinary traditions, a study may seem convincing even when its empirical basis is synthetic. The case thereby underscores the importance of transparency concerning the provenance of data and of methodological control mechanisms that extend beyond stylistic or rhetorical plausibility.

Chapter 14 – Case Study: 'Predictions of Digital Transformations'

This chapter analyses the production of a historical study of predictions concerning digital technology and their actual societal outcomes in the Netherlands between 1945 and 2025. The study reconstructs a canon of ICT developments presented in political and societal sources as potentially transformative. Using this canon, historical claims about the future are systematically compared with the actual societal developments that followed. To make this comparison analytically manageable, a 2×2 matrix is developed that combines two dimensions: the accuracy of the original prediction and the extent to which the predicted societal transformation was actually realised. This produces a typology of historical cases in which expectations and outcomes exhibit different relationships.

Generative AI played a role primarily in the exploratory and analytical phases of the research process. Language models were used to identify possible episodes for inclusion in the canon, to organise large quantities of historical and policy material, and to explore different classifications of technological claims about the future. AI was also used in the systematic application of three analytical lenses: a claim-oriented lens, a transformation lens and a causal lens examining the relationship between the two.

The case demonstrates that AI can support historical research by structuring and analysing extensive discursive corpora. At the same time, the assessment of historical adequacy remains dependent on human contextualisation and interpretative expertise. The analysis therefore

illustrates that AI adds value primarily in organising and comparing historical material, while the ultimate historiographical interpretation remains a human responsibility.

Chapter 15 – Case Study: 'Governing Within Biological Boundaries'

This chapter examines the production of a dissertation describing the influence of evolutionarily and biologically shaped factors on persistent problems within public administration. The study begins from the proposition that explanations of policy failure remain incomplete when they rely exclusively on economic, legal or religious-cultural models. A framework of ten evolutionary-biological mechanisms was therefore developed. Disregarding these mechanisms in policymaking could explain the persistence of certain problems in public administration. Empirically, the study is structured around six cases distributed across three fields of conflict and is methodologically elaborated as qualitative, abductive pattern-matching research.

Generative AI played a broad role in the early and analytical phases of the research process. AI was used for literature exploration, conceptual organisation, structuring the theoretical framework and developing interpretative analyses within an interdisciplinary corpus. The case thereby demonstrates that AI adds particular value where large quantities of dispersed literature must be brought together within a more coherent analytical framework.

At the same time, the chapter exposes a classical epistemic vulnerability of interpretative disciplines. Research involving meaning-making and narrative reconstruction is traditionally susceptible to selective interpretation, narrative coherence and confirmation bias. Generative AI can reinforce these vulnerabilities because language models optimise for plausible and consistent text and can readily elaborate implicit assumptions into an increasingly coherent narrative.

It was precisely this vulnerability that led to a methodological innovation in this case. Rather than using AI solely in a generative capacity, a second model was deployed in an adversarial and supervisory role. One model generated textual proposals and analyses, while another was used systematically to identify inconsistencies, weaknesses and alternative interpretations. This created a form of role-separated modelling in which different epistemic functions were deliberately kept distinct. The case thus demonstrates that AI can not only intensify risks within interpretative disciplines, but also enable new forms of methodological counter-checking.

Chapter 16 – Case Study: 'Polyglot Language Acquisition'

This chapter discusses the production of a design-oriented dissertation on polyglot language acquisition. The starting point of the study is the observation that language education is traditionally organised separately for each language, whereas many multilingual speakers make use of structural similarities between languages. The dissertation examines how these similarities can be used systematically within a learning architecture for the parallel acquisition of multiple languages. The study combines insights from linguistics, cognitive science and pedagogy and develops a design framework in which grammatical structures, word formation and syntactic patterns across languages are explicitly connected. Generative AI was used to develop and refine examples, exercises and illustrations of such structures.

The case shows that AI is particularly valuable in design-oriented research for elaborating and operationalising a conceptual design. Whereas traditional design processes are often constrained by the time required to develop examples and educational concepts, AI makes it

possible to translate structural language patterns rapidly into pedagogical applications. At the same time, the chapter makes clear that a generative process imposes new requirements on the conceptual stability of the design. When examples and formulations are generated on a large scale, there is a risk that terminology or pedagogical principles will gradually and subtly shift, potentially reaching the point at which the generated exercises become entirely unusable. The case makes clear that generating and broadly elaborating new concepts with AI is fundamentally different from producing artefacts with AI on a large scale and with consistency.

The case therefore demonstrates that generative AI functions in design-oriented research primarily as an instrument for concretising and elaborating an existing design structure, while the architecture of the learning process itself is designed and safeguarded by the researcher.

Chapter 17 – Case Study: 'AI-Driven Knowledge Production' (Meta-Case)

This chapter analyses the production of the present document as a reflexive case of AI-driven knowledge production. The study thereby functions as a mise en abyme: the research object — the impact of generative AI on scientific knowledge production — becomes visible in the process through which the dissertation itself was produced, the 'nurse-within-a-nurse effect' or 'Droste effect'. The analysis therefore focuses not on the content, but on the architecture of the research process within which these texts were produced.

The case shows that generative AI can accelerate academic production considerably. The first complete version of this dissertation was generated within two days during a single extended interaction with a language model and was immediately usable as an academic working document. This exceptional turnaround time is not, however, the study's most important insight. The speed of the process appears to have been explained primarily by the way in which the research process was organised.

Two forms of discipline proved crucial. Process discipline kept the generative trajectory focused on the research objective and prevented the writing process from becoming derailed by iterative rewrites or digressions. More important still was structural discipline: the systematic construction of an analytical framework, followed by its consistent application in the case chapters and in answering the empirical research questions, all within a predesigned manuscript architecture. This architecture determined the cadence of the entire document and ensured that the interaction with the language model remained substantively stable throughout an extended dialogue.

The case also demonstrates that an explicit research architecture can break circular validation. By distinguishing different epistemic roles — such as generation, analysis and verification — model output is prevented from implicitly functioning as its own evidence. Curatorial oversight of the process thus proves to be a necessary condition for reliable AI-assisted knowledge production.

In addition to analysing the production of this dissertation, the chapter identifies several implications of AI-driven knowledge production that extend beyond the individual case. The first concerns the role of architecture: language models produce locally coherent text, but do not independently safeguard the global argumentative structure of a manuscript. The second concerns curatorial oversight: because contextual loss and terminological drift can gradually erode the coherence of an argument, human curation is indispensable. The third concerns the use of literature: generative AI proved not only to be a source of risk, but also a tool for

systematically checking the correct use of key references. The fourth concerns the organisation of the research process: separating productive and evaluative roles proved more effective than combining both functions within a single model.

The chapter also identifies a series of practical lessons from the broader research programme: regeneration behaviour that may undermine carefully curated passages, model output that is often paraphrastic rather than textually exact, users being unable to rely on a language model's reading accuracy, and limitations in context windows that lead to semantic drift. In this case — the production of the present document — these problems could be mitigated to a considerable extent through process discipline and structural discipline.

Chapter 18 – Synthesis and Definitive Answers

This chapter brings together the findings from the five cases and compares them with the provisional conclusions from the second interim assessment in Chapter 12. An explicit delta approach is followed: for each empirical research question, the analysis examines the extent to which the case studies confirm, qualify or correct the earlier answer.

The analysis shows that generative AI changes the cognitive and methodological dimension of knowledge production. Across the different cases, AI proves capable of accelerating synthesis and analysis considerably, while also introducing new epistemic vulnerabilities. Language models produce persuasive and coherent narratives, while empirical uncertainties and alternative interpretations may recede into the background. At the same time, these systems enable new forms of verification, for example through the systematic checking of literature use and argumentative consistency.

The pedagogical and institutional analysis shows that AI changes the character of academic work. Across several cases, the researcher's role shifts from linear author to curator of a generative process. Research and writing thereby become less a linear production of text and more an iterative organisation of generation, analysis and validation.

In the normative and political dimension, the impact of AI proves strongly domain-specific. New dependencies arise primarily at the level of infrastructure and access to models, while many normative choices in the cases examined remain connected to traditional research practices.

On the basis of these findings, the central research question is answered as follows: generative AI changes scientific knowledge production by reconfiguring the architecture of the research process. Under certain conditions, this can accelerate research, broaden it interdisciplinarily and enrich it methodologically. These effects are not automatic, however; they depend heavily on how the generative process is organised.

The cases show that the quality of AI-driven knowledge production is determined not primarily by the capabilities of individual models, but by the research architecture within which they are deployed. Explicit role separation, curatorial oversight and transparent methodological procedures prove to be crucial conditions. The chapter concludes with a discussion of the methodological safeguards employed in the research and with recommendations for further research into discipline-specific effects of AI and robust research architectures for AI-assisted science.

Epilogue

This dissertation has examined the extent to which artificial intelligence changes knowledge production in its nature, quality and legitimacy. The five cases have shown that AI-driven knowledge production is both promising and vulnerable. It is promising because AI can accelerate processes radically, reveal patterns that would otherwise remain hidden and enable new forms of synthesis. It is vulnerable because this same power is accompanied by structural risks: black-box opacity, bias, hallucinations and the temptation of apparent plurality.

The decisive factor is consistently the human curator. Whereas AI performs the role of executor and pattern recogniser, it is the human who selects, corrects and legitimises. Without this role, the promises of AI collapse into unreliability; with it, AI can broaden the horizon of knowledge production. Transparency, methodological integrity and reasoned curation are not optional in this respect, but necessary conditions.

These findings lead to what may be understood as a ‘Socratic oath’ for the AI era. Just as physicians are bound by the principle of first doing no harm, every researcher who uses AI should swear: first, produce no false knowledge. This oath entails recognising the limits of both human and machine knowledge, transparently accounting for the choices made, and never delegating responsibility for interpretation to the technology.

The value of this oath lies in its preventive function. It prevents the power of AI from turning into intellectual arrogance and apparent efficiency from becoming more important than epistemological integrity. It thereby provides a safeguard against the erosion of scientific legitimacy.

With conscious commitment to the Socratic oath, AI-driven knowledge production can develop into a new and legitimate practice of knowing. Without such commitment, it risks becoming an instrument that undermines trust and erodes the epistemic foundations of science.

Table of Contents.

Chapter 1 – Introduction: AI and Knowledge Production

1.1 Introduction: AI and Knowledge Production

The rise of artificial intelligence poses a fundamental challenge to the way science is conducted. The role of computers in research has been a subject of reflection since the 1960s, but only with the emergence of generative systems and large-scale language models has a situation arisen in which the computer is no longer merely a computational tool or simulation machine, but a co-producer of knowledge. This makes the question of how AI is changing the epistemic core of science increasingly urgent: how we collect, analyse, explain and present data.

A problem with many discussions of AI is that they refer to 'the' impact of 'AI', as though it were a homogeneous phenomenon. In reality, both science and technology are diverse. Science encompasses different traditions and objectives; AI takes various forms, ranging from general-purpose language models to highly specialised diagnostic systems. Failing to recognise these differences creates the risk of lapsing into generalisations that obscure more than they clarify.

This chapter therefore establishes a framework that does justice to the diversity of knowledge production while remaining manageable. The distinction between four archetypes of knowledge production provides a concise yet meaningful ordering. Explicit attention is also paid to the distinction between generic and domain-specific AI, while the special status of computer science and AI development itself as a meta-domain is highlighted. Together, these elements form the foundation for the subsequent analysis of the impact of AI in this dissertation.

1.2 Four Archetypes of Knowledge Production

1.2.1 Natural Science Research

The natural sciences seek universal laws through measurement, experimentation and mathematical modelling (Popper, 1959). Reliability rests here on reproducibility and verifiable causality.

AI affects this domain in several ways. Systems for data analysis and simulation accelerate the discovery of patterns. Examples include applications in climate science and molecular biology (Jumper et al., 2021). At the same time, causality risks being replaced by correlation: algorithms identify patterns but do not always explain them. Black-box models are difficult to scrutinise, which conflicts with the classical ideal of transparency (Burrell, 2016).

1.2.2 Historical Research

Historical research is concerned with understanding unique events within their context. Source criticism and hermeneutic interpretation are essential (Gamer, 1960/2004). AI opens up new possibilities in this field. The digitisation of archives and automated text analysis make accessible corpora that were previously unmanageable (Berry, 2012). Yet there are also dangers. Automated clustering can diminish nuance. Because many AI systems are trained on contemporary language, there is a risk of anachronism: modern meanings may be projected onto the past. Systems may also fabricate

sources or quotations, so-called hallucinations, thereby threatening the core value of authenticity (Bender et al., 2021).

1.2.3 Anthropological Research

Anthropology revolves around proximity: through participant observation and 'thick description', researchers seek to understand culture from within (Geertz, 1973). AI can provide support by transcribing field notes, coding interviews and producing translations (Boellstorff, 2021). This accelerates analysis and enables cross-cultural comparison. Yet something essential is lost. The relational dimension of fieldwork — trust, presence and embodiment — cannot be automated. There is a risk of superficiality and homogenisation, as global data patterns may override local diversity (Couldry & Mejias, 2019).

1.2.4 Design-Oriented Research

Alongside explanation and understanding, there is also a form of knowledge production primarily concerned with creating solutions: design-oriented research (Simon, 1969; Schön, 1983). Success is often measured in terms of usability and effectiveness. AI has proved to be a powerful co-designer. It generates variants, optimises designs and simulates outcomes, ranging from medicines to computer chips (Dean et al., 2021). Yet this also introduces new risks. Researchers may become dependent on AI, causing their own professional expertise to erode. Designs are often opaque and difficult to justify. Moreover, because AI frequently optimises for efficiency, it risks disregarding moral and social values (Zhang & Dafoe, 2019).

1.3 Generic and Domain-Specific AI

It is crucial to distinguish between generic AI and domain-specific AI. Generic AI refers to large models trained on a wide range of internet sources. Their strength lies in their versatility: summarising, generating ideas and formulating texts. Their weakness is that these internet data are replete with noise, bias and disinformation (Bender et al., 2021). In the natural sciences, such models may provide inspiration, but they do not produce valid data. For historians, they increase the risk of anachronism and hallucinations. For anthropologists, they homogenise cultural differences. Domain-specific AI, by contrast, is trained on carefully selected datasets and focuses on a narrow field of application. Examples include AlphaFold in biology and medical image recognition. Such systems often prove more reliable than human analysis, but lack the breadth and flexibility of generic models. In history and anthropology, they mainly perform a supporting role, for example in OCR or transcription. This distinction demonstrates that 'AI' is not a uniform entity. Its impact varies according to the type of technology, the discipline and the core values of the field concerned.

1.4 The Meta-Domain: Computer Science and AI Itself

The four archetypes describe different modes of knowledge production. Yet one domain occupies a distinctive position above them: computer science, and specifically the development of AI itself. This domain is characterised by reflexivity: AI is both the object and the instrument of research. AI development accelerates itself. Models assist in designing new models, algorithms optimise algorithms, and chips are developed with the aid of AI that subsequently makes AI itself more powerful (Real et al.,

2019). This self-reinforcing cycle increases the capacity for innovation, but also raises questions about control and dependency. Errors, bias and blind spots may be amplified exponentially.

Moreover, AI development is not neutral. It has become a geopolitical arena of conflict: those who control AI also control knowledge production in other domains. AI is deployed in cyber warfare, surveillance and geopolitical competition (Kania, 2019). The meta-domain therefore affects not only the epistemic core of science, but also the wider social and political order.

1.5 Impact Dimensions and Relevance

The question of how artificial intelligence affects science can be approached from numerous perspectives: economic, technological, sociological or philosophical. For the purposes of this dissertation, six impact dimensions have been selected: the cognitive, methodological, pedagogical and institutional dimensions, power and ownership, and reliability and normativity.

This choice is not arbitrary, but rests on three considerations. First, these dimensions correspond directly to the core functions of science. Science is always a cognitive enterprise in which knowledge is sought, a methodological practice through which research is conducted, a pedagogical context in which new generations are educated, an institutional structure comprising universities, journals and peer review, a system embedded in power relations that determine access and direction, and a normative framework that establishes what counts as evidence and as responsible scientific practice. Other perspectives — such as economic effects or legal regulation — are certainly relevant, but largely derive from or overlap with these core functions.

Second, this classification enables comparison across disciplines. The natural sciences, humanities and design sciences differ considerably in their methods and objectives, but cognitive, methodological and institutional questions arise in each domain. By applying a consistent framework of dimensions, it becomes possible to see that AI affects every domain, albeit in different ways.

Third, these are precisely the dimensions around which societal and academic debates are concentrated. Discussions of hallucinations and bias concern reliability; lecturers' concerns about dissertations relate to the pedagogical dimension; questions about who controls AI concern power relations; and calls for algorithmic transparency raise methodological and normative issues. By placing these six dimensions at the centre of the analysis, this study engages not only with academic debates, but also with pressing societal concerns.

1.5.1 The Pedagogical Challenge of the AI Generation

For the current generation of researchers, artificial intelligence is primarily a new addition. The next generation of children and young people, however, is set to grow up in a world in which AI will be a taken-for-granted presence from the outset. This difference is not merely temporal, but touches the core of human development. The existing separation between human reasoning and technological support, once so fundamental to learning processes, threatens to disappear. As a result, the cognitive development of this new generation will take place in continuous interaction with systems that provide immediate answers and solutions.

This shift raises fundamental questions. If basic skills such as memorisation, arithmetic or language proficiency are systematically outsourced to AI, which capacities will remain distinctly human? And which forms of autonomy are undermined when learners no longer learn to cope with bounded knowledge, but have access from the outset to a virtually unlimited supply of information? There is a

risk that the formative value of learning — the experience of effort, making mistakes and gradually developing understanding — will be lost in an environment that continuously offers rapid and seemingly conclusive answers.

For schools, curricula and universities, this means that their role must be fundamentally reconsidered. Their task is shifting from transmitting knowledge that was difficult to obtain elsewhere towards cultivating critical faculties, metacognitive skills and a reflective attitude towards technology. The pedagogical project of the twenty-first century cannot be limited to instrumental proficiency in the use of AI tools, but must focus on forming independent thinkers who are capable of assessing, correcting and limiting this technology.

It is here that the so-called AI paradox comes into sharp relief. To use AI productively, one needs a foundation that has been formed independently of AI. Only those who have developed their own judgement and patterns of thought can use technology critically and creatively. Yet the generations now growing up lack precisely this experience of a world without AI. This makes it all the more urgent to create explicit space in primary, secondary and tertiary education for activities that are not directly mediated by AI, but in which learners have the opportunity to develop their own judgement, creativity and perseverance.

1.5.2 Cognitive Dimension

Science is first and foremost a cognitive enterprise: it revolves around seeking knowledge, recognising patterns and developing explanations. Artificial intelligence intervenes directly in these processes. Systems capable of summarising millions of texts or detecting unexpected patterns in datasets are changing the way scientists think and reason. Whereas knowledge was previously built up gradually through close reading and meticulous analysis, AI can generate an overview within moments. This also changes the relationship between detailed knowledge and synthesis: researchers become less collectors of data and more curators of AI-generated insights. The question, however, is whether this new balance leads to a deeper understanding or merely to superficial acceleration, with the attendant risk that nuance and interpretative richness will be lost.

1.5.3 Methodological Dimension

Alongside thinking about knowledge, science has always also been a practice of methods and techniques, and it is precisely here that AI intervenes profoundly. In the natural sciences, this means that experiments and simulations can be accelerated and scaled up by AI. In the humanities, AI makes it possible to search archives and corpora on an unprecedented scale. In the social sciences, automated transcription and pattern recognition are fundamentally changing the analysis of interviews and observations. In all these cases, however, the question arises whether the core methodological values of transparency, reproducibility and causality can be preserved. If AI produces results that can scarcely, if at all, be verified, the methodological basis of science risks shifting from explainability towards mere utility, a development that is both promising and dangerous.

1.5.4 Institutional Dimension

Science never takes place in a vacuum, but always within institutions: universities, research councils, journals and conferences. These institutions confer recognition and perform selection; they determine who gains access to the scientific community and who does not. AI simultaneously undermines and strengthens these institutions. On the one hand, it can accelerate peer review, broaden access to the literature and facilitate collaboration. On the other, it raises questions about authorship, originality and the value of publications that are no longer read manually by anyone. If AI takes over the writing and

assessment of articles, publication culture loses its ritual and social significance and must be legitimised anew.

1.5.5 Power and Ownership

None of these developments can ultimately be understood without considering power and ownership. The development and use of AI are concentrated among a small number of major technology companies and military or geopolitical actors. Access to AI, and therefore to the knowledge produced with it, is consequently distributed unequally. Universities and public institutions are becoming dependent on infrastructures they do not themselves own. This raises questions about sovereignty over knowledge, the ability to determine independent scientific agendas, and the vulnerability of science as a public good. AI is therefore not only an instrument of knowledge, but also an instrument of power that is reordering relations among states, companies and scientific communities.

1.5.6 Reliability and Normativity

All these issues relate to the final dimension: reliability and normativity. This concerns the core of what distinguishes science from other forms of knowledge: its orientation towards truth, evidence and accountability. Hallucinations, bias and black-box models undermine the reliability of AI-driven knowledge production. What does peer review still mean when the data and algorithms on which a conclusion rests are not transparent? How can authenticity be safeguarded in disciplines that depend on sources when AI can produce plausible but fictitious quotations? The challenge is not only technical but also normative: the scientific community must reformulate the standards of evidence and accountability that apply in an era in which knowledge is increasingly produced by hybrid human–machine systems.

1.6 Synthesis and Outlook

The four archetypes and the meta-domain establish a framework that captures the diversity of scientific knowledge production without becoming mired in an endless enumeration of methodological details. These archetypes demonstrate that science can take many forms: from the search for universal laws in the natural sciences, to the reconstruction of unique events in history, the exploration of human proximity and meaning in anthropology, and the design of solutions in engineering and design disciplines. Above these stands computer science, and in particular the development of AI itself, which has a distinct status because it changes the conditions of knowledge production across all domains simultaneously.

The selection of six impact dimensions makes it possible to analyse this diversity without losing sight of it. Cognitive, methodological, pedagogical, institutional, power-related and normative aspects touch upon the core of what makes science scientific. Distinguishing between them reveals that AI intervenes in each domain in its own way: sometimes as an accelerator, sometimes as a disruptive force, and often as both at once. What may lead to unprecedented discoveries in the natural sciences may result in anachronisms or hallucinations in historical research. What enhances innovative capacity in the design sciences may undermine the relational core of fieldwork in anthropology.

The distinction between generic and domain-specific AI reinforces this point. Generic models, trained on a multitude of internet sources, are broadly applicable but vulnerable because of their tendency

towards distortion and fabrication. Domain-specific systems, trained on carefully curated data, often prove more reliable within their niche, but lack the broad applicability that makes generic models so attractive. The impact of AI is therefore never uniform, but always depends on both the type of technology and the type of science in which it is used.

In this sense, the framework presented here can be understood as a matrix: the archetypes form the rows and the impact dimensions the columns. It is not a scheme that provides all the answers at a glance, but a structure that invites analysis. The following chapters will systematically examine, across the six dimensions, how AI is reshaping science, with the archetypes serving as illustrative semi-case material. Their discussion is based entirely on publicly available sources and demonstrates the breadth of AI's

References for Chapter 1

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic

Parrots: Can Language Models Be Too Big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. https://doi.org/10.1145/3442188.3445922

Berry, D. M. (2012). Understanding Digital Humanities. Palgrave Macmillan.

Boellstorff, T. (2021). Virtual Society: Technology, Cyberculture, and Ethnography. Princeton University Press.

Burrell, J. (2016). How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms. Big Data & Society, 3(1), 1–12. https://doi.org/10.1177/2053951715622512

Couldry, N., & Mejias, U. A. (2019). The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism. Stanford University Press.

Dean, J., Corrado, G., & Shazeer, N. (2021). The Deep Learning Revolution and Its Implications for AI. Communications of the ACM, 64(6), 58–65. https://doi.org/10.1145/3448250

Gadamer, H.-G. (2004). Truth and Method (2nd rev. ed., J. Weinsheimer & D. G. Marshall, Trans.). Continuum. (Original work published 1960).

Geertz, C. (1973). The Interpretation of Cultures. Basic Books.

Jumper, J., Evans, R., Pritzel, A., et al. (2021). Highly Accurate Protein Structure Prediction with AlphaFold. Nature, 596, 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kania, E. (2019). AI Weapons and China's Military Innovation. Texas National Security Review, 2(1), 42–53.

Popper, K. (1959). The Logic of Scientific Discovery. Hutchinson.

Real, E., Liang, C., So, D. R., & Le, Q. V. (2019). AutoML-Zero: Evolving Machine Learning Algorithms From Scratch. Proceedings of the 37th International Conference on Machine Learning, 8007–8019.

Schön, D. A. (1983). The Reflective Practitioner: How Professionals Think in Action. Basic Books.

Simon, H. A. (1969). The Sciences of the Artificial. MIT Press.

Zhang, B., & Dafoe, A. (2019). Artificial Intelligence: American Attitudes and Trends. Oxford University Pres

Chapter 2 – Research Questions

2.1 Introduction: From Framework to Questions

The previous chapter developed a framework that does justice to the diversity of science while providing a means of understanding the impact of artificial intelligence. Four archetypes of knowledge production and one meta-domain were distinguished, and six impact dimensions were identified that together touch upon the core functions of science. This has produced a structure that reveals where AI intervenes, but not yet which questions should be asked about those interventions. The next step is therefore to move from reflection to explicit formulation: which issues should guide this dissertation, and why are they particularly relevant now?

The need to formulate these questions becomes apparent from the ways in which AI is already changing scientific practice. For centuries, researchers have produced knowledge in fragments, often in the form of small, specialised contributions. This has been useful, but it also raises the question of whether science in the age of AI can still afford to remain confined to such 'postage-stamp insights'. As Ioannidis (2005) demonstrated, many published research findings may be incorrect, further calling into question the value of fragmented contributions. AI makes it possible to identify connections that exceed the scope of human perception. This opens up prospects for integration, but also confronts us with the question of how valuable specialisation remains in a context in which synthesis can be generated by machines.

Moreover, the literature review, which lies at the heart of much research, is acquiring a different character. Whereas it once took months to gain a thorough understanding of the state of knowledge within a field, algorithms can search thousands of articles and uncover patterns within minutes. This raises the question of whether the classical form of literature review will remain relevant, or whether it will be transformed into a process of evaluating and correcting AI-generated overviews.

Empirical research is also affected. Data science and machine learning make it possible to identify patterns in large datasets at a speed that was previously inconceivable. Although this appears to accelerate the pace of discovery, the emphasis simultaneously shifts from explanation towards prediction. This raises the question of how science can uphold its commitment to causality and understanding in an era in which correlations are becoming increasingly dominant.

Even the social organisation of science is under pressure. If AI can act as a peer reviewer — potentially in multiple instances, with several models assuming different roles — the system of evaluation that has underpinned academia since the Enlightenment is fundamentally revised. What does academic recognition mean when assessment by human colleagues is replaced or supplemented by machine evaluation? It is conceivable not only that AI might replace the human reviewer, but also that several systems might simultaneously assume different roles. One model could act as a critic, another as a defender, and a third as an editor weighing the respective arguments. The classical situation in which one or several peers assess an article would then be transformed into a machine simulation of debate. In such a constellation, what remains of the idea that peer review is a conversation among disciplinary colleagues whose judgements are grounded in shared traditions and human experience?

The discussion therefore also concerns the meaningfulness of academic rituals themselves. If students use AI to write dissertations and lecturers use AI to assess them, the question arises as to what purpose the writing of dissertations still serves. And if scientific articles are largely generated and peer reviewed by machines, who will continue to read them, and why? The radical scenario is one in which scientific

texts are ultimately scarcely read by humans at all. When articles are written by machines and assessed by other machines, a circular system emerges in which the communicative function of science — the exchange of ideas and the collective pursuit of truth — is effectively eroded.

Finally, there is the AI paradox: to use AI productively, one must first have been educated without it. Only then does one possess the critical faculties required to employ the instrument productively and responsibly. Yet the generations now growing up no longer have experience of a world without AI. How can they learn to exercise independent judgement in an environment in which AI has been present from the outset?

Even when AI is used critically, however, the fundamental unreliability of these systems continues to operate in the background. As Bender et al. (2021) warn, hallucinations, bias and fabricated references are not marginal phenomena, but inherent risks of large-scale language models. Lawrence (2003) further emphasises that publication pressure, as a political dynamic within academia, already produces distortions, while Fanelli (2010) demonstrates empirically that this pressure increases systematic bias in scientific research. The question is not whether these problems will occur, but how science can address them without losing its normative orientation towards truth and evidence.

2.2 Scientific Relevance of the Questions

These observations show that AI is not merely an instrument that makes science more efficient, but a force that changes the very logic of knowledge production. The questions are scientifically relevant because they concern the core of what defines science: the relationship between data and theory, between detail and synthesis, and between human and machine.

The existing literature on AI in science is often fragmented. Some studies focus on specific disciplines, such as biology or history, and show how AI accelerates or supports processes within them. Other contributions address ethical issues such as bias and privacy. What is lacking is an integrated perspective that encompasses the full breadth of science and systematically maps the diverse effects of AI. The distinction between archetypes and impact dimensions is precisely what makes such an overview possible.

The factor of acceleration is also significant. AI and data science are giving empirical research an enormous boost. Whereas researchers once needed months or years to identify patterns in complex datasets, algorithms can now produce results within days or even hours. This acceleration changes the logic of scientific production: the duration of projects, the timing of publications and even the manner in which careers are built all shift accordingly. Yet the literature often lacks systematic reflection on this phenomenon. As a result, insufficient attention is paid to the ways in which AI affects not only the content of science, but also its rhythm and pace.

It is here in particular that the relevance of this research becomes apparent as extending beyond technical applications. The issue is not only what AI can do, but how science must reinvent itself. As Kuhn (2012) made clear, science consists of paradigms, and fragmentation or turning points must always be viewed in historical perspective. Originality can no longer be defined solely by the number of publications or citations, but requires the capacity to break down barriers between disciplines and establish connections that previously remained invisible. In this context, the interpretative approaches introduced by Geertz (1973), with their emphasis on thick description, offer an alternative to fragmentation, while Gadamer's (2004) hermeneutic perspective underlines the legitimacy of humanities research beyond the confines of positivist methodology.

2.3 Societal Relevance of the Questions

The question of how artificial intelligence affects knowledge production is not solely a scientific issue. It also concerns the ways in which societies relate to knowledge, authority and education. Science has never been an enterprise confined entirely within the walls of universities. Its findings enter policy, education, healthcare and culture, and help shape how societies understand themselves. When AI fundamentally changes knowledge production, this therefore has direct societal consequences.

A first concern is the reliability of the knowledge on which citizens, policymakers and professionals rely. When AI is involved in generating or assessing scientific insights, while remaining susceptible to hallucinations, bias and opaque decision-making, the credibility of science is at stake. As Burrell (2016) explains, black-box models are fundamentally opaque and thereby undermine reliability. At a time when disinformation and distrust of experts are already widespread, an insufficiently critical use of AI may erode that trust even further.

A second societal issue concerns education. If pupils and students routinely use AI from the earliest stages of their education, their relationship with knowledge changes. The classical notion of learning as a process of practice, memory formation and gradual mastery comes under pressure when answers are immediately available. For society as a whole, this is not a peripheral matter: it concerns the kind of citizens, professionals and thinkers who will be formed in the future. This makes clear that the pedagogical dimension is not merely an internal concern of science, but a societal issue that affects the quality of democratic culture and citizenship.

There is also the question of power and ownership. AI systems are largely developed and controlled by a limited number of technology companies and major powers. This means that the infrastructure on which knowledge production increasingly depends is held in private and geopolitical hands. For a society that has traditionally regarded science as a public good, this raises questions about autonomy and sovereignty. Those who have access to powerful AI models also help determine who can participate in the production of knowledge and who cannot.

Finally, there is a cultural dimension. If scientific publications are largely produced and assessed by machines, and if students write dissertations with the aid of AI that lecturers in turn use to evaluate those dissertations, there is a danger that the symbolic value of these practices will disappear. Dissertations and articles are not merely instruments for transmitting or assessing knowledge, but also rituals that confer meaning on the status of science and education. If such rituals are further eroded because scientific texts are increasingly produced by machines for other machines, the communicative fabric of science may also begin to unravel. Articles and dissertations have traditionally functioned not only as evidence of competence or vehicles for data, but also as invitations to dialogue. If no one reads them because AI assumes the role of both writer and reader, science loses one of its most essential social practices: the collective construction of meaning through language.

As Müller (2020) emphasises in his overview of ethical and normative questions concerning AI systems, these developments have direct societal relevance that extends beyond technical considerations. For all these reasons, it is societally necessary to formulate the questions surrounding AI and knowledge production with precision. This is not a technical discussion about tools, but a broader reflection on the future of knowledge as a public value, the education of new generations and the distribution of power in a digital world.

2.4 Derivation of the Research Questions

The preceding analysis makes clear that artificial intelligence is not merely an instrument that renders existing scientific practices more efficient, but a force that recalibrates the very logic of knowledge production. The questions posed in this dissertation therefore cannot be confined to technical or disciplinary details: they must address the fundamental shifts that occur when AI enters the processes through which knowledge is sought, organised, legitimised and communicated.

An important starting point is the observation that academic culture has long been characterised by fragmentation. Researchers are frequently assessed on their ability to produce small, narrowly defined contributions that perform well primarily in quantitative terms, such as publication and citation counts. As Moosa (2018) analyses in detail, this 'publish or perish' system has led to a proliferation of what may be termed 'postage-stamp insights': findings that may be correct and carefully substantiated, but contribute little to the larger whole. As a result, subjects that require broad, multidisciplinary approaches often remain outside the field of view.

AI brings this problem into sharper relief. On the one hand, it can enormously accelerate the production of such fragments: articles can be written, assessed and published more quickly, further intensifying the publication culture. On the other hand, AI creates opportunities to break down barriers between disciplines. Algorithms that analyse corpora across disciplinary boundaries can reveal connections that would remain hidden within a fragmented system. This points towards an alternative standard of scientific originality: not the quantity of contributions, but the quality of synthesis and the imaginative capacity to build bridges where disciplines have previously failed to meet.

The author's own cases included in this dissertation — earlier research projects on polyglot learning and on biology and social governance — illustrate this tension. These are subjects that struggle to develop under a regime of postage-stamp insights because they span several disciplines and therefore do not fit easily within existing publication channels. It is precisely AI that has made it possible to overcome these barriers and explore such questions productively.

Against this background, the research questions of this dissertation are formulated. They are sufficiently broad to encompass the diverse impact of AI, while focusing explicitly on the core functions of science and the need to break down multidisciplinary barriers.

Main research question: *How does the use of artificial intelligence affect the processes, institutions and epistemology of scientific knowledge production, and what scope does it create for overcoming fragmentation and fostering multidisciplinary originality?*

Sub-questions:

1. **Cognitive and methodological** — How does AI change the ways in which knowledge is sought, organised and verified, compared with traditional forms of literature review and empirical methods, and how does this relate to the tension between fragmentation and synthesis?
2. **Pedagogical and institutional** — What consequences does AI have for educational practices, the role of dissertations and publications, and the institutions that sustain scientific recognition and selection, and to what extent do these changes reinforce or correct the regime of postage-stamp insights?
3. **Power and normativity** — How does AI affect the distribution of power and ownership within science, and what new safeguards are required to maintain reliability, transparency and ethical

standards, particularly when originality is no longer sought in small fragments but in broad syntheses?

2.5 Position of the Research Questions within this Dissertation

The formulation of the main research question and the three sub-questions does not mark the end of the introduction, but rather the beginning of the dissertation's further development. They determine the structure and direction of the analysis. The main research question can only be answered at the end, when the individual strands converge. The three sub-questions guide the intermediate stages. They correspond to three clusters of themes addressed in the successive chapters: the cognitive and methodological shifts in knowledge production, the pedagogical and institutional changes accompanying those shifts, and, finally, the questions of power and normativity arising from them. Each chapter contributes to answering one or more of the sub-questions, and their sequence reflects a progression from thought (cognitive), through education and organisation (pedagogical and institutional), to power and reliability (normative).

These questions concern not only the ways in which AI affects existing practices, but also the possibility that AI compels a reconsideration of what scientific originality means. In an academic system often dominated by fragmentation and publication pressure, AI reveals both the limitations of this regime and the possibility of breaking through it. The aim is not to deny the value of specialist contributions, but to examine how science can avoid narrowing into postage-stamp insights and how it can create space for broader perspectives and multidisciplinary connections.

The chosen approach reflects this tension. The four archetypes of science and the meta-domain of computer science function as semi-case material: they are based on the existing literature and demonstrate how AI affects the core functions of science in different ways. These archetypes show that fragmentation is not the only possible perspective: through comparison and contrast, connections become visible that often remain hidden within individual disciplines. In addition, two concrete cases are introduced: the earlier research projects on polyglot learning and on biology and social governance. They demonstrate how AI can contribute in practice to research situated precisely at the boundaries between disciplines and which therefore remains outside the purview of the traditional publication system.

This also clarifies the scope of the dissertation. It is not a technical account of algorithmic innovations, a discipline-specific contribution confined to a single field, or a policy advisory document prescribing immediate solutions. Its purpose is reflective and framework-setting: to provide an analysis that maps the diversity of scientific knowledge production, shows how AI is reshaping that diversity, and explores the scope this creates for a reorientation towards originality defined not by the quantity of publications, but by the quality of connection and synthesis.

References for Chapter 2

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. https://doi.org/10.1145/3442188.3445922

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society, 3(1), 1–12. https://doi.org/10.1177/2053951715622512

Fanelli, D. (2010). Do pressures to publish increase scientists' bias? An empirical support from US States Data. PLoS ONE, 5(4), e10271. https://doi.org/10.1371/journal.pone.0010271

Gadamer, H.-G. (2004). Wahrheit und Methode (6th ed.). Tübingen: Mohr Siebeck. (Original work published 1960).

Geertz, C. (1973). The interpretation of cultures. New York: Basic Books.

Ioannidis, J. P. A. (2005). Why most published research findings are false. PLoS Medicine, 2(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Kuhn, T. S. (2012). The structure of scientific revolutions (50th anniversary ed.). Chicago: University of Chicago Press.

Lawrence, P. A. (2003). The politics of publication. Nature, 422(6929), 259–261. https://doi.org/10.1038/422259a

Moosa, I. A. (2018). Publish or perish: Perceived benefits versus unintended consequences. Cheltenham: Edward Elgar.

Müller, V. C. (2020). Ethics of artificial intelligence and robotics. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Fall 2020 Edition). https://plato.stanford.edu/archives/fall2020/entries/ethics-ai/

Chapter 3 – Methodological Rationale

3.1 Introduction: The Methodological Challenge

Formulating precise research questions is only the first step; equally important is the question of how those questions can be answered in a methodologically responsible manner. The challenge is distinctive here because artificial intelligence is both an object of analysis and an instrument that helps make the research possible. This means that the methodological choices are not merely technical or procedural, but themselves form part of the problem under investigation.

Traditional methodological frameworks provide only partial guidance. Quantitative approaches are too narrow to encompass the full scope of the questions, while purely qualitative strategies are insufficient to capture the scale and speed of AI applications (Creswell & Creswell, 2018). Discipline-specific methods — historical, sociological or natural-scientific — are likewise insufficiently suited to the ambition of looking across disciplinary boundaries. This research must therefore position itself as a reflective and systematic exploration that combines different sources and strategies.

This hybrid position entails particular responsibilities. On the one hand, the research must avoid lapsing into unconstrained speculation; on the other, it must not allow itself to be restricted by methodological orthodoxy that fails to do justice to the nature of the questions. The central methodological task is therefore to design a framework that produces valid and reliable insights while remaining sufficiently open to accommodate the complexity of AI's impact on knowledge production.

3.2 Positioning of the Research

To clarify this position, the model developed by Saunders et al. (2019), commonly referred to as the research onion, provides a useful point of reference. This model demonstrates that research always involves choices across several layers: from its philosophical foundations, through strategy and methods, to specific techniques for data collection and analysis. Within this model, the dissertation does not follow a single, unambiguous route, but adopts a combination suited to the nature of the research questions.

Its philosophical foundation is essentially critical realist (Bhaskar, 2008). It is assumed that scientific knowledge practices have real structures and effects, but that our access to them is always mediated — and, in this case, increasingly so by AI. This implies that truth and reliability cannot simply be taken as given, but must continually be questioned and justified anew.

At the strategic level, the research adopts a reflective and systematically comparative approach. The framework of archetypes and impact dimensions serves as an analytical instrument: it makes it possible to map the impact of AI broadly without lapsing into an unwieldy collection of detailed studies. The discussion of these archetypes may be understood as semi-case analysis: it is grounded in the existing literature and demonstrates how AI affects different core functions of science (Knorr-Cetina, 1999).

Five concrete cases are also included: earlier research projects in which AI was used intensively, concerning respectively navigation behaviour in homing pigeons, the prediction of IT- success, social governance and polyglot language acquisition. These cases are not intended as design interventions or as strictly autoethnographic material, but as reflexive sources that reveal how AI helps shape the

production of knowledge in practice. They provide insight into the opportunities and limitations that arise when one actually attempts to generate new knowledge with the aid of this technology.

This combination of conceptual analysis, literature review, archetypal comparison and case material makes it possible to address the research questions in their full breadth. It also legitimises an approach that cannot be captured by any single traditional method, yet remains academically rigorous through transparency regarding the provenance of sources, triangulation of insights and explicit reflection on validity and reliability.

3.3 Research Approach

3.3.1 Research Architecture

The methodological core of this dissertation consists of a research architecture that systematically maps the impact of artificial intelligence. This architecture takes the form of a matrix connecting two axes. The first axis comprises the three impact dimensions distinguished earlier: cognitive and methodological, pedagogical and institutional, and power and normativity. The second axis comprises the four archetypes of science and the meta-domain of computer science: the natural sciences, history, anthropology, design-oriented research, and computer science/AI itself.

Crossing these two axes produces a framework in which each scientific domain is examined through the same three lenses. This prevents the analysis from becoming fragmented or impressionistic and enables comparison between the archetypes. At the same time, it does justice to the distinctive character of each archetype: questions of causality and reproducibility in the natural sciences differ fundamentally from questions of authenticity and anachronism in history, or from the relational core of fieldwork in anthropology. The matrix requires these differences to be made explicit, while its parallel structure ensures that the results can ultimately be brought together in a synthesis.

The role of the meta-domain of computer science is distinctive in this respect. Because this domain is not merely another archetype alongside the others, but also provides the technology that affects all other domains, it functions as both a mirror and a touchstone (Floridi, 2014). Whereas the natural sciences and history show how AI intervenes in existing practices, computer science reveals how AI produces and reinforces itself. This domain therefore constitutes a critical boundary condition for the analysis as a whole.

The architecture also determines the chapter structure of the dissertation. The three impact dimensions are first discussed separately, with attention to their significance and interrelationships. This is followed by the chapters centred on the archetypes. Each of these archetype chapters is systematically analysed across the three dimensions. This creates a dual perspective: first, an in-depth treatment of each dimension, followed by a broader examination of each scientific domain. The result is a layered and consistent approach that ensures both theoretical precision and practical applicability.

3.3.2 Data Collection and Analysis for Each Sub-Question

The three sub-questions formulated in Chapter 2 guide the further development of this dissertation. Each requires a specific combination of sources and analytical techniques, but all three are embedded in the research architecture that crosses the impact dimensions with the scientific archetypes. This ensures that the treatment of the sub-questions remains consistent and mutually comparable.

The first sub-question, concerning the cognitive and methodological dimension, requires insight into how AI changes the processes through which knowledge is sought, organised and verified. Data collection for this question consists primarily of a systematic literature review: academic articles and meta-studies concerning the use of AI in literature reviews, data processing and pattern recognition (Cooper, 2016). This literature is supplemented with secondary examples of AI applications in different fields, such as molecular biology and climatology. In addition, observations from the author's earlier projects are used as reflexive material to illustrate the tension between fragmentation and synthesis.

The second sub-question, which covers the pedagogical and institutional dimension, requires a broader range of sources. In addition to academic literature, policy documents and guidelines issued by universities and research organisations concerning AI in education and publishing are used (Selwyn, 2019). Experiences with dissertations and articles produced partly with AI are also included in order to gain insight into the changing role of academic rituals.

The third sub-question, which concerns power and normativity, requires a more critical selection of sources. The relevant literature addresses AI governance, ownership of models and data, and ethical reflections on reliability and transparency (Crawford, 2021). Empirical examples of concentrations of power, such as the role of major technology companies in developing generative models, are supplemented with policy documents issued by governments and international organisations. The author's own experiences of dependency on commercial tools are also included to demonstrate how these power structures operate in practice.

This approach produces, for each sub-question, a combination of literature, secondary examples and reflexive cases. The systematic structure of the matrix ensures that these sources are not used arbitrarily, but are consistently organised by dimension and archetype. The research thereby acquires both breadth and depth: breadth because it encompasses the full diversity of science, and depth because each sub-question is developed in relation to the others.

3.3.3 Evaluation Points for Each Scientific Domain

The matrix that crosses the three impact dimensions with the four archetypes and the meta-domain provides a robust structure, but a purely schematic approach would not do sufficient justice to the complexity of AI's impact. Several evaluation points have therefore been formulated that recur systematically in each scientific domain. They function as critical lenses that sharpen the analysis and ensure that essential issues are not overlooked.

The first evaluation point concerns the acceleration brought about by AI. In virtually every domain, data processing, literature reviews and even publication cycles proceed considerably faster when AI is used (Kitchin, 2014). This acceleration creates opportunities for new discoveries, but also raises questions about the value of knowledge produced in ever shorter periods of time.

The second evaluation point concerns the multiplication of voices in evaluation processes. Whereas peer review was traditionally a conversation among a small number of disciplinary colleagues, AI makes it possible to deploy several models simultaneously, each in a different role. This raises the question of whether such a machine-generated dialogue improves the quality of evaluation or instead undermines the human character of scientific dialogue.

The third evaluation point is the communicative function of scientific texts. Dissertations, articles and monographs are more than instruments for conveying data or analyses; they are rituals that confer meaning on scientific community and dialogue (Becher & Trowler, 2001). If AI can not only generate

but also evaluate texts, there is a danger that these texts will circulate among machines without being read by humans.

The fourth evaluation point concerns the tension between postage-stamp insights and synthesis. Publication culture still too often rewards small, narrowly focused contributions, whereas AI is capable of revealing patterns across disciplinary boundaries. The question is whether this results in deeper synthesis or merely accelerates fragmentation (Huutoniemi et al., 2010).

Finally, there is the AI paradox: the notion that AI can only be used critically and productively by those who have been educated in a context without it. For generations that experience AI as a taken-for-granted presence from the outset, the question arises of how they can develop the necessary critical distance.

By allowing these evaluation points to recur systematically in each scientific domain, the analysis acquires not only greater breadth but also greater depth. They ensure that the discussion of the impact dimensions does not remain confined to general observations, but is consistently tested against issues that touch the core of scientific practice.

3.4 Validity and Reliability

A methodological rationale cannot be limited to outlining an architecture and developing plans for analysis. Equally important is the question of how the quality of this research can be assured. Validity and reliability cannot be taken for granted here, because the use of artificial intelligence as both a research instrument and an object of research introduces new uncertainties.

The validity of this research rests on the extent to which it succeeds in revealing the core functions of knowledge production in the context of AI. This requires not only an accurate representation of the state of the literature, but also a considered application of the impact dimensions and evaluation points to the different archetypes. Internal validity is strengthened by consistently applying the same lenses across all scientific domains, thereby enabling comparison and synthesis. External validity is safeguarded by the diversity of the archetypes (Maxwell, 2012).

The reliability of this research requires particular attention because AI systems themselves are susceptible to hallucinations, bias and opaque decision-making (Bender et al., 2021). The use of AI as an instrument for literature searches or text production is therefore always accounted for transparently: sources are explicitly identified, generated passages are critically assessed against the existing literature, and fabricated references are systematically excluded. Where AI played a role in earlier case projects, this is not concealed but described explicitly, including the limitations it introduced.

Reliability is further strengthened through triangulation: insights derived from the literature review, the archetypes and the case material are placed alongside one another and critically weighed against each other. An additional measure is the use of the evaluation points formulated earlier as quality criteria.

Finally, there is the issue of reflexivity. Because the researcher has personal experience of AI-driven knowledge production, there is a continuing risk of blind spots or confirmation bias. This is addressed by presenting these experiences not as neutral data, but as reflexive material explicitly embedded within the broader context of the literature and archetypal analysis. The aim is not to generalise from

a subjective experience, but to use that experience as a prism through which broader patterns can be discerned.

3.5 Concepts and Definitions

An analysis extending across the full breadth of scientific knowledge production requires a carefully constructed conceptual framework.

Artificial intelligence (AI). In this dissertation, AI is understood as a broad family of systems that perform a variety of tasks: pattern recognition, simulation, speech and image processing, and text generation. This family includes machine learning, deep learning, natural language processing, computer vision and generative models (Russell & Norvig, 2021). This diversity is crucial because the impact differs across scientific domains: simulation in the natural sciences, digital archives in the humanities, transcription tools in anthropology, optimisation in design-oriented disciplines, and self-development in computer science. This chapter focuses primarily on this technical and functional diversity; its epistemic consequences are elaborated in later chapters.

Knowledge production. In this dissertation, knowledge production refers to the totality of processes through which new insights are generated, validated and disseminated. It encompasses not only data collection and analysis, but also the institutional and communicative practices that determine which knowledge is recognised as legitimate (Knorr-Cetina, 1999).

Originality. Originality does not refer primarily to defining a niche or producing incremental contributions, but to the capacity to connect disciplines and perspectives. Breaking down barriers and revealing relationships that remain hidden within a fragmented system constitute the core of originality (Huutoniemi et al., 2010).

Postage-stamp insights. This term refers to the phenomenon whereby researchers build careers by publishing small, tightly delimited contributions whose value derives primarily from the logic of publication pressure and citation indicators (Becher & Trowler, 2001). Such contributions are often correct and carefully produced, but make only a limited contribution to broader coherence or socially meaningful knowledge.

Synthesis. Synthesis is defined here as the capacity to bring together insights from different disciplines or research traditions into a whole whose value exceeds the sum of its parts (Repko & Szostak, 2021). Whereas multidisciplinarity often remains a matter of juxtaposition, synthesis refers to integration and meaningful connection.

Multidisciplinarity. Multidisciplinarity refers to the involvement of multiple disciplines in addressing a question, but not in the minimal sense of mere juxtaposition. It is understood here as a step towards integration and synthesis (Frodeman, 2017).

3.6 Synthesis and Outlook

The preceding sections have established the methodological foundation of this dissertation. First, the distinctive position of this research was clarified: reflective in nature, embedded in a critical analysis of the role of AI in scientific knowledge production, and informed by literature and case material. The research architecture was then developed as a matrix crossing dimensions and archetypes. This was followed by a concrete account of how the sub-questions are answered, using evaluation points as critical lenses. Validity and reliability are safeguarded through transparency, triangulation and reflexivity. Finally, the key concepts underpinning the analysis were defined.

The remainder of this dissertation first examines the three impact dimensions separately (Chapters 4–6), followed by the archetypes (Chapters 7–11), a synthesis (Chapter 12) and five cases (Chapters 13–17). Chapter 18 concludes by answering the main research question.

References for Chapter 3

Becher, T., & Trowler, P. (2001). Academic tribes and territories: Intellectual enquiry and the culture of disciplines (2nd ed.). Open University Press.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. https://doi.org/10.1145/3442188.3445922 \.

Bhaskar, R. (2008). A realist theory of science (rev. ed.). Routledge.

Cooper, H. (2016). Research synthesis and meta-analysis: A step-by-step approach (5th ed.). Sage

Crawford, K. (2021). Atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press.

Creswell, J. W., & Creswell, J. D. (2018). Research design: Qualitative, quantitative, and mixed methods approaches (5th ed.). Sage.

Floridi, L. (2014). The fourth revolution: How the infosphere is reshaping human reality. Oxford University Press.

Frodeman, R. (2017). Sustainable knowledge: A theory of interdisciplinarity. Palgrave Macmillan.

Huutoniemi, K., Klein, J. T., Bruun, H., & Hukkinen, J. (2010). Analyzing interdisciplinarity: Typology and indicators. Research Policy, 39(1), 79–88. https://doi.org/10.1016/j.respol.2009.09.011

Kitchin, R. (2014). The data revolution: Big data, open data, data infrastructures and their consequences. Sage.

Knorr-Cetina, K. (1999). Epistemic cultures: How the sciences make knowledge. Harvard University Press.

Maxwell, J. A. (2012). A realist approach for qualitative research. Sage.

Pickstone, J. V. (2000). Ways of knowing: A new history of science, technology, and medicine. University of Chicago Press.

Repko, A. F., & Szostak, R. (2021). Interdisciplinary research: Process and theory (4th ed.). Sage.

Russell, S., & Norvig, P. (2021). Artificial intelligence: A modern approach (4th ed.). Pearson.

Saunders, M., Lewis, P., & Thornhill, A. (2019). Research methods for business students (8th ed.). Pearson.

Selwyn, N. (2019). Should robots replace teachers? AI and the future of education. Polity Press.

Chapter 4 – Cognitive and Methodological Impact

4.1 Introduction

The first sub-question of this dissertation is: *How does artificial intelligence affect the core cognitive and methodological processes of knowledge production?* This question concerns the foundations of scientific work: the ways in which data are collected, analysed and interpreted, and how these processes shape hypothesis formulation, testing and validation. This chapter focuses on the cognitive and methodological dimension. It examines how AI is redefining knowledge practices, the opportunities this creates for acceleration and innovation, and the risks arising from bias, oversimplification and loss of nuance.

4.2 AI and the Redefinition of Knowledge Practices

Within a short period of time, artificial intelligence has developed from an auxiliary tool into a structural element of the scientific process. This is most visible in the natural sciences. Erduran (2024) shows how AI models such as AlphaFold have transformed biology: rather than researchers first formulating a hypothesis and then testing it, patterns generated by the machine become the starting point for further interpretation. This represents a shift in the cognitive logic of research, as algorithmic predictions become part of theory formation.

A similar reordering is taking place in other disciplines. Fabiano (2024) discusses the use of AI in systematic literature reviews. This involves more than merely accelerating a time-consuming process. What changes is the nature of the question being asked. Whereas a review has traditionally been intended to provide an overview of what is known, AI can identify patterns and gaps on a scale that previously remained invisible. The methodological function of the review thereby changes: from a summary of the state of knowledge to an algorithmically generated meta-overview.

This shift is also apparent in bibliometric applications. Saiednia et al. (2024) show how AI can be used to map publication trends and citation networks. The methodological question consequently shifts from the reliability of individual studies to the reliability of the algorithms that organise those studies. Bernard (2025) makes this concrete in his discussion of Elicit, an AI tool that supports systematic reviews. He emphasises that the selection criteria applied by the algorithm are often less transparent than traditional search strategies, calling the reliability of the results into question.

4.3 Opportunities and Possibilities

The cognitive and methodological benefits of AI appear evident. Hao et al. (2024) demonstrate empirically that researchers who use AI often achieve greater impact. This is because they establish connections between datasets and disciplines more rapidly. Methodologically, this amounts to a radical broadening of scope: AI makes it possible to discover patterns that remain hidden within individual disciplines.

A second opportunity concerns the replication problem. Ioannidis (2005) argued forcefully that a substantial proportion of published findings are probably incorrect, partly because of selective reporting and methodological shortcomings. AI can help address this problem by enabling large-scale

reanalysis. The Open Science Collaboration (2015) demonstrated how low reproducibility was in psychology. Whereas such studies traditionally take years to complete, AI can accelerate and scale up reanalyses. This opens up the possibility of collective quality control that was previously unattainable.

AI also creates cognitive space. Fabiano (2024) emphasises that automating repetitive elements of systematic reviews enables researchers to focus on interpretation and theory formation. AI therefore contributes not only to the speed of research, but also to a revaluation of the human role: critically assessing relevance and synthesising insights.

4.4 Challenges and Risks

These opportunities are accompanied by considerable risks. Messeri and Crockett (2024) warn that although AI accelerates knowledge production, it may diminish the depth of scientific reflection. Researchers 'do more, but learn less': intensive engagement with sources and data gives way to algorithmic intermediate steps, weakening their cognitive connection with the material.

A second risk concerns oversimplification. Recent evaluations of LLM systems show that they often reduce scientific texts to overly simplistic summaries in which crucial details disappear (LiveScience, 2025). What initially appears efficient may result in a loss of methodological precision. The danger is that researchers draw conclusions from algorithmic summaries without fully understanding the underlying complexity.

A third risk concerns the fabrication of bibliographic references. Generative AI models can produce references ranging from entirely accurate to wholly fictitious, a phenomenon that creates fundamental problems for scientific reliability. This problem manifests itself across a spectrum. At one end, models may generate entirely correct references when certain standard works occur very frequently in their training data. At the other, they produce completely invented article titles attributed to non-existent authors. Between these extremes lies a grey area of partially correct references: genuine authors' names linked to fictitious publications, existing journals containing invented articles, correct bibliographic details that nevertheless refer to the wrong pages or sections, or accurate titles accompanied by incorrect publication years.

The cause lies in the nature of language models themselves. These systems learn to recognise patterns across millions of text corpora and thereby develop the ability to predict what a bibliographic reference should look like: which styles of author names are conventional, how journal titles are formatted, and which volume numbers sound plausible. For the model, however, references are not links to actual sources, but merely statistically probable textual patterns. The system has no access to actual bibliographic databases and therefore cannot verify whether a constructed reference genuinely exists or points to the correct location. Relatively little systematic research is available on the precise proportions of correct, partially correct and wholly fabricated references. Existing studies suggest that the problem is substantial, but its scale varies considerably depending on the particular model, prompt and field of study. This lack of systematic knowledge makes the risk even more problematic: researchers cannot estimate how likely errors are and must therefore verify every AI-generated reference individually.

The problem is particularly insidious because partially correct references often appear convincingly accurate and are difficult to distinguish from genuine sources without systematic checking. For researchers using AI in literature reviews, this creates a structural threat to source attribution, a core practice underpinning scientific legitimacy.

There is also the structural issue of bias. Kahneman and Tversky (1979) demonstrated how heuristics produce systematic distortions in human decision-making. Nickerson (1998) showed that confirmation bias is ubiquitous. Fanelli (2010) demonstrated that positive findings are published more frequently than negative ones. AI risks reproducing or even amplifying all these biases because models are trained on existing data in which such distortions are already embedded. Methodologically, this means that AI may create an illusion of objectivity while in fact reproducing existing biases.

4.5 Evaluation Points

The cognitive and methodological impact of AI can be assessed most clearly by systematically considering several recurring evaluation points. These points function as lenses that reveal where opportunities turn into risks.

The first lens concerns the acceleration of research. Data processing and literature searches that previously took weeks or months can be completed by AI within days or even hours. In the natural sciences, this offers tangible benefits for research involving enormous datasets, as in genomics. At the same time, there is a danger that speed will be mistaken for quality. As the pressure to publish increases, an accelerated process may lead to superficial interpretation. The question is therefore whether AI contributes to greater depth or instead diminishes scientific rigour.

The second lens concerns the tension between fragmentation and synthesis. Publication culture continues to reward small, incremental contributions, the so-called postage-stamp insights. AI may reinforce this tendency by making it easier to generate a continuous succession of narrowly defined papers. Yet the same instrument can also reveal patterns and connections that invite synthesis. An AI-supported bibliometric analysis might, for example, identify an unexpected relationship between climate research and epidemiology. The question is whether scientists will use such connections to achieve integration, or whether institutional pressures will compel them to divide these patterns into new niches.

The third lens concerns the multiplication of reviewers. AI makes it possible for several models to act simultaneously as virtual peer reviewers, each assuming a different role: critical reader, editor or methodological adviser. This appears to enrich quality control, but raises the question of whether these voices are genuinely independent. If all the models have been trained on largely the same corpus, the result may be echoes rather than plurality. For the cognitive and methodological dimension, this means that an apparent expansion of checks and balances may also reinforce existing biases.

Taken together, these evaluation points make clear that the impact of AI cannot be captured through simple oppositions between opportunity and threat. It consistently gives rise to tensions in which the direction taken by science will determine the value of AI.

4.6 Conclusion

The first sub-question was: *How does artificial intelligence affect the core cognitive and methodological processes of knowledge production*? The analysis in this chapter shows that AI fundamentally reconfigures these processes. On the one hand, the technology expands the possibilities: acceleration makes large-scale replication feasible, synthesis becomes more attainable as patterns across disciplines are revealed, and virtual reviewers appear to enable new forms of quality control.

On the other hand, these same developments entail risks. Acceleration may lead to superficiality; under the existing publication culture, synthesis may give way to further fragmentation; and the multiplication of reviewers may prove illusory when all voices share the same training biases. What becomes visible at the cognitive and methodological level is therefore not linear progress, but a set of tensions that will shape the future of science.

The provisional answer to the first sub-question is consequently a nuanced one: artificial intelligence considerably expands the possibilities for knowledge production, but their value depends entirely on the extent to which scientists recognise these evaluation points and counterbalance them with methodological rigour. Only when acceleration is accompanied by greater depth, fragmentation is transformed into synthesis, and plurality of voices is genuinely critical in character can AI contribute productively to cognitive and methodological progress.

References for chapter 4.

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. Cureus, 15(2), e35179. https://doi.org/10.7759/cureus.35179

Bernard, C. (2025). Using AI for systematic review: the example of Elicit. BMC Medical Research Methodology, 25(1), 1–10. https://doi.org/10.1186/s12874-025-02528-y

Borji, A. (2023). A categorical archive of ChatGPT failures. arXiv preprint, arXiv:2302.03494.

Erduran, S. (2024). The impact of artificial intelligence on scientific practices. International Journal of Science Education, 46(3), 325–343. https://doi.org/10.1080/09500693.2024.2306604

Fabiano, C. (2024). How to optimize the systematic review process using AI tools. JCPP Advances, 4(1), e12234. https://doi.org/10.1002/jcv2.12234

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. PLoS ONE, 5(4), e10068. https://doi.org/10.1371/journal.pone.0010068

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. npj Digital Medicine, 6(1), 75. https://doi.org/10.1038/s41746-023-00819-6

Hao, D., Li, Y., & Wang, X. (2024). AI expands scientists' impact but contracts science's focus. arXiv preprint. https://arxiv.org/abs/2412.07727

Ioannidis, J. P. A. (2005). Why most published research findings are false. PLoS Medicine, 2(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47(2), 263–291. https://doi.org/10.2307/1914185

LiveScience. (2025, January 15). AI chatbots oversimplify scientific studies and gloss over critical details — the newest models are especially guilty. https://www.livescience.com/technology/artificial-intelligence/ai-chatbots-oversimplify-scientific-studies-and-gloss-over-critical-details-the-newest-models-are-especially-guilty

Messeri, L., & Crockett, Z. (2024, March 7). Doing more, learning less: The risks of AI in research. Yale News. https://news.yale.edu/2024/03/07/doing-more-learning-less-risks-ai-research

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology, 2(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716. https://doi.org/10.1126/science.aac4716

Chapter 5 – Pedagogical and Institutional Impact

5.1 Introduction

The second sub-question of this dissertation is: *How does artificial intelligence affect the pedagogical and institutional practices of knowledge production?* This question concerns two closely interconnected domains. On the one hand, it addresses the pedagogical dimension: the education of students and researchers within an academic culture increasingly permeated by AI. On the other hand, it concerns the institutional dimension: the structures of recognition, publication and peer review that shape scientific practice. This chapter examines how AI creates opportunities for innovation and inclusivity in both domains, while also generating risks of superficiality, dependency and loss of legitimacy.

5.2 AI in Education and Training

The idea that AI could fundamentally transform education has received considerable attention in recent years. Luckin et al. (2016) emphasise AI's potential to personalise learning: systems can adapt to the level, pace and interests of individual students. AI could thereby disrupt traditional pedagogical models and create greater scope for differentiation. Holmes et al. (2021) identify a similar promise in their UNESCO report, but add that the technology may simultaneously create new forms of inequality. Not every student, after all, has equal access to high-quality digital infrastructure.

Critical perspectives emphasise that the use of AI in education often fits within a broader neoliberal logic. Knox (2020) shows that AI systems are not neutral: they privilege measurability and efficiency, thereby reducing education to a series of outputs that can readily be recorded but do not necessarily encompass the full breadth of learning. Selwyn (2019) makes a related argument, maintaining that AI in education is not only a pedagogical issue, but also a political and ethical one. The question is not merely what students learn, but also who determines which forms of knowledge and which skills are considered relevant.

The pedagogical dimension also relates to the AI paradox. Younger generations are growing up in a context in which AI is a taken-for-granted presence, but may consequently lack the critical distance required to recognise the technology's limitations and biases. Students who have never written a dissertation without AI may find it difficult to distinguish between their own work and material produced by a machine. This raises questions about the development of academic judgement and the transmission of research skills.

5.3 AI and Academic Institutions

Not only education, but also the institutions of science are undergoing change under the influence of AI. Publication culture provides a first example. Bastow, Dunleavy and Tinkler (2014) show that the social sciences have increasingly come to rely on impact metrics that accord greater importance to quantitative visibility than to qualitative content. In such a context, AI can considerably accelerate the production of articles and reports, but may also contribute to the proliferation of what were previously characterised as postage-stamp insights.

The peer review system, traditionally at the heart of scientific quality assurance, is likewise under pressure. Biagioli (2002) reminds us that, historically, peer review has not merely functioned as a mechanism of quality control, but also as an instrument of power determining which voices gain access to the scientific community. AI risks reinforcing this power dynamic by automating evaluation processes. Müller-Birn et al. (2020) describe how AI tools are already being used to screen manuscripts and support reviewers. This may increase efficiency, but it also raises fundamental questions of legitimacy: who evaluates the evaluator when the evaluator is itself an algorithm?

A second institutional shift concerns the criteria governing academic careers. In their Hong Kong Principles, Moher et al. (2020) called for a reorientation towards integrity, transparency and open science. AI could act as a catalyst in this respect: it can promote reuse and open data and create greater scope for collaboration. Yet dependency on commercial tools also entails a risk. When access to advanced AI systems is distributed unequally, this may reinforce existing power relations and create new forms of exclusion.

5.4 Opportunities and Possibilities

When the pedagogical and institutional role of AI is assessed in the spirit of the Hong Kong Principles, it becomes apparent that AI could potentially contribute to the integrity and openness of science. Pedagogically, AI can help students make their learning processes more transparent. Automatic logs of writing and search activities, for example, can reveal which steps were taken during a literature review and which choices were made along the way. This would enable lecturers to assess not only students' final results, but also the processes through which they were achieved, thereby embedding integrity within education itself.

At the institutional level, AI can advance ambitions concerning open science. The automatic annotation and indexing of datasets make it easier to render research findings reusable. Translation and summarisation tools can make knowledge accessible to a broader public, bringing the ideal of societal impact closer to realisation (Holmes et al., 2021). AI can also contribute to transparency in peer review, for example by detecting textual overlap or irregularities in data analysis. In this way, AI can relieve human reviewers of routine investigative work and create scope for deeper substantive assessment.

There is also an inclusivity perspective. Luckin et al. (2016) emphasise that AI in education concerns not only efficiency, but also accessibility. For students from diverse backgrounds or with disabilities, AI can lower barriers by automatically adapting, reading aloud or translating texts. This aligns with the pursuit of equitable and inclusive learning environments.

5.5 Challenges and Risks

These opportunities are accompanied, however, by substantial risks that can likewise be understood in light of the Hong Kong Principles. Transparency and integrity cannot be taken for granted when AI is used. On the contrary, AI may create an illusion of transparency. An automatically generated report on reused data may give the impression that open-science principles have been observed, while the underlying dataset is incomplete or poorly documented. This creates a paradox: the more processes AI automates, the less clear it becomes whether researchers themselves genuinely understand and account for what is taking place.

Moreover, the use of AI may reinforce existing inequalities within academia. Access to powerful AI models often depends on commercial providers. Universities and researchers with limited resources may consequently be excluded from the latest tools, creating a divide between institutions with and without technological infrastructure. Instead of open science, a new regime of closed science may emerge, in which the concentration of power among technology companies determines the rules of the game (Crawford, 2021).

There are also pedagogical risks. When students rely too heavily on AI-generated texts, the learning process may be reduced to the consumption of outputs. Selwyn (2019) and Knox (2020) warn that AI-supported education is too often organised around measurability and standardisation, with the result that the critical and creative dimensions of academic education are lost. The AI paradox emerges once again: students who have never worked without AI may lack the frame of reference needed to recognise the technology's limitations.

5.6 Evaluation Points

The pedagogical and institutional impact of artificial intelligence cannot be understood without addressing three underlying tensions. These evaluation points are not isolated observations, but critical lenses that reveal how AI affects the foundations of academic culture.

The communicative function of scientific texts. Scientific texts have never been merely vehicles for information. They function as rituals that hold the academic community together: writing and reading articles, dissertations and monographs is also a process of socialisation through which students and researchers learn what counts as evidence, which styles of argumentation are legitimate, and how knowledge is embedded within broader traditions (Becher & Trowler, 2001). When AI takes over both the production and the assessment of texts, this ritual function risks being hollowed out. If articles are written primarily by algorithms and subsequently assessed and classified by other algorithms, a circulation emerges that sidelines human beings. The question is then not whether texts continue to be produced, but whether they still function as instruments of academic dialogue and education.

Originality versus postage-stamp insights. A second tension concerns originality and fragmentation. In principle, AI can contribute to synthesis: bibliometric analyses and literature tools can reveal patterns that researchers may connect across disciplinary boundaries. Yet the institutional logic of 'publish or perish' operates in the opposite direction. Bastow, Dunleavy and Tinkler (2014) showed how systems of career advancement and recognition have become increasingly tied to metrics such as impact factors and the h-index. Within such a regime, AI may instead function as a production machine for small, narrowly delimited contributions that perpetuate the system. Whether AI genuinely promotes originality therefore depends not on the technology itself, but on the institutional context in which it is deployed. Only when universities and funding bodies, in the spirit of the Hong Kong Principles (Moher et al., 2020), place greater emphasis on integrity and collaboration than on quantity can AI fulfil its potential as an instrument of synthesis.

The AI paradox. The most pressing evaluation point in this chapter is the AI paradox: the idea that productive use of AI requires one to have been educated in an academic context without it. Students who rely on AI systems from the beginning of their education often lack experience of the laborious processes of literature searching, source criticism and constructing arguments. Yet precisely this experience is needed to assess what AI omits or distorts. A similar dynamic operates institutionally: writing a dissertation or undergoing peer review has traditionally functioned as a rite of passage

through which researchers internalise not only knowledge, but also academic norms and values. When AI largely assumes these tasks, academic culture loses a crucial formative function.

The paradox becomes even sharper when its intergenerational dimension is considered. Researchers educated without AI possess a frame of reference from which to assess the technology's limitations critically. Younger cohorts, who experience AI as a taken-for-granted presence from the outset, risk becoming dependent on an instrument they neither fully understand nor sufficiently question. This may create tensions within the academic community: who is regarded as authoritative, who as excessively dependent, and who determines the standard of critical distance? The AI paradox is therefore not merely a pedagogical issue, but a structural problem for the future of academic legitimacy. It compels us to consider new forms of AI literacy and metacognitive training, as well as institutional safeguards that preserve rituals of critical formation.

5.7 Conclusion

The second sub-question was: *How does AI affect the pedagogical and institutional practices of knowledge production?* The analysis shows that AI presents both opportunities and risks in both areas. Pedagogically, AI can make learning more inclusive and personalised, but may also lead to a loss of critical formation and to new forms of surveillance. Institutionally, AI can make publications more transparent and peer review more efficient, but may also reinforce the fragmentation and instrumentalisation of knowledge.

The provisional answer to this second sub-question is therefore that AI not only accelerates pedagogical and institutional practices of knowledge production and makes them more efficient, but also undermines the conditions from which these practices derive their legitimacy. Only if texts retain their communicative function, originality takes precedence over fragmentation, and the AI paradox is explicitly recognised and addressed can AI contribute productively to an academic culture that is both conducive to learning and legitimate.

5.8 New Rituals in an AI Era

The analysis thus far has primarily emphasised how AI places existing pedagogical and institutional rituals under pressure. Dissertations, peer review and scientific publications lose their traditional meaning when production and assessment are largely taken over by algorithms. Yet it would be too simplistic to regard this solely as a loss. The history of academia shows that rituals continually change. Peer review, now a cornerstone of legitimacy, emerged only in the twentieth century from a combination of book censorship and collegial oversight (Biagioli, 2002). Likewise, doctoral ceremonies were originally religious in character and acquired their modern academic form only later. Rituals are therefore not fixed entities, but dynamic practices that adapt to changing circumstances.

Against this background, AI can be viewed not only as a threat, but also as a catalyst for new rituals. One possibility is the emergence of collaborative transparency. Instead of a finished text presented by a single author, the path towards publication could be made visible in a log documenting human choices and AI interventions. The ritual would then shift from the singular final product to the process of collaborative creation.

A second possibility is community review. Whereas peer review traditionally takes place anonymously and behind closed doors, AI could be used to enable larger groups of researchers to annotate, analyse

and discuss a text. Human and machine feedback could thus combine to form a new ritual in which transparency and plurality are more important than formal accreditation by a small number of anonymous reviewers.

New rituals may also emerge in education. Writing a dissertation may lose some of its formative power, but this does not mean that academic formation itself will disappear. Instead, the critical evaluation of AI output could develop into a new rite of passage. Students would no longer be assessed solely on their ability to produce a text, but on their capacity to distinguish what is reliable, valid and relevant in a world saturated with automatically generated knowledge.

These possibilities suggest that resistance to AI cannot be explained solely by a sense of loss, but also by uncertainty about what will replace existing practices. If AI can help shape new rituals that better reflect the values of integrity, transparency and inclusivity, the pedagogical and institutional culture of academia may even benefit. This nevertheless requires deliberate choices: rituals do not arise spontaneously, but are constructed and legitimised by communities.

References for Chapter 5

Bastow, S., Dunleavy, P., & Tinkler, J. (2014). The impact of the social sciences: How academics and their research make a difference. SAGE. https://doi.org/10.4135/9781473957947

Becher, T., & Trowler, P. R. (2001). Academic tribes and territories: Intellectual enquiry and the culture of disciplines (2nd ed.). Open University Press.

Biagioli, M. (2002). From book censorship to academic peer review. Emergences: Journal for the Study of Media & Composite Cultures, 12(1), 11–45. https://doi.org/10.1080/1045722022000003430

Crawford, K. (2021). Atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press.

Holmes, W., Bialik, M., & Fadel, C. (2021). Artificial intelligence in education: Promises and implications for teaching and learning. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000376709

Knox, J. (2020). Artificial intelligence and education futures: Critical perspectives on learning, knowledge and culture. Routledge.

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). Intelligence unleashed: An argument for AI in education. Pearson.

Moher, D., Bouter, L., Kleinert, S., Glasziou, P., Sham, M. H., Barbour, V., … & Dirnagl, U. (2020). The Hong Kong principles for assessing researchers: Fostering research integrity. PLoS Biology, 18(7), e3000737. https://doi.org/10.1371/journal.pbio.3000737

Müller-Birn, C., Dobusch, L., & Seitz, S. (2020). Peer review in the age of artificial intelligence. Science and Public Policy, 47(6), 749–759. https://doi.org/10.1093/scipol/scaa039

Prinsloo, P., & Slade, S. (2017). An elephant in the learning analytics room: The case for critical research in learning analytics. Learning Analytics: Fundaments, Applications, and Trends, 105–128. Springer. https://doi.org/10.1007/978-3-319-52977-6_6

Selwyn, N. (2019). Should robots replace teachers? AI and the future of education. Polity Press.

Chapter 6 – Power, Normativity and the Nature of AI-Driven Knowledge Production

6.1 Introduction

The third sub-question in this dissertation is: How does artificial intelligence reconfigure power relations, and what normative order is both desirable and feasible for AI-driven knowledge production? This question touches upon the core of scientific practice. Whereas power and normativity have traditionally often been regarded as external conditions — contexts within which science takes place but which do not alter the core of knowledge production itself — in the case of AI these dimensions prove to be constitutive. Access to infrastructure, the distribution of data and the ways in which norms are embedded in systems determine not only who can produce knowledge, but also which knowledge is regarded as valid. This chapter demonstrates that power and normativity fundamentally shape both the nature and the quality of AI-driven knowledge production.

6.2 Power as a Condition of Knowledge Production

AI-driven knowledge production depends on large-scale infrastructures: computing power, energy, cloud services and specialised talent. Training foundation models requires investments that only a handful of actors can afford. As Brookings and GovAI (2023) have demonstrated, these high fixed costs generate economies of scale and consequently concentrate power among a small number of companies and research consortia. The UK Competition and Markets Authority warned that vertical alliances between cloud providers, chip manufacturers and model developers further restrict access.

This concentration of power has direct consequences for the nature of knowledge production. Researchers without access to such infrastructure must rely on pre-trained models. They thereby lose influence over crucial early stages of research: the formulation of questions, the selection of datasets and the choices embedded in model design. Knowledge production consequently becomes conditional and relational: not everyone can participate, and participation depends on positions of power within a technological ecosystem.

6.3 Embedded Normativity

Normativity in AI is not merely an external criterion against which the technology is assessed, but a constitutive element of the technology itself. Whereas traditional science can rely on shared methodological standards that are evaluated retrospectively through peer review, replication or ethics committees, in AI norms are often embedded in datasets and models in advance. As Biagioli (2002) has shown, academic communication and evaluation processes perform not only an informational function, but also a ritual and legitimising role in knowledge production. AI-driven knowledge production is therefore shaped from the outset by both implicit and explicit choices.

A first problem concerns selection and exclusion within datasets. Every dataset is a reduction of reality: certain sources, languages or groups are included, while others are not. If a model is trained predominantly on English-language scientific literature, it primarily learns to reproduce and legitimise those perspectives. This is not merely a matter of bias, but also a normative choice concerning what

counts as knowledge and who is represented within that process. The claim to universality projected by many AI models obscures the fundamentally contingent nature of their foundations.

A second problem lies in defining what counts as legitimate output. Documentation practices such as Model Cards (Mitchell et al., 2019) and Datasheets for Datasets (Gebru et al., 2018) frequently specify the purposes for which a model is intended and how it may be used responsibly. Although these practices appear to promote transparency, they also relocate normative power: developers determine which uses are 'appropriate' and which applications fall outside the prescribed boundaries. For knowledge production, this means that scientists are often required to operate within boundaries drawn by others, without being able to renegotiate those boundaries independently.

A third problem concerns epistemic dependency. Researchers who use an existing model inevitably adopt its embedded assumptions. It is often no longer possible to determine which selection criteria were applied, which data were removed or which optimisations were performed. Knowledge production thereby becomes dependent on norms that researchers neither selected themselves nor are able to control. Responsibilities traditionally attributed to the researcher shift towards invisible actors and processes within the AI-development chain.

Concrete examples illustrate how this may become problematic. In the natural sciences, an AI system analysing climate models may exclude certain historical datasets, producing patterns that appear robust but are in fact contingent. In the humanities, a language model may implicitly determine that certain dialects or genres are irrelevant, thereby reinforcing a canon without researchers explicitly intending to do so. In the social sciences, sentiment analysis of Twitter may be presented as objective knowledge about 'society', while the underlying normative claim is that Twitter is representative of that society.

The central point is that AI-driven knowledge production derives its legitimacy not only from classical criteria such as transparency and reproducibility, but also from choices invisibly embedded in data and models. As long as those choices are not made explicit and subjected to discussion, the quality of the knowledge produced remains fundamentally uncertain.

6.4 Decontextualisation and Data Scarcity

One of the most fundamental challenges to the quality of AI-driven knowledge production lies in the ways in which data are collected and reused. Data originally produced within a specific context — such as medical records, social-media posts or ethnographic observations — lose their original meaning once they are deployed within another framework. This process of decontextualisation is not new, but AI gives it a systematic and large-scale character. Models are trained on datasets that largely conceal their provenance, causing crucial information about circumstances, purpose and interpretative frameworks to be lost.

The consequences for the quality of knowledge production are immediately apparent. Without context, the reliability of the patterns identified by AI becomes uncertain: correlations may emerge that were meaningful in the original setting but prove misleading in a new domain. The claim to universality often projected by AI models therefore masks the fragility of their foundations.

Moreover, data themselves are increasingly becoming a scarce resource. High-quality, well-documented and representative datasets are rare. Competition for access to such data concentrates

power among those actors capable of collecting, purchasing or restricting access to this resource. As Couldry and Mejías (2019) demonstrate, data infrastructures are not merely a technical matter, but create new forms of extraction and appropriation that are directly connected to global power relations. Power over data consequently means power over knowledge: those with access can produce knowledge, while those who are excluded remain dependent on the models and assumptions of others.

The combination of unavoidable decontextualisation and the scarcity of usable data makes clear that AI-driven knowledge production is vulnerable. This is not only because models may contain distortions or biases, but also because the conditions under which data become available determine which questions can be asked in the first place. Power over data thus translates directly into power over knowledge.

6.5 The Black-Box Character and Normative Tensions

The most fundamental problem for the legitimacy of AI-driven knowledge production is the black-box character of the technology. Users and scientists alike often lack insight into the datasets on which models were trained, the precise algorithmic choices embedded within them, or the mechanisms through which an output is produced. This lack of transparency means that every application entails uncertainty: not only about the outcome, but also about the epistemic basis on which that outcome rests.

The normative tensions described in this chapter are, in fact, manifestations of this black-box character. The conflict between transparency and commercial secrecy arises because many models remain closed on the grounds of intellectual property and competitive advantage. This makes it impossible to assess systematically whether outputs are reliable or which biases are present.

The debate surrounding open versus closed source similarly concerns the extent to which the box can be opened. Open-source models appear more accessible, but their training is likewise often based on datasets whose provenance and selection cannot be fully reconstructed. Openness therefore does not necessarily guarantee insight, while it may expose new risks, such as misuse by malicious actors.

Even the debate on sustainability illustrates how the black-box character operates. As Strubell, Ganesh and McCallum (2019) have demonstrated empirically, the energy and resource consumption involved in training AI systems remains largely hidden behind technical and commercial barriers and therefore forms part of their epistemic opacity. Users see the outputs, but not the ecological cost attached to them. This creates a false sense of neutrality: knowledge production appears immaterial, while in reality it is firmly embedded in material costs and power structures. Zuboff (2019) situates these developments within the broader context of surveillance capitalism, in which data infrastructures create new forms of dependency and extraction with direct consequences for knowledge production.

As a result, every attempt to evaluate AI-driven knowledge production inevitably struggles with this black-box character. As long as the box cannot be opened, claims concerning validity, reliability and universality remain fundamentally uncertain. Transparency instruments and regulatory initiatives such as the European Union's Artificial Intelligence Act (2024) and the OECD Principles on Artificial Intelligence (2019) may partially illuminate the box, but they do not alter the fact that the epistemic core of AI systems remains largely inaccessible. This is why AI-driven knowledge production, however promising, remains subject to fundamental uncertainty concerning its nature and quality.

6.6 Generative and Dedicated AI: Different Implications for Knowledge Production

Thus far, this chapter has frequently referred to generative AI models trained on enormous text and image corpora. Such models, including GPT systems, make the black-box character and the problem of embedded normativity particularly visible. Their training data are extensive but undocumented, their assumptions largely opaque, and their outputs appear universal even though they are in fact based on an uncontrollable selection. For the quality of AI-driven knowledge production, this means that generalisability and validity are invariably surrounded by uncertainty: users are often unable to reconstruct how a particular conclusion was reached.

This picture is not representative of all AI applications, however. Many domains employ dedicated AI systems trained on specific and clearly delimited datasets, such as medical imaging data, astronomical observations or experimental laboratory data. In these cases, the context of the data is often better documented, the normative choices involved in selection and annotation are more explicit, and the black box can in principle be smaller and more readily auditable. The problem consequently shifts: less emphasis falls on radical opacity, and more on access to scarce datasets and the power of institutions that own or regulate such data.

This comparison demonstrates that AI-driven knowledge production cannot be assessed in uniform terms. In generative models, the principal vulnerability lies in the epistemic opacity of the underlying corpora. In dedicated models, it lies in the concentration of power surrounding data infrastructures and in the normative choices made during labelling and interpretation. Each configuration therefore presents its own problems, and both must be explicitly considered when assessing the quality and legitimacy of AI-driven knowledge production.

6.7 Provisional Conclusion

The analysis of power and normativity shows that AI-driven knowledge production is fundamentally different in nature from traditional forms of science. Power is not merely contextual, but determines who can participate and which questions can be asked. Normativity is not external, but embedded in data, models and infrastructures. Universality cannot be taken for granted, but is fragile because representation and legitimacy are distributed unevenly. Knowledge production is not autonomous, but relational and conditional: dependent on access, inclusion and sustainability.

As UNESCO (2021) emphasises in its recommendations on the ethics of artificial intelligence, normative questions concerning AI-driven knowledge production extend beyond purely academic or technical considerations and form part of global governance debates. The provisional answer to the third sub-question is therefore that AI-driven knowledge production can be legitimate only when the constitutive role of power and normativity is explicitly recognised and regulated. Quality can no longer be assessed solely in terms of methodological rigour or peer review, but must also be evaluated in relation to access to infrastructure, plurality of perspectives and transparency concerning embedded norms. Only by applying these broader criteria can knowledge production in an AI era be both robust and equitable.

6.8 Qualification and Mitigation

The analysis in this chapter has emphasised that power, normativity and the black-box character are constitutive of AI-driven knowledge production. The foundations of scientific legitimacy therefore appear to be under constant pressure. It is nevertheless important to qualify this conclusion. The

question is not only which problems exist, but also how serious they are in practice and to what extent they can be mitigated.

Policymakers and research institutions will inevitably make pragmatic trade-offs. If AI produces reliable results in 99.5 per cent of cases and only 0.5 per cent are problematic, there may be a strong temptation to accept the remaining uncertainty. After all, margins of error have always existed throughout the history of science: instruments are imperfect, measurements are fallible, and human researchers likewise make interpretative choices that may introduce distortion. The central question is therefore not whether AI must be error-free, but which margin of error is considered socially and scientifically acceptable. In the experimental natural sciences, this may involve a small degree of statistical uncertainty, whereas in medical or legal applications even rare errors may have serious consequences.

A range of strategies has now been developed to mitigate the risks of AI-driven knowledge production. At the technical level, instruments such as Model Cards (Mitchell et al., 2019) and Datasheets for Datasets (Gebru et al., 2018) systematically document which data were used and which assumptions were made. Bias-detection techniques and fairness metrics can help to reveal distortions. At the institutional level, forms of auditing are being developed in which models are periodically subjected to external review. The NIST Artificial Intelligence Risk Management Framework (2023) provides policy instruments intended to address systematically the reliability and risks of AI-driven knowledge production. Hybrid forms of human–machine collaboration may also increase reliability: AI generates patterns, while human researchers critically evaluate and contextualise them.

The use of labelled data requires particular attention. Many dedicated AI systems are based on supervised learning: models are trained using datasets labelled by experts, such as medical scans classified by radiologists as indicating the ‘presence’ or ‘absence’ of a tumour. This appears more transparent than training on undocumented corpora, but normative choices are involved here as well: who determines what counts as the correct label, how consistently are labels applied, and what happens to cases that do not fit the predefined categories? In this context, mitigation primarily means making labels explicit, subjecting them to intersubjective scrutiny and allowing scope for revision.

This qualification shows that the problems outlined in this chapter are serious, but not fatal. Despite these concerns, AI-driven knowledge production is likely to become widespread. The question is not whether the technology will be used, but under which conditions and with which safeguards. The challenge is to strike a balance: accepting margins of error where this is responsible, while simultaneously establishing mechanisms to limit structural distortion, abuses of power and black-box uncertainty. Only in this way can AI-driven knowledge production develop into a practice that is not only efficient and powerful, but also scientifically legitimate and socially responsible.

References for Chapter 6

Biagioli, M. (2002). From book censorship to academic peer review. Emergences, 12(1), 11–45.

Brookings Institution / GovAI. (2023). Analyses of foundation model concentration and market power [Policy reports].

Couldry, N., & Mejías, U. A. (2019). The costs of connection: How data is colonizing human life and appropriating it for capitalism. Stanford University Press.

European Union. (2024). Artificial Intelligence Act. Official Journal of the European Union.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. Communications of the ACM, 64(12), 86–92.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019).

Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–229.

NIST. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology.

OECD. (2019). OECD Principles on Artificial Intelligence. Organisation for Economic Co-operation and Development.

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the ACL, 3645–3650.

UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. United Nations Educational, Scientific and Cultural Organization.

Zuboff, S. (2019). The age of surveillance capitalism. PublicAffairs.

Intermission – Interim Assessment and Agenda

Following the discussion of the three impact dimensions — cognitive and methodological (Chapter 4), pedagogical and institutional (Chapter 5), and power and normativity (Chapter 6) — it is useful to take provisional stock. The sub-questions have received an initial treatment, but have not yet been answered in their full breadth. What has become clear is that AI-driven knowledge production presents itself as a hybrid phenomenon: at once promising and problematic, enriching and undermining. In the preceding chapters, these tensions were examined from the perspective of one dimension at a time. The purpose of this interlude is to bring those insights together in three taxonomies — of opportunities, risks and challenges — which collectively provide the analytical basis for the archetype chapters that follow.

I. Taxonomy of opportunities. The first taxonomy concerns the opportunities offered by AI-driven knowledge production in general. These opportunities are not merely by-products of technological progress, but affect the core processes of science.

1. **Acceleration and scale.** AI makes it possible to produce knowledge at a pace and on a scale that were previously inconceivable. Data processing and literature reviews no longer proceed slowly and linearly, but cyclically and dynamically: hypotheses can be generated, tested and revised more rapidly. This acceleration opens the way not only to greater efficiency, but also to new forms of scientific practice in which feedback loops are shorter and discoveries circulate more quickly.
2. **Discovery and synthesis.** AI also functions as an epistemic tool that reveals unexpected patterns. Whereas human researchers are constrained by disciplinary frameworks and cognitive limitations, AI can uncover relationships within heterogeneous datasets that would otherwise remain hidden. Breakthroughs have already been achieved in the biological sciences in predicting protein structures, while new relationships have been identified within extensive corpora in the humanities. For AI-driven knowledge production, this means that fragmentation can be transcended and synthesis enabled on a scale that was previously unattainable.
3. **Pedagogical and institutional innovation**. A third cluster of opportunities lies in the pedagogical and institutional sphere. AI can function as a personalised learning and guidance instrument and as a multiplier of evaluative perspectives. Traditional forms of peer review can be enriched by parallel contributions from AI systems analysing argumentative consistency or methodological robustness. Instruments such as Model Cards and Datasheets also enable new forms of transparency and systematic accountability. In terms of knowledge production, this means that the institutions and practices in which science is embedded also acquire scope for innovation.

II. Taxonomy of risks. These opportunities are accompanied by equally structural risks. These risks do not affect the margins, but the core of what makes scientific knowledge legitimate.

1. **Black boxes and loss of transparency.** AI models often function as black boxes. Users generally do not know exactly which data were used or how algorithms reach decisions. This undermines reproducibility, a fundamental criterion of scientific validity. For AI-driven knowledge production, this means that accountability for

results becomes dependent on infrastructures and datasets that are often inaccessible or commercially closed.

2. **Decontextualisation and loss of nuance**. Data collected in specific contexts lose their meaning when they are used in new environments without adequate interpretation. AI intensifies this process by constructing patterns on a large scale from data detached from their original settings. Outcomes consequently acquire the appearance of universal applicability, while in reality they remain contextually contingent. This threatens the reliability of knowledge produced through AI.
3. **Ritualisation and loss of communicative meaning**. In addition to its informational function, scientific communication also performs a ritual role: it legitimises researchers as members of a community and transmits values such as rigour, transparency and dialogue (Biagioli, 2002). When AI takes over both the production and assessment of dissertations and articles, this ritual dimension risks being hollowed out. Texts then circulate primarily between humans and machines without exercising the same community-forming force. For AI-driven knowledge production, this entails a shift in legitimacy: from intersubjective evaluation towards technological output.
4. **Concentration of power and dependency.** Because high-quality data and computing power are scarce and expensive, they become concentrated among a limited number of companies and institutions. This concentration of power means that epistemic agendas are implicitly or explicitly shaped by commercial or geopolitical interests. Researchers dependent on external models consequently lose part of their autonomy.
5. **Fabrication and reliability of references**. One particular risk, closely connected to the black-box character but requiring separate attention, concerns the fabrication of bibliographic references. Large language models can generate convincing literature reviews while producing references that are partly or wholly fictitious. For AI, sources are not meaningful markers within scientific dialogue, but textual patterns. This undermines a core scientific practice: legitimising new knowledge by explicitly relating it to the existing literature. For AI-driven knowledge production, this poses a direct threat to the reliability and credibility of literature reviews.

III. Taxonomy of challenges. The third taxonomy brings together the fundamental challenges revealed by the combination of opportunities and risks. These are questions that will shape the future organisation of AI-driven knowledge production.

1. **Recalibrating validity and reliability.** The classical quality criteria of science can no longer be taken for granted. Validity and reliability must be redefined in a context in which the instruments themselves are epistemically opaque. New forms of accountability, such as auditing procedures and metadata documentation, are needed to safeguard trust in AI-driven knowledge production.
2. **Institutional integration without erosion of core rituals**. AI must be integrated into education, supervision and publication without eliminating the social and ritual functions of these practices. The challenge is not only technical, but also cultural: how can the academic community preserve its formative rituals as AI assumes an increasing number of tasks?
3. **Context-specific standards for error margins**. No system is infallible, and AI is no exception. The question is therefore not whether errors occur, but which margin of error is acceptable in which context. In fundamental research, some uncertainty may be acceptable as long as robust trends remain visible. In medical or legal contexts, by

contrast, even extremely rare errors may be unacceptable. This differentiation requires explicit standard-setting for each domain and application.

IV. Bridging function. These three taxonomies complete the interim assessment. They bring together the insights from the three dimension chapters and sharply illuminate the tensions characteristic of AI-driven knowledge production. The taxonomies serve as an analytical framework for the chapters that follow, in which the archetypes of science take centre stage. For each archetype, the analysis will systematically examine which opportunities are realised in practice, which risks manifest themselves most strongly, and which challenges are most urgent.

This interlude is therefore more than a summary: it marks a transition. Whereas the preceding chapters treated the dimensions separately, the whole now becomes visible in a structured form. At the same time, the interlude looks ahead: the question is no longer whether AI-driven knowledge production is changing science, but how those changes take concrete form within different knowledge practices.

Chapter 7 – Natural Science Research and AI-Driven Knowledge Production

7.1 Introduction

The natural sciences are traditionally characterised by their strong emphasis on empirical verification, reproducibility and causality. These very characteristics make the domain particularly interesting for analysing the role of artificial intelligence (AI). AI is deployed here on a large scale, in fields ranging from molecular biology to astronomy and climate science. Unlike in more interpretative disciplines, its use is generally directed towards pattern recognition, data analysis and hypothesis formation, often in the form of dedicated AI systems trained on domain-specific datasets.

The central question in this chapter is how the opportunities, risks and challenges of AI-driven knowledge production manifest themselves within the natural sciences. This concerns not only the practical usefulness of AI, but above all the extent to which its results are scientifically legitimate and reliable.

To provide an empirical foundation for this question, the chapter adopts an asymmetrical approach. AlphaFold in molecular biology serves as the primary case and is examined in depth across all elements of the taxonomy. This choice is not arbitrary: AlphaFold exemplifies the tensions and possibilities of dedicated AI systems in the natural sciences and thereby provides a rich empirical anchor for conceptual analysis. Examples from astronomy and climate science are used additionally to support the generalisability of the findings, without claiming to treat these domains in equal depth.

The analysis applies the taxonomies developed in the interlude and illustrates how the cognitive, institutional and normative dimensions of AI-driven knowledge production are expressed within this archetype.

7.2 Opportunities

Acceleration and scale. One of the most tangible opportunities offered by AI in the natural sciences is the acceleration of data processing. The most striking example is AlphaFold, a DeepMind system capable of predicting three-dimensional protein structures with remarkable accuracy (Jumper et al., 2021). Such predictions previously required years of experimental research, whereas AlphaFold made hundreds of thousands of structures available within a matter of months. This acceleration has led not only to greater efficiency, but also to a democratisation of knowledge: the resulting databases are publicly accessible, giving researchers worldwide access to information that was previously restricted to specialised laboratories.

The AlphaFold breakthrough merits closer analysis. Jumper et al. (2021) write in Nature: “The ability to predict a protein’s shape computationally from its amino acid sequence would be a boon to life sciences and medicine. It would vastly accelerate efforts to understand the building blocks of cells and enable quicker and more advanced drug discovery.” This is not mere rhetoric. Their system achieved a

median score of 92.4 GDT in the CASP14 test[1], whereas the best previous methods had remained below 60. Yet Jumper et al. are also candid about the limitations. They state: “AlphaFold produces a per-residue estimate of its confidence on a scale from 0-100. This confidence measure is calibrated to correspond approximately to the model’s expected per-residue accuracy” (p. 587). This is crucial: the system knows where it is uncertain. This contrasts with the black-box criticism frequently directed at AI. Although AlphaFold is complex, it is not entirely opaque — it communicates its own uncertainty.

The contribution of AI is equally evident in astronomy. Modern telescopes generate enormous quantities of data that cannot possibly be analysed manually. AI models are used to identify patterns in light curves and detect exoplanets (Valenti et al., 2021). This radically expands the scale of observation and analysis: whereas only a few dozen planets could previously be discovered each year, AI makes it possible to evaluate thousands of potential candidates.

AI offers comparable benefits in climate science. Machine-learning models are used to refine weather forecasts and improve climate projections (Rolnick et al., 2019). Because these models can be recalculated more rapidly than traditional simulations, they enable a more dynamic process of hypothesis formation and scenario analysis. This allows researchers to respond more quickly to new data and thereby shorten the empirical cycle.

Discovery and synthesis. In addition to acceleration, AI enables new forms of discovery and synthesis. AlphaFold not only produces results more rapidly, but also generates hypotheses about structures that have not yet been established empirically. This expands the scope of scientific imagination: AI functions here as a source of new questions that can subsequently be tested experimentally.

The same applies in astronomy. AI systems have identified gravitational lenses and patterns in cosmological datasets that were previously invisible (Fluke & Jacobs, 2020). These discoveries shift the horizon of research by creating new opportunities to test theories concerning dark matter and cosmological expansion.

Climate science demonstrates how AI can contribute to synthesis. By integrating diverse datasets — ranging from satellite imagery to historical measurement series — models can generate more robust predictions. AI functions here not only as an accelerator, but also as an instrument that brings fragments of data together into a more coherent whole.

Pedagogical and institutional innovation. Finally, AI offers opportunities for pedagogical and institutional innovation. Databases such as the Protein Data Bank, which AlphaFold has expanded considerably, serve as educational and research tools for students and scientists worldwide. Quality assurance also benefits: AI systems can screen datasets for inconsistencies or anomalies, allowing human reviewers to concentrate on substantive evaluation. This produces a form of layered quality assurance that enhances both efficiency and reliability.

[1] A median score of 92.4 GDT on the CASP14 test refers to AlphaFold’s performance in the international Critical Assessment of Protein Structure Prediction (CASP) competition. The GDT_TS score (Global Distance Test – Total Score) measures the structural similarity between a predicted protein structure and one determined experimentally, on a scale from 0 to 100. A score above 90 is regarded as achieving experimental accuracy, meaning that the calculated three-dimensional shape of the protein scarcely differs from the structure established in the laboratory using physical methods such as X-ray diffraction or cryogenic electron microscopy. By achieving a median score of 92.4 across all test proteins, AlphaFold therefore attained a level of precision virtually equivalent to laboratory accuracy — a leap from previous leading models, which scored below 60 GDT, to near-perfect structural predictions — widely recognised as a breakthrough in structural biology.

7.3 Risks

The natural sciences unmistakably demonstrate the power of AI, but they are not immune to risks. Within this domain, attention shifts from interpretative uncertainties towards questions concerning infrastructure, data and transparency.

Black boxes and limited explainability. Even when models are dedicated and domain-specific, the black-box problem remains. AlphaFold's accuracy is impressive, but the route by which it arrives at its results can be understood only in part (Jumper et al., 2021). In astronomy, too, AI systems identify patterns in light curves that researchers cannot always explain in physical terms (Valenti et al., 2021). This raises the question of whether knowledge lacking explainability can have the same status as knowledge produced through classical methods.

Jenna Burrell's influential article 'How the Machine "Thinks"' (2016) distinguishes three forms of algorithmic opacity: 'opacity as intentional corporate or state secrecy', 'opacity as technical illiteracy', and 'opacity from the mismatch between mathematical optimization and human semantic interpretation' (p. 3)[2]. AlphaFold is primarily affected by the third form. Burrell writes: 'When we are seeking to understand a decision made by a machine learning algorithm, the challenge is not only in the inscrutability of any given operation but in the interplay of weights and connections across the model' (p. 10). This is precisely what can be observed in AlphaFold: the architecture has been published and the code is open source, but the way in which 21 million parameters interact to produce a protein structure remains largely inaccessible to human interpretation. Burrell's analysis nevertheless requires qualification. She wrote her article in 2016, before the development of attention mechanisms and interpretability tools that are now standard. AlphaFold's confidence scores constitute a form of what Burrell calls 'the provision of a rationale' — not perfect, but substantially better than purely black-box systems.

Decontextualisation and data quality. The reliability of AI in the natural sciences stands or falls with the quality of the datasets. Climate models, for example, are more robust in regions with a high density of measurements, but exhibit greater uncertainty in areas with limited data coverage (Rolnick et al., 2019). The appearance of universality thus conceals inequalities in data collection and may lead to an overestimation of the precision of predictions.

Concentration of power and dependency. Another risk concerns access to computing power and infrastructure. Training advanced AI models requires investments that only a small number of actors can afford. Brookings and GovAI (2023) warn that this creates oligopolistic structures in which a limited number of companies partly determine the epistemic agenda. For researchers in the natural sciences, this creates dependency: those without access to such infrastructure can use only pre-trained models, with all the limitations and assumptions embedded within them.

Fabrication of references. Although knowledge production in the natural sciences is generally grounded in empirical data, vigilance is also required when generative models are used for literature research. Studies show that language models regularly produce fictitious or incorrect references (Borji,

[2] Burrell (2016) distinguishes three sources of algorithmic opacity. The first is intentional secrecy ('intentional corporate or state secrecy'), in which companies or governments deliberately conceal the operation of their algorithms for commercial or strategic reasons. The second is technical illiteracy ('technical illiteracy'): most users, policymakers and even researchers lack the mathematical or computational expertise required to understand complex models. The third is a structural semantic gap ('the mismatch between mathematical optimization and human semantic interpretation'): algorithms optimise numerical functions, whereas humans assign meaning through language and context. Consequently, even with full access to the code, part of the system's operation remains inherently difficult to translate into human concepts.

2023). When such instruments are used uncritically in secondary processes, such as writing reviews or preparing research proposals, they may undermine the reliability of scientific communication.

7.4 Challenges

The analysis of opportunities and risks makes clear that the natural sciences constitute a distinctive case within the broader context of AI-driven knowledge production. On the one hand, this domain convincingly demonstrates how AI can contribute to acceleration, discovery and synthesis. On the other hand, specific challenges emerge that are not merely incidental, but structurally connected to the nature of this archetype. Three clusters of challenges merit particular attention: methodological, institutional and normative.

Methodological challenges: validity and explainability. The core values of the natural sciences are transparency, reproducibility and causality. AI applications conform to these only partially. The results of systems such as AlphaFold are reproducible, but their explainability remains limited (Jumper et al., 2021). In astronomy, algorithms identify patterns that are consistent, but whose physical basis is not always clear (Fluke & Jacobs, 2020). This raises the question of how validity should be defined: is a prediction that is empirically correct but not theoretically understood legitimate scientific knowledge? The answer to this question will largely determine how AI can become established within this domain.

Epistemological tension — expertise versus automation. In Rethinking Expertise (2007), Harry Collins and Robert Evans analyse what they call 'interactional expertise': 'enough expertise to interact interestingly with participants' without oneself being a practitioner. This concept is highly relevant to AI in the natural sciences. They warn: 'The danger is that we come to treat the computer's correlations as understanding'. This goes to the heart of the AlphaFold dilemma. The system has no 'understanding' of biochemistry in the human sense, yet it produces useful predictions. Collins and Evans would ask whether this constitutes contributory expertise — a genuine contribution to knowledge — or merely a form of 'mimicry'. Their answer would be nuanced. They acknowledge that 'expertise is not just about possessing information but about being able to make judgments' (p. 45). AlphaFold does not make judgements in the traditional sense, but the researchers who use it do. This suggests a new form of hybrid expertise in which humans and machines operate complementarily.

Institutional challenges: infrastructure and access. A second cluster of challenges concerns access to infrastructure. The computing power and datasets required for large-scale AI applications are often controlled by a limited number of companies or research consortia. Brookings and GovAI (2023) point to the oligopolistic structure that consequently arises. For researchers, this limits autonomy: they may test or replicate results, but rarely possess the resources needed to train a model of comparable scale themselves. Institutionally, this requires mechanisms that guarantee equitable access, for example through public investment in data infrastructures or international agreements on data sharing.

Normative challenges: error margins and legitimacy. Finally, there are normative questions concerning error margins and legitimacy. In climate models, for example, even a small margin of error may have major consequences for policy decisions (Rolnick et al., 2019). This requires explicit standard-setting: what degree of uncertainty is acceptable, and who determines this? The problem becomes even more complex when AI is used to generate predictions that extend beyond the existing empirical basis. The legitimacy of such knowledge rests not only on statistical robustness, but also on acceptance by the scientific community. Normative judgements are therefore unavoidable: reliability is not merely a technical property, but also a socially recognised quality criterion (Collins & Evans, 2007).

7.5 Conclusion

The analysis of the natural sciences clarifies how artificial intelligence affects knowledge production within this archetype. To maintain analytical precision, the three sub-questions are explicitly restated and provisionally answered.

Sub-question 1: *How does AI change the cognitive and methodological dimension of knowledge production?* In the natural sciences, AI reinforces existing methodological core values such as reproducibility and empirical testing. Examples including AlphaFold (Jumper et al., 2021) and AI applications in astronomy (Fluke & Jacobs, 2020; Valenti et al., 2021) demonstrate that predictions are repeatable and testable, but often remain only partially explainable. The methodological challenge therefore lies not in a loss of reproducibility, but in a shift from explanatory to predictive knowledge. This means that the legitimacy of scientific knowledge is being redefined: whereas the natural sciences traditionally seek causal understanding ('why does X happen?'), AI-driven knowledge production often accepts instrumental predictability ('what happens if X?'). Accurate prediction without mechanistic understanding does not undermine the reliability of results, but it does alter the nature of what counts as a 'scientific explanation'. The provisional answer to the first sub-question is that AI accelerates and expands the empirical cycle in the natural sciences, but shifts the legitimacy of knowledge from causal understanding towards predictive accuracy as its primary basis.

Sub-question 2: *How does AI affect the pedagogical and institutional dimension of knowledge production?* AI makes scientific knowledge more accessible and expands its institutional applicability. The enrichment of public databases, such as AlphaFold's expansion of the Protein Data Bank, democratises access to knowledge and increases the educational value of AI systems. Institutionally, however, new dependencies are also emerging. Access to computing power and high-quality datasets is becoming concentrated among a limited number of companies and consortia (Brookings Institution / GovAI, 2023). This means that the democratisation of knowledge is conditional: although AI-generated datasets are widely available, the capacity to produce new knowledge remains confined to actors with access to large-scale infrastructure. Institutionally, power is shifting from individual researchers and smaller institutes towards large consortia and technology companies that own the infrastructure. The provisional answer to the second sub-question is that AI enhances pedagogical possibilities and facilitates knowledge sharing, but that these benefits are accompanied by structural dependencies that restrict researchers' autonomy. Accessibility of results is not the same as democratic participation in knowledge production.

Sub-question 3: *How does AI reconfigure power relations, and what normative order is both desirable and feasible?* In the natural sciences, the power dimension manifests itself primarily in access to infrastructure and datasets. Reliability and legitimacy are no longer exclusively methodological issues, but are also determined by who has access to data and computing power. Normatively, this necessitates explicit choices concerning error margins, the legitimacy of predictions and criteria governing access to infrastructure. Power thereby shifts from epistemic to infrastructural control. Those who determine which datasets are collected, who has access to computing power and which questions can be addressed computationally also help to define the boundaries of what can be known. The normative challenge is not merely to define criteria for 'good AI science', but to create the conditions under which diverse actors can participate in knowledge production. The provisional answer to the third sub-question is that power and normativity within this archetype are shaped primarily by access to infrastructure and institutional arrangements, and that legitimacy can be guaranteed only when these conditions are explicitly regulated.

In summary. In the natural sciences, AI is transforming knowledge production profoundly, but at a different locus from the one often assumed. The core values of empirical testing and reproducibility remain intact and are even reinforced. The vulnerability lies instead in the shift from explainability towards predictability, in institutional dependency on infrastructures, and in normative choices concerning access and error margins. AI-driven knowledge production in the natural sciences is therefore both powerful and conditional: the quality of the knowledge is high, but it depends entirely on the institutional and normative conditions that make this domain possible.

References for Chapter 7

Borji, A. (2023). A categorical archive of ChatGPT failures. arXiv preprint arXiv:2302.03494.

Brookings Institution / GovAI. (2023). Analyses of foundation model concentration and market power.
Collins, H., & Evans, R. (2007). Rethinking expertise. University of Chicago Press.

Fluke, C. J., & Jacobs, C. (2020). Surveying the reach and maturity of machine learning and artificial intelligence in astronomy. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(2), e1349.

Jumper, J., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589.

Rolnick, D., et al. (2019). Tackling climate change with machine learning. arXiv preprint arXiv:1906.05433.

Valenti, S., et al. (2021). Machine learning applications for exoplanet detection. Astronomy & Astrophysics, 645, A62. https://doi.org/10.1051/0004-6361/202039448

Chapter 8 – Historical Research and AI-Driven Knowledge Production

8.1 Introduction

Historical scholarship differs fundamentally from the natural sciences. Whereas the latter emphasise reproducibility and causal explanation, historical research revolves around authenticity and context. A document, testimony or artefact acquires meaning only when it is carefully embedded within a broader narrative. The legitimacy of knowledge in this domain therefore rests on source criticism, interpretation and narrative accountability.

It is precisely here that the use of artificial intelligence (AI) raises new questions. AI can unlock large archives and reveal patterns, but it does so by datafying sources: texts, images and sounds are converted into data points detached from their original meaning. This creates major opportunities, but also risks. The central question is whether the reliability and legitimacy of historical knowledge production can be preserved as AI increasingly assumes tasks traditionally performed by historians.

This chapter analyses the opportunities, risks and challenges of AI in historical scholarship. Three specific areas of application take centre stage: the digitisation and analysis of Holocaust archives, the opening up of colonial archives, and the phenomenon of 'deepfake history'. These cases illustrate how AI is changing the conditions of historical knowledge production.

8.2 Opportunities

Acceleration and scale. AI can considerably accelerate the work of historians. For centuries, handwritten documents were accessible only to specialists, but systems such as Transkribus can automatically transcribe millions of pages (Kahle et al., 2017). In the case of Holocaust archives, this means that diaries, letters and witness statements previously dispersed across numerous archives can now be searched and compared. Historians can identify patterns in persecution, resistance or survival that previously remained hidden. This increase in scale is not only quantitative, but also qualitative. Whereas an individual researcher might study several thousand pages over the course of a career, AI can search millions of documents for specific names, places or events. This creates new opportunities for prosopographical research and network analysis, allowing relationships between people and events to be mapped on a large scale.

Discovery and synthesis. AI makes it possible to establish connections between dispersed and diverse sources. In colonial archives, for example, algorithms can identify names or places recurring across different documents, even when they appear in different languages or variant spellings. This may generate new historical insights into trade networks, family ties or administrative structures that would otherwise remain fragmented. The strength of pattern recognition lies in its capacity to survey large quantities of material. In his work on macroanalysis in literary studies, Jockers (2013) shows how large-scale textual analysis reveals patterns that remain hidden through close reading. For historical scholarship, this means that AI can identify trends in newspaper archives, pamphlets or correspondence that would have escaped the attention of individual researchers. Although Jockers' work focuses primarily on literary texts, his methodological argument — that large-scale patterns cannot replace narrative interpretation but can enrich it — applies directly to historical research.

Accessibility and democratisation. Digitisation and AI-supported access can make archives available to researchers who previously lacked access to them. Terras (2011) demonstrates how even amateur digitisation of texts can contribute to broader accessibility. This is crucial for historical scholarship: not only professional historians, but also local communities, relatives of victims and interested members of the public can now consult sources that were previously available only for physical inspection in specific archives. This democratisation also has an institutional dimension. Smaller universities and researchers in the Global South who lack the funds for extended archival visits in Europe or North America can now access source material through digital platforms. This may shift the balance of power in historical research and enable greater diversity of perspectives.

8.3 Risks

Loss of context and anachronism. The greatest risk of AI in historical scholarship is the loss of context. When a document is digitised and processed as a data point, it loses its material and institutional context. A letter acquires meaning through knowledge of who wrote it, to whom, under what circumstances and for what purpose. AI systems can transcribe the text, but lack the historical awareness required to reconstruct these contextual layers. Zwierlein (2016) warns of the danger of anachronism when contemporary categories are projected onto historical material. AI models are trained on present-day language and concepts and may therefore establish connections that had no historical meaning. An algorithm searching for 'revolution' in eighteenth-century documents, for example, may recognise the word without understanding that its meaning at the time differed fundamentally from its twenty-first-century meaning. In his work on distant reading, Underwood (2019) emphasises that pattern recognition becomes meaningful only when embedded within interpretative frameworks. For historians, this means that AI-generated patterns in newspapers, pamphlets or correspondence may be suggestive, but cannot be accepted without thorough historical interpretation. The risk is that researchers under time pressure or lacking critical reflection will too readily treat such patterns as facts.

Canon formation and power relations. Not all archives receive equal treatment in digitisation projects. The best-preserved and most readily accessible collections are often processed first — frequently European archives. Putnam (2016) emphasises that digitisation always casts 'shadows': what is digitally available becomes dominant, while what has not been digitised fades from view. This is particularly problematic in colonial archives. The perspectives of colonial rulers are often well documented and preserved, whereas local voices are fragmentary or entirely absent. When AI is used to open up these archives, European perspectives risk becoming canonised unless deliberate investment is made in digitising and contextualising sources from former colonies. Ketelaar (2001) refers in this context to 'tacit narratives': implicit stories that archives already carry through their organisation and description. AI amplifies this effect because algorithmic choices in metadata and search structures determine which histories come to the fore and which recede from view.

Fabrication of references and deepfake history. Generative AI introduces a new and fundamental risk: the creation of fictitious sources or inaccurate quotations (Borji, 2023). This affects not only secondary processes such as literature reviews, but the very foundations of historical reliability. A discipline that derives its legitimacy precisely from source attribution faces an existential threat when sources can no longer be verified. More serious still is the phenomenon of 'deepfake history': hyperrealistic reconstructions of voices, faces or historical scenes produced using AI. Such reconstructions may have educational value — for example, by bringing historical figures 'to life' — but they also create a dangerous space in which the distinction between fact and fiction becomes blurred. For historical

scholarship, which already struggles with the denial of events such as the Holocaust, this is not a theoretical risk but an acute threat to the societal legitimacy of historical knowledge.

8.4 Challenges

Methodological challenges: authenticity and narrativity. Historical research requires more than the identification of patterns: it requires the construction of a narrative that meaningfully connects sources. AI can identify correlations, but lacks historical awareness and interpretative intuition. The challenge is to ensure that researchers do not mistake AI-generated results for ready-made knowledge, but continually translate them critically into authentic and contextually accountable narratives. Jockers (2013) and Underwood (2019) both show that distant reading reveals patterns of literary interest, but that these acquire meaning only when embedded within interpretative frameworks. This methodological tension applies equally to historical research. Patterns in newspaper archives, pamphlets or correspondence may be suggestive, but historians must continually translate them into narratives that do justice to the complexity and contingency of the past. The challenge is to use AI as an instrument of pattern recognition without eroding the historian's hermeneutic task — understanding meaning in context. This requires methodological rigour and explicit accountability concerning the role played by AI in the research process.

Institutional challenges: representation and access. The digitisation and opening up of archives do not take place in a vacuum. Decisions about which archives are processed, which metadata are added and who is granted access to the results determine the horizon of historical research. Terras (2011) shows how amateur digitisation can influence canon formation, as some texts acquire a second digital life while others disappear. In colonial archives, this occurs on a larger scale. European institutions often determine which sources are made accessible and how they are described. Ketelaar's (2001) concept of 'tacit narratives' is particularly relevant here: through their organisation and description, archives invariably carry implicit narratives. AI amplifies this effect: algorithmic choices in metadata and search structures determine which histories come to the fore and which recede from view. The institutional challenge is to organise digitisation and AI-supported access in such a way that diversity of perspectives is preserved. This requires international collaboration in which not only European and North American institutions set the agenda, but archives and researchers from former colonies are also actively involved in designing metadata schemas and search algorithms.

Normative challenges: reliability and responsibility. The use of AI raises questions of responsibility. If an AI system incorrectly transcribes a Holocaust testimony or categorises a colonial source using inaccurate metadata, who bears responsibility for the error? Verhoeven (2016) emphasises that digital methods create normative obligations for researchers and institutions. It is not sufficient to say that 'the technology did it' — the decision to use AI and the manner in which it is deployed remain the responsibility of people and institutions. This is particularly sensitive in historical scholarship, because errors may have direct consequences for historical truth and public debate. An incorrectly transcribed date in a Holocaust document may be exploited by deniers. A colonial source linked through inaccurate metadata to the wrong place or period may alienate local communities from their own history. The normative challenge is therefore twofold: AI systems must be carefully designed and validated, while ultimate responsibility must remain clearly with human historians. This requires transparency regarding the use of AI, documentation of decisions and errors, and openness about the limitations of algorithmic methods.

8.5 Conclusion

The analysis of historical scholarship shows that AI affects this archetype in fundamental ways. To formulate the findings precisely, the three sub-questions are explicitly answered for this domain.

Sub-question 1: *How does AI change the cognitive and methodological dimension of knowledge production?* AI changes the cognitive practice of historical research through acceleration, expansion of scale and pattern recognition. Systems such as Transkribus make millions of pages searchable, while algorithms can identify connections between dispersed sources. This broadens the horizon of what historians can investigate and enables new forms of network analysis and prosopography. Methodologically, however, a fundamental tension arises between pattern recognition and interpretation. Where AI identifies correlations, the historian must construct meaning. The challenge is to use AI without losing authenticity and context. Jockers (2013) and Underwood (2019) show that large-scale analysis becomes meaningful only when embedded within hermeneutic interpretation — an insight directly applicable to historical research. The provisional answer is that AI expands historians' cognitive capacity, but that the methodological legitimacy of knowledge production remains dependent on critical source analysis and contextual accountability on the part of the researcher.

Sub-question 2: *How does AI affect the pedagogical and institutional dimension of knowledge production?* Pedagogically, AI creates new possibilities for teaching history. Through digital platforms, students can explore large archives and independently identify patterns that previously remained hidden. This may lead to a more active and research-oriented form of history education. Institutionally, however, the balance of power shifts in problematic ways. Terras (2011) and Putnam (2016) show how digitisation influences canon formation: what has been digitised becomes visible, while what has not been digitised recedes from view. In colonial archives, this means that European perspectives remain dominant unless active investment is made in digitising sources from former colonies. The challenge is to organise AI-supported access in a manner that ensures equitable representation and access. This requires international collaboration and explicit attention to who sets the agenda in digitisation projects. The provisional answer is that AI can promote institutional innovation and inclusion, but that this is sustainable only when representation and balanced access are actively safeguarded.

Sub-question 3: *How does AI reconfigure power relations, and what normative order is both desirable and feasible?* Power manifests itself in decisions concerning digitisation, metadata and access. These determine which narratives circulate and which disappear. Ketelaar (2001) shows how archives always carry implicit narratives through their organisation. AI amplifies this effect: algorithmic choices in search structures determine which histories come to the fore. Normatively, it is essential that transparency and responsibility remain central. Verhoeven (2016) emphasises that digital methods create normative obligations for researchers and institutions. It is not sufficient to deploy AI without explicitly accounting for the choices made and the limitations involved. The provisional answer is that AI can function legitimately within this domain only when the constitutive role of power is explicitly recognised and regulated, and when responsibility for interpretation remains unequivocally with the historian. This requires transparency concerning the use of AI, documentation of decisions and openness about limitations.

In summary. For historical scholarship, AI offers enormous opportunities for increasing scale, enabling discovery and improving accessibility. Yet the risks are concentrated in the areas of contextual loss, canon formation and reliability. The challenges centre on preserving authenticity and narrativity within a research environment increasingly shaped by patterns and algorithms. AI can enrich historical scholarship, but the legitimacy of historical knowledge production remains dependent on human interpretation and critical accountability.

References for Chapter 8

Borji, A. (2023). A categorical archive of ChatGPT failures. arXiv preprint arXiv:2302.03494.
Jockers, M. (2013). Macroanalysis: Digital Methods and Literary History. University of Illinois Press.

Kahle, P., Colutto, S., Hackl, G., & Fiel, S. (2017). Transkribus – A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. ICDAR 2017 Proceedings, 19–24.

Ketelaar, E. (2001). Tacit Narratives: The Meanings of Archives. Archival Science, 1, 131–141.

Putnam, L. (2016). The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast. American Historical Review, 121(2), 377–402.

Terras, M. (2011). Digital Curiosities: Resource Creation Via Amateur Digitisation. Literary and Linguistic Computing, 26(4), 425–438.

Underwood, T. (2019). Distant Horizons: Digital Evidence and Literary Change. University of Chicago Press.

Verhoeven, D. (2016). Doing Digital Humanities: An Extended Review. Digital Scholarship in the Humanities, 31(1), 1–15.

Wevers, M., & Smits, T. (2020). The politics of the machine: Digital humanities and the automation of text analysis. Digital Scholarship in the Humanities, 35(1), 194–214.

Zwierlein, C. (2016). Anachronism and Antiquarianism: Historical Knowledge and its Uses. Routledge.

Chapter 9 – Anthropological Research and AI-Driven Knowledge Production

9.1 Introduction

Following the analysis of historical scholarship, which focused on archives, canon formation and the interpretation of sources from the past, this chapter turns to a different archetype of knowledge production: anthropology. Unlike history, anthropology is not primarily concerned with documents or artefacts, but with living communities and their cultural practices. Whereas historical research depends on sources that already exist, anthropological research is methodologically embedded in fieldwork, participation and prolonged interaction with people in their social context.

This shift towards a research tradition that is relational and experiential means that the use of artificial intelligence (AI) generates different opportunities and risks here. In historical scholarship, the principal vulnerability lay in the representation of sources and the canonisation of particular perspectives. In anthropology, the central question shifts to the relationship between researcher, community and technology: how do the quality and legitimacy of knowledge production change when AI mediates between researchers and the people they study?

AI is already being used in this domain in a variety of ways. Digital ethnography employs AI to analyse patterns in social media. Speech and image recognition support the documentation of endangered languages and rituals. Remote sensing and geographical information systems (GIS) assist in the study of landscapes and communities. These applications offer undeniable new possibilities, but also create tensions. Are communities reduced to data points? Which power relations become embedded in the ways AI collects and structures information? And to what extent can a discipline that derives its legitimacy from proximity and empathy rely on a technology that instead creates distance?

This chapter examines the opportunities, risks and challenges that AI presents in anthropology. The central question remains whether, and under which conditions, AI-driven knowledge production within this archetype can be scientifically legitimate.

9.2 Opportunities

Digital ethnography and expansion of scale. Anthropological fieldwork has traditionally been small-scale and labour-intensive. A researcher spends months or years within a community, takes notes, conducts interviews and observes rituals. This method produces rich, contextual knowledge, but is limited in scale. AI makes it possible to overcome these limitations in part through digital ethnography: the study of online communities and social interactions on platforms such as Reddit, WhatsApp or Twitter. Pink et al. (2016) show in their standard work on digital ethnography how online communities create new fields of meaning that are as legitimate as physical communities. AI can analyse these digital spaces for patterns in language use, network structures and collective meaning-making. Varis and Blommaert (2015) demonstrate how social media generate new forms of conviviality and collective structure that can be made visible through AI-supported analysis. This expansion of scale means that anthropologists are no longer restricted to a single village, neighbourhood or community, but can examine transnational networks, diasporic communities or online subcultures that were previously methodologically inaccessible.

Transcription and thematic coding. A substantial part of anthropological fieldwork consists of transcribing interviews and coding observations. This is time-consuming and labour-intensive work. AI can accelerate this process considerably. Speech-recognition software can transcribe interviews automatically, while natural language processing can identify thematic patterns across large quantities of text. This does not mean that AI takes over the anthropologist's interpretative task, but it does make more time available for the analysis itself. In their work on ethnography in virtual worlds, Boellstorff et al. (2012) emphasise that technological tools do not replace the researcher, but do enable larger quantities of data to be processed without sacrificing interpretative depth.

Documentation of endangered languages and rituals. AI offers distinctive possibilities for linguistic-anthropological research. Endangered languages spoken by only a small number of people can be documented using speech recognition and audiovisual analysis. Kandel et al. (2019) show how GIS and AI-based landscape analysis can enrich anthropological perspectives on space and place. The same applies to rituals and ceremonies that are rarely performed and are at risk of disappearing. AI can help identify patterns in spoken language, gestures or musical structures that are difficult for human researchers to distinguish. This can contribute to the preservation of cultural heritage and to empowering communities seeking to document their own traditions and transmit them to future generations.

9.3 Risks

Loss of relational proximity. The greatest risk of AI in anthropology is the loss of relational proximity. Anthropological research traditionally rests on thick description (Geertz, 1973): detailed, contextual accounts of social practices that are possible only through prolonged participation. Hammersley and Atkinson (2019) emphasise in their classic work on ethnographic methods that the legitimacy of anthropological knowledge derives from the researcher's direct, embodied experience in the field. AI creates distance. An algorithm may identify patterns in social-media interactions, but it misses the nuances of body language, silences, irony and implicit meanings that can be perceived only through direct interaction. The risk is that anthropological knowledge becomes impoverished, reduced to statistics and correlations without the richness of contextual interpretation that characterises the discipline.

Reduction of communities to data points. A related risk is that communities are reduced to datasets. Whereas traditional anthropological research approaches people as complex, acting subjects, AI-driven analysis risks objectifying them as data points. This affects the ethical core of the discipline, which is grounded in respect for the autonomy and dignity of the communities being studied. Adams et al. (2020) warn in their work on AI-driven ethnography about the ethical implications of large-scale data extraction. When people's posts, photographs and conversations are analysed without their explicit consent, a form of surveillance emerges that is fundamentally at odds with the reciprocity and mutuality anthropologists seek to uphold.

Reproduction of colonial power relations. Historically, anthropology has been burdened by a colonial past in which Western researchers studied non-Western communities from an asymmetrical position of power. The discipline has critically reflected on this history and sought to develop more egalitarian research methods. AI risks reproducing these power relations. Who determines which data are collected, how they are analysed and what happens to the results? If AI systems are developed and controlled by Western technology companies and universities, a situation once again arises in which knowledge about communities is extracted without those communities themselves having ownership

or control. This is not only ethically problematic, but also undermines the legitimacy of the knowledge produced.

Consent and reuse of data. In traditional fieldwork, informed consent is a core principle: people must know that they are being studied and must agree to participate. In the context of AI, this becomes more complicated. Social-media data are often publicly accessible, but this does not automatically mean that users consent to their posts being analysed for research purposes. Moreover, datasets collected for one research purpose may be reused for other purposes without the original participants being aware of this. Widlok (2017) emphasises that reciprocity and consent are core values in anthropology and that these values come under pressure when AI enables data to be reproduced and reused on an unlimited scale.

9.4 Challenges

Methodological challenges: thick description versus pattern recognition. The methodological challenge for anthropology is how thick description and pattern recognition can be connected. AI is highly effective at identifying patterns in large datasets, but anthropological knowledge requires more: contextual interpretation, understanding of implicit meanings and attention to the uniqueness of specific situations. Hammersley and Atkinson (2019) emphasise that ethnography centres on understanding meaning in context. This is fundamentally different from identifying statistical correlations. The challenge is to use AI as a tool for organising and structuring data without losing the interpretative depth that makes anthropological knowledge legitimate. One possible solution lies in hybrid methods, in which AI is used for an initial exploration of large datasets, after which the anthropologist examines in greater depth particular patterns or anomalies identified by the algorithm. This nevertheless requires explicit methodological justification and transparency concerning the role played by AI at every stage of the research.

Institutional challenges: access and representation. Access to AI tools is distributed unequally. Western universities and research institutions often possess the resources needed to deploy advanced software and computing power, whereas researchers in the Global South or from the communities themselves frequently lack those resources. This reinforces existing inequalities in who produces knowledge and whose perspectives are represented. Pink et al. (2016) emphasise that digital ethnography should be accessible to diverse researchers, but in practice control over the technology remains concentrated among a small number of actors. The institutional challenge is to democratise AI tools and methodologies so that researchers in non-Western contexts can also use them on terms they determine themselves. There is also the challenge of representation in training data. AI models are often trained on Western languages and cultural contexts, which means they are less effective in analysing non-Western languages, dialects or cultural practices. This may produce systematic distortions in what is made visible and what remains invisible.

Normative challenges: reciprocity and collective ownership. The normative challenge for anthropology is how reciprocity and collective ownership can be safeguarded in a context of AI-driven knowledge production. Traditional fieldwork rests on a relationship of give and take: researchers gain access to a community and, in return, share knowledge, provide practical assistance or contribute to local projects. AI renders this model problematic. Data can be extracted and analysed without direct interaction or reciprocity. Widlok (2017) emphasises that consent should apply not only at the moment of data collection, but throughout the entire data lifecycle within an AI ecosystem. This means that communities must retain control over how their data are used, who has access to them and how the

results are shared. One possible solution lies in participatory design: involving communities themselves in the design of AI systems and research protocols. This nevertheless requires time, resources and a fundamental shift in the organisation of research — from extractive to collaborative.

9.5 Conclusion

The analysis of anthropology shows that AI not only offers this archetype new possibilities, but also directly affects the discipline's core values: proximity, context and relational responsibility. To maintain analytical precision, the three sub-questions are explicitly answered.

Sub-question 1: *How does AI change the cognitive and methodological dimension of knowledge production?* AI broadens the methodological horizon of anthropological research. Transcription and coding tools accelerate the processing of interviews and observations, digital ethnography enables large-scale analysis of online communities, and multimodal AI can systematically analyse audiovisual material (Adams et al., 2020; Varis & Blommaert, 2015). Yet this also presents the greatest challenge: the interpretative proximity developed by the anthropologist during fieldwork cannot be fully replaced by algorithmic patterns. Hammersley and Atkinson (2019) emphasise that ethnography revolves around thick description — detailed, contextual interpretation made possible only through prolonged participation.

The provisional answer is that AI expands the cognitive reach of anthropological research, but that the legitimacy of knowledge production remains dependent on hermeneutic interpretation by the researcher. Pattern recognition can provide an initial exploration, but the meaning of those patterns must always be constructed by the anthropologist in dialogue with the community.

Sub-question 2: *How does AI affect the pedagogical and institutional dimension of knowledge production?* Pedagogically, AI offers new possibilities for teaching anthropology. Through digital platforms, students can become acquainted with diverse communities and independently identify patterns in online interactions. Pink et al. (2016) show that digital ethnography enables students to conduct fieldwork in virtual spaces at their own pace. Institutionally, however, new inequalities emerge. Access to AI tools and computing power is distributed unevenly, benefiting Western universities over institutions in the Global South. There is also a risk that knowledge will be extracted from communities without those communities themselves retaining ownership or control over the results.

The provisional answer is that AI contributes pedagogically and institutionally to inclusion and accessibility, but that this is sustainable only when infrastructures and access are organised equitably. This requires the democratisation of AI tools and participatory design in which communities themselves are involved in the research process.

Sub-question 3: *How does AI reconfigure power relations, and what normative order is both desirable and feasible?* AI profoundly affects the power relationship between researcher and researched community. Whereas traditional anthropology rests on reciprocity and consent, AI can reproduce and reuse data indefinitely without the community retaining control over them. Widlok (2017) emphasises that reciprocity is a core value that must be systematically embedded in AI-driven research practices. Normatively, it is essential that consent apply not only at the moment of data collection, but throughout the entire data lifecycle. Communities must retain control over how their data are used, who has access to them and how results are shared. This requires new institutional arrangements centred on collective ownership and participatory design.

The provisional answer is that AI fundamentally reconfigures power relations in anthropological research, and that a legitimate normative order is possible only when reciprocity, consent and collective ownership are systematically safeguarded. This is not only an ethical obligation, but also a condition for the quality and reliability of anthropological knowledge.

In summary. For anthropology, the central question is how AI affects the relationship between researcher and community. AI can increase the reach and efficiency of research, but simultaneously risks undermining the proximity and empathy that make anthropological knowledge legitimate. The challenge is to use AI as a tool without eroding the discipline's relational foundations. This requires methodological rigour, institutional reform and a normative framework centred on reciprocity and collective ownership.

References for Chapter 9

Adams, O., Michaud, A., & Cohn, T. (2020). Massively Multilingual Adversarial Speech Recognition. Proceedings of ACL 2020, 368–379.
→ Relevance: illustrates how AI-supported speech recognition can assist in documenting small and endangered languages.

Boellstorff, T., Nardi, B., Pearce, C., & Taylor, T. L. (2012). Ethnography and Virtual Worlds: A Handbook of Method. Princeton University Press.
→ Relevance: foundational to digital ethnography; demonstrates how online communities constitute contexts for fieldwork.

Geertz, C. (1973). The Interpretation of Cultures. Basic Books.

Hammersley, M., & Atkinson, P. (2019). Ethnography: Principles in Practice (4th ed.). Routledge.
→ Relevance: a classic methodological work on the foundations of ethnography; provides an important contrast with AI-driven methods.

Kandel, A. W., Klein, A., & Wessel, M. (2019). Remote sensing and the anthropology of landscape. Journal of Anthropological Research, 75(2), 178–196.
→ Relevance: discusses how GIS and AI-based landscape analysis affect anthropological perspectives on space.

Pink, S., Horst, H., Postill, J., Hjorth, L., Lewis, T., & Tacchi, J. (2016). Digital Ethnography: Principles and Practice. Sage.
→ Relevance: a standard work on digital ethnography, directly applicable to AI applications in online fieldwork.

Varis, P., & Blommaert, J. (2015). Conviviality and collectives on social media: Virality, memes, and new social structures. Multilingual Margins, 2(1), 31–45.
→ Relevance: shows how social media create new fields of collective meaning to which AI-supported analysis can be applied.

Widlok, T. (2017). Anthropology and the Economy of Sharing. Routledge.
→ Relevance: emphasises reciprocity and consent as core values in anthropology; relevant to the normative dimensions of AI use.

Chapter 10 – Design-Oriented Research and AI-Driven Knowledge Production

10.1 Introduction

Whereas the natural sciences seek to explain, history seeks to understand the past, and anthropology seeks to comprehend living communities, design-oriented research centres on making. This archetype produces knowledge not through observation or interpretation alone, but by creating artefacts that both function in practice and generate theoretical insight. Herbert Simon (1996) characterised this process as 'the sciences of the artificial': knowledge arises through designing systems that change reality rather than merely describe it.

Computer science and the engineering sciences have long embraced this paradigm, but design-oriented research also plays an increasingly important role in domains such as education, healthcare and public administration. Research in these fields examines not only how reality is, but above all how it might function better. The value of design-oriented research therefore lies in the interaction between creation and reflection: artefacts are designed, tested, improved and theoretically grounded.

The advent of artificial intelligence is transforming this archetype profoundly. AI functions not only as a tool for simulating, analysing or optimising designs, but also as a co-designer that proposes solutions itself or automates entire design trajectories. The researcher's role consequently shifts from designer to curator: the human selects from and accounts for the multitude of AI-generated possibilities.

This shift raises fundamental questions. When an artefact has been designed partly by AI, who is its author? How can the quality of a design be assessed when the design process is not fully transparent? And under which conditions can an AI-generated design be regarded as scientifically legitimate? This chapter examines the opportunities, risks and challenges presented by AI in design-oriented research.

10.2 Opportunities

AI as co-designer and accelerator. One of the most promising applications of AI in design-oriented research is its role as co-designer. Generative AI systems can produce countless design variants on the basis of specified parameters. In architecture, algorithms can generate thousands of floor plans that meet particular requirements concerning spatial efficiency, natural light and accessibility. In product design, AI systems can optimise forms for strength, weight or aesthetics.

This considerably accelerates the design process. Whereas designers previously needed weeks or months to explore different options, AI can present a broad spectrum of possibilities within hours or days. Simon (1996) had already emphasised that design is a process of satisficing — searching for solutions that are 'good enough' given the constraints. AI expands the space within which this search can take place and makes it possible to identify solutions that a human designer might otherwise have overlooked.

Simulation and prototyping. AI enables advanced simulation. Before a physical prototype is constructed, a digital version can be tested under different conditions. In the engineering sciences, AI-driven simulations can predict the behaviour of materials, structures or circuits. In the social sciences, agent-based models can simulate how people respond to new policy interventions or urban designs.

Zimmerman et al. (2007) emphasise in their work on design research in HCI (Human–Computer Interaction) that prototyping is essential to the development and testing of design knowledge. AI makes it possible to intensify this process through more variants, faster iterations and more detailed feedback. This improves the quality of the final design and reduces the risk of costly errors.

Optimisation and innovation. AI can optimise existing designs by varying parameters and evaluating performance. In product design, this may produce forms that are more efficient, lighter or stronger than those conceived by a human designer. In urban planning, algorithms can optimise traffic flows or minimise energy consumption. More importantly, AI can contribute to innovation by generating unconventional solutions. In his classic work on the reflective practitioner, Schön (1983) describes how designers learn through experimentation and reflection on unexpected outcomes. AI can enrich this process by proposing solutions that fall outside the designer's conventional patterns of thought, potentially leading to breakthroughs in design and engineering.

10.3 Risks

Black-box designs and loss of transparency. The greatest risk of AI in design-oriented research is the emergence of black-box designs: artefacts whose origins are unclear. When a generative algorithm produces a form that meets all specified requirements but cannot explain the logic underlying that form, a fundamental problem arises for the scientific legitimacy of the design. Design-oriented research requires design choices to be accountable. In their influential article on design science, Hevner et al. (2004) argue that artefacts must not only work, but also contribute to theoretical understanding. This requires transparency concerning the design process: why was this form selected rather than another? Which trade-offs were made? In AI-generated designs, such transparency is often absent.

Loss of craft and intuition. A related risk is that the role of human intuition and craft may be eroded. In his classic work on user-friendly design, Norman (2013) emphasises that good designs often emerge from years of experience and a developed sense of what works. When AI assumes much of the design work, this expertise risks becoming impoverished. This entails not only a loss of craftsmanship, but also a loss of knowledge transmission. Students who learn to design with AI as a co-designer may not develop the same depth of understanding of design principles as students required to work everything out themselves. The pedagogical challenge is to use AI without undermining the formation of future designers.

Ownership and intellectual property rights. When an artefact has been designed partly by AI, the question arises as to who its author is. Legally and institutionally, this is problematic. Can a scientist claim a patent for a design generated largely by an algorithm? Who is responsible if an AI-generated design fails or causes harm? Sanders and Stappers (2008) emphasise in their work on co-creation that design processes are increasingly collaborative, with multiple actors contributing. AI adds a new layer: the technology is not a passive tool but an active contributor. This requires new legal and institutional frameworks for assigning ownership and responsibility.

Homogenisation and lack of diversity. A further risk is that AI-generated designs become homogeneous. Algorithms are trained on existing designs and may therefore tend to remain within familiar patterns. This can result in a lack of diversity and cultural specificity in design. Rittel and Webber (1973) introduced the concept of wicked problems: complex design problems without a single definitive solution, in which the quality of the response depends on the values and perspectives adopted. AI systems are often unable to accommodate this plurality of values and may therefore generate solutions that appear technically optimal but are culturally or socially undesirable.

10.4 Challenges

Methodological challenges: the validity of design knowledge. The methodological challenge for design-oriented research is how to safeguard the validity of AI-generated designs. Hevner et al. (2004) formulate seven guidelines for design science research, including the requirement that artefacts be evaluated rigorously. But how can a design be evaluated when its design process is not transparent? One possible solution lies in 'adversarial testing': systematically testing AI-generated designs against critical criteria and explicitly documenting weaknesses and limitations. This nevertheless requires time and resources and conflicts with the speed at which AI can produce designs. The tension between efficiency and rigour is acute here. Schön (1983) emphasises that reflection is essential to designers' learning. The challenge is to integrate this 'reflective practice' into AI-driven design processes, so that researchers not only produce designs but also learn from the process and contribute to theoretical knowledge.

Institutional challenges: publication and recognition. Institutionally, the question arises of how AI-generated designs should be published and recognised. Traditional scientific publications require the methodology and design choices to be accounted for. This is often problematic in AI-generated designs: how can a process be described when it is partly algorithmic and partly inexplicable? The question also arises whether existing peer review mechanisms are suitable for assessing AI-generated work. Reviewers must assess not only the artefact itself, but also the process and the role played by AI within it. This requires new competencies and possibly new publication formats in which accountability for the process takes centre stage. Sanders and Stappers (2008) emphasise that co-creation requires new forms of documentation. The same applies to AI as co-designer: the output alone is insufficient; the process must be documented transparently so that others can assess how the design was produced and what role AI played within it.

Normative challenges: ethics and values in design. The normative challenge is how values and ethics can be embedded explicitly in AI-driven design processes. Norman (2013) emphasises that design is always normative: every design contains implicit assumptions about how users ought to behave and which goals are valuable. When AI generates designs, these normative choices are often made implicitly through the parameters specified by the designer or through the training data on which the algorithm was trained. This can result in designs that are technically optimal but ethically problematic — for example, systems that are efficient but infringe privacy, or products that are inexpensive but environmentally damaging. Rittel and Webber (1973) emphasise that wicked problems require participatory design involving diverse stakeholders. The challenge is to integrate this participation into AI-driven processes so that values and ethics are explicitly articulated and embedded in the design.

10.5 Conclusion

The analysis of design-oriented research shows that AI offers enormous opportunities within this archetype while also introducing fundamental tensions. To formulate the findings precisely, the three sub-questions are explicitly answered for this domain.

Sub-question 1: *How does AI change the cognitive and methodological dimension of knowledge production?* AI changes the cognitive practice of design-oriented research by shifting the designer's role from maker to curator. Whereas designers previously conceived and developed forms themselves, AI now generates numerous variants from which selections must be made. Simon (1996) described design as a process of satisficing; AI expands the search space but simultaneously introduces the risk of black-box solutions. Methodologically, the challenge is to safeguard transparency and accountability. Hevner et al. (2004) emphasise that design science artefacts must be evaluated rigorously and contribute to theoretical understanding. In the case of AI-generated designs, this is possible only when the design process is explicitly documented and the role of AI clearly articulated. The provisional answer is that AI expands designers' cognitive capacity by enabling acceleration, optimisation and innovation, but that methodological legitimacy remains dependent on transparency concerning the design process and reflective accountability on the part of the researcher.

Sub-question 2: *How does AI affect the pedagogical and institutional dimension of knowledge production?* Pedagogically, AI offers new possibilities for teaching design. Students can experiment with advanced tools and iterate rapidly between idea and prototype. Schön (1983) emphasises that designers learn through reflection on their own actions; AI can accelerate this process by providing immediate feedback on design choices. Yet there is a risk that students will become dependent on AI without understanding the underlying principles. Norman (2013) warns that good design emerges from experience and intuition — qualities that can be developed only through making and failing oneself. Institutionally, the question arises of how AI-generated work should be published and recognised. Existing publication formats are not suited to documenting hybrid design processes in which humans and machines collaborate. Sanders and Stappers (2008) emphasise that new forms of documentation are needed in which accountability for the process takes centre stage. The provisional answer is that AI can be pedagogically valuable provided that the formation of designers is not hollowed out, and that institutional reforms are required to recognise and facilitate hybrid authorship.

Sub-question 3: *How does AI reconfigure power relations, and what normative order is both desirable and feasible?* Power manifests itself in control over AI tools and training data. Large technology companies and well-funded universities have access to advanced generative systems, while smaller institutions and designers in the Global South often lack such resources. This reinforces existing inequalities in who is able to design and whose perspectives are represented. Normatively, it is essential that values and ethics be embedded explicitly in AI-driven design processes. Rittel and Webber (1973) emphasise that wicked problems require participatory design involving diverse stakeholders. This becomes even more urgent when AI generates designs: without explicit articulation of values, artefacts may emerge that appear technically optimal but are ethically or culturally problematic. The provisional answer is that AI reconfigures power relations in design-oriented research by concentrating access to resources, and that a legitimate normative order requires transparency, participation and the explicit articulation of values to occupy a central place in the design process.

In summary. For design-oriented research, AI offers enormous opportunities as a co-designer and accelerator, but also introduces fundamental tensions concerning transparency, authorship and ethics. The challenge is to use AI without eroding the reflective practice and craft expertise that characterise this archetype. This requires methodological rigour through transparent documentation, institutional

reforms through new publication formats, and normative explication through the incorporation of values into the design process. AI can enrich design-oriented research, but only when humans retain curatorial oversight and assume responsibility for the choices made. The curator's role is crucial here: not blindly trusting AI output, but critically selecting, evaluating and accounting for it.

References for Chapter 10

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS Quarterly, 28(1), 75–105.

Norman, D. A. (2013). The Design of Everyday Things: Revised and Expanded Edition. Basic Books.
Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. Policy Sciences, 4(2), 155–169.

Sanders, E. B.-N., & Stappers, P. J. (2008). Co-creation and the new landscapes of design. CoDesign, 4(1), 5–18.

Schön, D. A. (1983). The Reflective Practitioner: How Professionals Think in Action. Basic Books.

Simon, H. A. (1996). The Sciences of the Artificial (3rd ed.). MIT Press.

Zimmerman, J., Forlizzi, J., & Evenson, S. (2007). Research through design as a method for interaction design research in HCI. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 493–502.

Chapter 11 – The Meta-Domain of Computer Science (AI) and AI-Driven Knowledge Production

11.1 Introduction

In the preceding chapters, the use of artificial intelligence (AI) was examined within separate archetypes of science: the natural sciences, history, anthropology and design-oriented research. In each case, AI was shown to change the conditions of knowledge production, but the analysis consistently began from existing disciplines. Another perspective is also possible, however: AI as a research domain in its own right.

Computer science and AI have developed into autonomous fields of knowledge with their own methods, institutions and publication cultures. At the same time, AI is more than merely one archetype alongside the others: it is also the infrastructure through which knowledge production proceeds in other domains. This dual character makes AI a meta-domain: it investigates both reality and itself. Recent developments such as automated machine learning (AutoML), in which algorithms develop new algorithms, or AI that optimises chip architectures for future AI models (Mirhoseini et al., 2021), demonstrate how reflexive this domain has become.

This chapter analyses what AI-driven knowledge production looks like when AI is both instrument and object. This reflexivity makes the meta-domain a critical test case: here, the tensions that remained implicit in other archetypes become explicit and constitutive. If AI can investigate and validate itself, what does this mean for the legitimacy of knowledge?

11.2 Opportunities: The Utopian Horizon

AutoML and self-improving systems. One of the most spectacular developments in AI is the emergence of AutoML: systems that automatically design and optimise new machine-learning models (He et al., 2021). This represents a fundamental shift: AI is used not only to solve problems, but also to improve itself. Vilalta and Drissi (2002) introduced the concept of meta-learning, in which algorithms learn from their own learning processes. This self-improvement may enable exponential progress. Whereas human researchers previously required months or years to develop a new model, AutoML can test and optimise thousands of architectures within days or weeks. This creates the possibility of a positive feedback loop: better AI systems can design still better AI systems.

Open platforms and democratisation. The meta-domain also contains a powerful movement towards openness. Platforms such as arXiv make it possible to publish research without the lengthy turnaround times of traditional peer review. Open-source libraries such as TensorFlow and PyTorch enable researchers worldwide to build state-of-the-art models without depending on proprietary software. Wikipedia functions as a distinctive example of an alternative validation model. Whereas traditional science relies on hierarchical peer review, Wikipedia operates through horizontal, collective correction. Errors are rapidly identified and corrected by a community of volunteers. This model is not perfect, but it demonstrates that legitimacy can also arise from communities rather than exclusively from formal institutions. Such openness can democratise knowledge production. Researchers in the Global South, small universities and even individual enthusiasts gain access to tools and knowledge that were previously reserved for an elite of well-funded institutions.

New pedagogy: learning with and about AI. The meta-domain enables new forms of education. Students can use AI to learn about AI: a language model as tutor, a code generator as assistant. This may lead to adaptive education in which content and pace are tailored to the individual student. Moreover, AI itself can become an object of study in ways that were previously impossible. Students can experiment with training processes, architectures and datasets and immediately observe how their choices affect performance. This makes abstract theoretical education concrete and interactive.

11.3 Risks: The Dystopian Horizon

Circular validation and internal benchmarks. The greatest risk of the meta-domain is circular validation: AI is tested on tasks shaped by AI. Benchmarks such as ImageNet or GLUE were designed to measure model performance, but are now themselves partly optimised by the models (Raji et al., 2021). This creates a form of self-confirmation in which progress is defined by criteria determined by the same technology. The question is whether this still possesses external validity. A model that performs exceptionally well on ImageNet does not automatically perform well in real-world applications. The gap between benchmark performance and practical usefulness may be considerable, but is obscured by the focus on internal validation.

Concentration of power and oligopoly. Although open platforms exist, the development of state-of-the-art AI remains highly concentrated among a small number of major actors: Google DeepMind, OpenAI, Meta and Anthropic. This concentration is not accidental but structural: developing foundation models requires enormous investments in computing power, data and talent that only large companies can afford. In their influential paper ‘Stochastic Parrots’, Bender et al. (2021) warn of the dangers of this concentration of power. Those who control the infrastructure determine which research questions are asked, which datasets are used and which norms are presented as ‘neutral’. This is not only an economic problem, but also an epistemic one: knowledge is produced by a small elite that sets the agenda.

Black boxes in extremis. In other domains, the black-box character of AI is problematic; in the meta-domain, it becomes existential. Deep-learning models produce results that even their designers can no longer fully explain. Castelvecchi (2016) asks whether the black box of AI can ever be opened, or whether opacity is a fundamental characteristic of the technology. Interpretability continues to lag behind performance: as long as a model scores better on benchmarks, it is regarded as progress, even when its internal operation remains largely unknown. This raises the fundamental question of whether knowledge production that cannot be explained, but nevertheless performs, can still be regarded as legitimate science.

Fragmentation of legitimacy. The meta-domain has no uniform standard of validation. arXiv preprints are assessed differently from peer-reviewed articles. Open-source projects apply different quality criteria from industrial laboratories. Wikipedia relies on collective correction, while scientific journals employ hierarchical review. This plurality can be productive, but also risks fragmentation. What happens when different validation practices produce contradictory conclusions? Who then determines what is valid? Traditional science possessed a degree of consensus regarding its criteria; in the meta-domain, that consensus is absent.

11.4 Challenges

Methodological challenges: redefining the criteria. The methodological challenge is to develop criteria that break the cycle of circular validation and guarantee external testability. It is not sufficient for a model to perform well on internal benchmarks; it must also be robust across diverse real-world contexts. Raji et al. (2021) advocate transparent and reproducible AI evaluation in which not only performance, but also limitations, biases and failure modes are documented. This requires a cultural shift: publishing not only success stories, but also systematic analyses of where and why models fail.

Institutional challenges: balancing open and closed systems. Open platforms such as arXiv and Wikipedia demonstrate that collective validation and knowledge sharing can work, but remain vulnerable in relation to closed industrial systems that control crucial infrastructure. The institutional challenge is to find a balance: how can open validation be combined with robust, controlled standards that counteract the concentration of power? One possible direction lies in public–private partnerships through which access to computing power and data is democratised while quality standards are determined collectively. This nevertheless requires political will and institutional reforms extending beyond what individual researchers can achieve.

Normative challenges: responsibility in a recursive ecosystem. When AI models design and validate other AI models, the question of who is responsible for the outcome becomes blurred. For AI-driven knowledge production, this means that responsibility must be explicitly recalibrated: it rests not only with researchers, but also with developers, institutions and communities that shape the systems. In their analysis of values in machine-learning research, Birhane et al. (2022) show that normative choices often remain implicit. The challenge is to make these choices explicit and document them transparently, so that others can assess whether the values embedded in systems are acceptable.

11.5 Qualification: Counterexamples and Alternatives

The meta-domain need not be viewed exclusively in dystopian terms. Powerful counterexamples demonstrate that alternative forms of knowledge production are possible.

Wikipedia is the most striking example. It shows that horizontal, collective validation can work and that legitimacy need not derive exclusively from hierarchical institutions. Millions of articles are maintained by volunteers who correct errors, add sources and debate controversial issues. The system is not perfect — biases exist in what is documented and in who contributes — but it serves as living proof that knowledge production can be organised without central control.

Open-source communities such as those surrounding TensorFlow, PyTorch and Hugging Face demonstrate that collective contributions can produce high-quality software. Code is shared openly, tested and improved by thousands of developers worldwide. Quality emerges not through formal certification, but through use, feedback and iteration.

arXiv democratises access to scientific knowledge. Researchers can share their work immediately without waiting for lengthy peer review. This accelerates the circulation of knowledge and makes science more accessible. There are, of course, risks associated with unreviewed claims, but the transparency of the system allows readers to assess reliability for themselves.

These counterexamples make clear that the future of AI-driven knowledge production is not predetermined. Both utopia and dystopia remain possible; the direction taken depends on the choices made now.

11.6 Conclusion

The meta-domain of AI and computer science is distinctive because AI is both an instrument and an object of knowledge production. This makes the legitimacy of AI-driven knowledge production precarious, but also particularly instructive. The three sub-questions can be provisionally answered for this domain as follows.

Sub-question 1: *How does AI change the cognitive and methodological dimension of knowledge production?* AI makes knowledge production reflexive: models develop, test and improve one another within frameworks they have themselves helped to shape. This leads to circular validation, in which progress is defined by benchmarks optimised by the same technology (Raji et al., 2021). Cognitively, this offers speed and scale, but methodologically it raises questions about external testability. Legitimate knowledge production therefore requires new criteria, in which performance alone is no longer sufficient without explainability and robustness beyond the system's own frameworks. Castelvecchi (2016) asks whether the black box can ever be opened; the provisional answer is that we must at least strive for transparency concerning limitations and failure modes. The provisional answer is that AI fundamentally reshapes cognitive practice through reflexivity, but that methodological legitimacy can be safeguarded only when circular validation is broken through external testing and transparent documentation.

Sub-question 2: *How does AI affect the pedagogical and institutional dimension of knowledge production?* Pedagogically, AI is radically transforming education. Students increasingly learn about AI through AI itself: a language model as tutor and a code generator as assistant. This may make learning more adaptive, but also increases the risk that critical distance will be lost. He et al. (2021) show how AutoML makes it possible to build models without in-depth knowledge; the question is whether this constitutes empowerment or intellectual dependency. Institutionally, a hybrid field is emerging: open platforms such as arXiv and open-source communities promote speed and accessibility, while industrial laboratories monopolise infrastructure and data. Wikipedia provides an instructive counterexample in which validation takes place through a horizontal, collective correction mechanism. It demonstrates that legitimacy can also arise from communities rather than exclusively from hierarchical institutions or private actors. The provisional answer is that AI offers considerable pedagogical potential provided that critical formation is not hollowed out, and that institutional reforms are needed to establish a balance between open and closed systems, speed and rigour, and democratisation and quality assurance.

Sub-question 3: *How does AI reconfigure power relations, and what normative order is both desirable and feasible?* The power structure in this domain is highly concentrated. Access to data, computing power and talent lies largely with a small group of actors that also formulate the norms of transparency and fairness (Bender et al., 2021). This creates a recursive power relationship in which resources and rules coincide. Birhane et al. (2022) show that values in machine-learning research often remain implicit. A desirable normative order requires mechanisms that explicitly decentralise power and make accountability enforceable. This can be operationalised through: 1). Transparent interpretation: AI results must be accompanied by explicit accounts of choices and limitations; 2) Reflexive positioning: researchers must make their own position within the human–machine network explicit, in line with autoethnographic and reflexive methodology; 3) Network tracing: in the spirit of Actor–Network Theory, all actors contributing to knowledge production — datasets, algorithms, infrastructures, developers and users — must be made visible.

The provisional answer is that AI renders power relations within the meta-domain recursive, and that a legitimate normative order requires transparency, reflexivity and network tracing to be systematically embedded in research practices and institutional arrangements.

In summary. The meta-domain brings together the tensions of all the archetypes and simultaneously serves as a testing ground. It makes clear that legitimacy does not arise automatically, but must be actively organised. Both utopia and dystopia remain possible, and the future of AI-driven knowledge production will depend on the extent to which new norms are structurally embedded. Wikipedia, open-source communities and arXiv demonstrate that alternative validation practices are possible. This offers hope, but no guarantee. The challenge is to strengthen and institutionalise these counterexamples so that the democratisation of knowledge production becomes a reality rather than remaining a utopian aspiration. The meta-domain is not only a mirror of the other archetypes, but also a laboratory in which the future of science is being shaped. The choices made here — concerning openness, validation, power and responsibility — will determine the legitimacy of AI-driven knowledge production across all domains.

References for chapter 11

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 610–623. https://doi.org/10.1145/3442188.3445922

Birhane, A., Prabhu, V. U., & Kahembwe, E. (2022). The values encoded in machine learning research. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, 173–184. https://doi.org/10.1145/3531146.3533083

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature*, 538(7623), 20–23. https://doi.org/10.1038/538020a

He, X., Zhao, K., & Chu, X. (2021). AutoML: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212, 106622. https://doi.org/10.1016/j.knosys.2020.106622

Mirhoseini, A., et al. (2021). A graph placement methodology for fast chip design. *Nature*, 594(7862), 207–212. https://doi.org/10.1038/s41586-021-03544-w

Raji, I. D., Buolamwini, J., & Gebru, T. (2021). AI benchmarking: Towards transparent and reproducible AI evaluation. *arXiv preprint*, arXiv:2102.05623. https://arxiv.org/abs/2102.05623

Vilalta, R., & Drissi, Y. (2002). A perspective view and survey of meta-learning. *Artificial Intelligence Review*, 18(2), 77–95. https://doi.org/10.1023/A:1019956318069

Chapter 12 – A Second Interim Assessment

12.1 Introduction – Interim Assessment 2

The preceding chapters have established a broad basis for answering the research questions. Chapters 4 to 6 first examined the three central impact dimensions of artificial intelligence: the cognitive and methodological dimension, the pedagogical and institutional dimension, and the dimension of power and normativity. These dimensions were subsequently applied to four archetypes of science — the natural sciences, history, anthropology and design-oriented research — while Chapter 11 presented the meta-domain of AI and computer science as a distinctive mirror that brings together and intensifies the tensions found across all the archetypes.

It is now time to take stock for a second time. Interim Assessment 1, following the discussion of the three impact dimensions, identified the opportunities, risks and challenges raised by AI-driven knowledge production in general. Interim Assessment 2, the present chapter, goes one step further by integrating the insights derived from the dimensions and archetypes. This produces an analytical framework that reveals the core patterns occurring across the full breadth of science and technology.

This interim assessment is deliberately not an endpoint. The strength of academic rigour lies precisely in building conclusions in stages. The implications of the dimensions and archetypes are first considered in relation to one another; only then are they applied to concrete cases. Chapters 13 to 17 introduce five such cases: five academic monographs written as dissertations. They constitute the practical test of the framework developed here.

The purpose of this chapter is therefore not to answer the main research question definitively, but to make the contours of an integrated answer visible. Interim Assessment 2 outlines the overarching patterns, systematically compares the archetypes and provisionally answers the three sub-questions. It thereby lays the foundation for the cases and, ultimately, for the definitive synthesis in Chapter 18.

12.2 Overarching Patterns and Distinguishing Characteristics

The analysis of the four archetypes and the meta-domain reveals several patterns that recur across multiple domains, while also showing that each archetype has its own distinctive tensions and possibilities. It is important to emphasise both sides: shared tendencies must not obscure the fact that each archetype possesses its own logic and vulnerabilities.

12.2.1 Commonalities: Universal Paradoxes

The analysis of the five domains reveals three fundamental paradoxes that occur across all archetypes. These paradoxes cannot be resolved, but only managed: they are inherent in AI's dual nature as both an accelerator and a force that restructures knowledge production.

Paradox 1: AI as an accelerator and a threat to legitimacy. The most fundamental paradox is that AI functions as an accelerator of knowledge production in virtually every domain, while simultaneously placing the legitimacy of that knowledge under pressure. This paradox manifests itself differently in each archetype:

1. Natural sciences: AlphaFold accelerates protein-structure prediction by a factor of 1,000, but its predictions are difficult to validate without experimental confirmation. Greater speed here means more hypotheses, but not automatically more reliable knowledge.
2. History: Transkribus can search within days archives that previously required years of work, but the patterns identified by AI may be anachronistic — relationships that had no historical meaning. Speed increases the risk of contextual loss.
3. Anthropology: Digital ethnography makes it possible to analyse thousands of social-media interactions, whereas traditional fieldwork is limited to a few dozen interviews. Yet this scale comes at the expense of the thick description that underpins anthropological legitimacy.
4. Design-oriented research: Generative AI can produce thousands of design variants within hours, but without transparency concerning the design process, it becomes difficult to account for the choices made. Speed undermines reflective practice.
5. Meta-domain: AutoML can develop within days models that previously required months, but when AI validates itself through benchmarks optimised by AI, circular validation arises without external testing.

The acceleration is undeniably valuable — more patterns, broader datasets and faster iterations — but precisely these gains undermine the mechanisms on which legitimacy has traditionally rested. Reproducibility, peer review and hermeneutic interpretation require time; AI imposes a choice between speed and rigour that cannot be resolved, but can only be managed through explicit accountability.

Paradox 2: Autonomy versus dependency. A second paradox is that AI makes scientists more autonomous by reducing the time spent on routine tasks and creating more scope for interpretation and synthesis, while simultaneously making them more dependent on technological infrastructures beyond their control. This manifests itself in each archetype as follows:

1. Natural sciences: Researchers can conduct more complex experiments through AI-driven simulations, but depend on cloud platforms such as AWS and Google Cloud and on proprietary software whose operation is not transparent.
2. History: Digital archives make sources accessible worldwide, but historians depend on digitisation projects that determine which archives become available. European collections dominate, while local sources from former colonies remain underrepresented.
3. Anthropology: AI-supported transcription gives anthropologists more time for analysis, but makes them dependent on systems trained on dominant languages and cultural contexts. Minority languages and dialects remain difficult to process.
4. Design-oriented research: Designers can iterate more rapidly using generative AI, but depend on algorithms whose decision-making logic is opaque. The designer becomes a curator, but loses control over the process.
5. Meta-domain: Researchers can build state-of-the-art models using open-source libraries, but the development of foundation models remains confined to a small number of large companies with enormous computing resources. Autonomy and dependency coexist.

AI expands the possibilities available to individual researchers — autonomy — while simultaneously concentrating control over infrastructure among a small number of actors — dependency. This is not a

temporary problem that can be solved; it is structural. The democratisation of tools coincides with the oligopolisation of infrastructure.

Paradox 3: Inclusion versus exclusion. A third paradox concerns the promise that AI democratises knowledge production while simultaneously creating new forms of exclusion. This takes the following forms in each archetype:

1. Natural sciences: Open-access publications and preprint servers such as arXiv lower barriers, but access to computing power for simulations remains unevenly distributed. Small universities in the Global South cannot compete with well-funded Western laboratories.

2. History: Digitisation makes archives accessible to researchers who lack the funds for physical archival visits, but reproduces power relations: European perspectives are digitised and therefore visible, while non-European sources remain out of view.

3. Anthropology: Digital ethnography enables fieldwork without physical presence, making it accessible to researchers with disabilities. Yet AI tools are trained on Western datasets and languages, leaving non-Western communities systematically underrepresented.

4. Design-oriented research: Generative AI tools such as DALL-E and Midjourney appear accessible to everyone, but obtaining the best results requires advanced prompt-engineering skills and an understanding of design principles. The 'democratisation' remains superficial.

5. Meta-domain: Open-source libraries such as PyTorch make AI development accessible, but training foundation models requires millions of dollars in computing resources. Knowledge sharing is open, but infrastructure remains closed.

AI lowers some barriers — access to tools and digital archives — while raising others, including technical expertise, computing power and representation within training data. AI is therefore not simply more inclusive or more exclusive; rather, it draws new dividing lines that supplement rather than replace the old ones. Those who were already privileged benefit most, while those who were already marginalised fall further behind.

Synthesis of the paradoxes. These three paradoxes are universal because they arise from AI's dual nature: on the one hand, it is a technology that increases capacity; on the other, it is an infrastructure that creates new dependencies and inequalities. They cannot be resolved through better technology or more intelligent implementation, but can only be managed through explicit recognition and institutional safeguards. The paradoxes demonstrate that AI is not a neutral amplifier, but a transformative force that redraws the conditions of knowledge production. This makes discussions of 'responsible AI use' complex: the issue is not avoiding risks, but managing inherent tensions that will not disappear.

12.2.2 Differences: Domain-Specific Tensions

Whereas the paradoxes are universal and emphasise the similarities between the four-plus-one domains, several domain-specific differences can also be identified.

1. **Natural sciences:** relatively unproblematic integration. In the natural sciences, AI has been integrated relatively unproblematically because the methodological norms of this archetype — reproducibility, transparency of experiments and peer review — largely remain intact. Dedicated AI applications such as AlphaFold have demonstrated that AI can discover patterns

that previously appeared impossible to resolve without undermining the epistemic foundations of the discipline. The risks are largely confined to dependency on infrastructure and potential bias in datasets, but these are often mitigated by robust experimental standards.

2. **History:** authenticity under pressure. In historical scholarship, the central tension lies in preserving authenticity and context. AI can search large archives and reveal unexpected connections, but also risks interpreting them anachronistically. A model recognises patterns, but does not understand their historical meaning. Human hermeneutics therefore remains indispensable, and the challenge is to use AI without losing the interpretative proximity that safeguards the legitimacy of historical knowledge.

3. Anthropology: preserving relational context. In anthropology, the central challenge is that relational proximity and contextual richness are difficult to automate. AI can accelerate transcription and support thematic coding, but lacks the embodied experience of fieldwork. The distinctive question is not so much whether AI is efficient, but whether the core of anthropological research — participation, reciprocity and proximity — can be preserved when technology mediates between researcher and community.

4. **Design-oriented research**: accounting for creativity. Design-oriented research presents the sharpest tension between opportunities and risks. AI can act as a co-designer that generates numerous variants and accelerates iteration. At the same time, the core of this archetype — explicitly accounting for design choices — comes under pressure. Artefacts may be efficient but difficult to explain. The legitimacy of design knowledge depends on transparency and accountability, precisely the qualities that AI problematises.

5. **Meta-domain:** circular validation as a constitutive problem. The meta-domain is unique because instrument and object coincide. AI investigates and validates itself, intensifying all the tensions identified above. Circular validation, concentration of power and the black-box character are not ancillary risks here, but constitutive features of knowledge production. What remains implicit in the other archetypes becomes explicit and unavoidable in this domain.

12.2.3 Comparing the Archetypes: Opportunities, Risks and Challenges

To provide a clear view of the differences and similarities, the five domains are systematically compared here across the dimensions of opportunities, risks and challenges:

Archetype	**Opportunities**	**Risks**	**Challenges**
Natural sciences	• Dedicated AI (AlphaFold)• Accelerated data processing• Robust reproducibility	• Dependency on infrastructure (cloud services, chips)• Bias in datasets	• Safeguarding openness and reproducibility despite the concentration of power
History	• Accessible digital archives• Pattern recognition in sources•	• Anachronism• Loss of authenticity and context• Canon formation through digitisation	• Preserving hermeneutic interpretation and critical contextualisation

Archetype	Opportunities	Risks	Challenges
	Opening up dispersed collections		
Anthropology	• Accelerated transcription and coding• Analysis of larger datasets• Digital ethnography	• Loss of relational proximity• Reduction of context to patterns• Ethical problems concerning consent	• Balancing the advantages of scale with the preservation of contextual richness
Design-oriented research	• AI as co-designer• Accelerated iteration• Simulation and prototyping	• Black-box designs• Loss of accountability• Diffuse ownership	• Developing standards for explainability• New forms of publication and recognition
Meta-domain	• AutoML and meta-learning• Self-improvement• Open platforms (arXiv, Wikipedia)	• Circular validation• Concentration of power• Black boxes in extremis• Fragmentation of legitimacy	• Formulating new standards (transparent interpretation, reflexive logs and network tracing)

This matrix shows that AI does not exert a uniform influence. The natural sciences exhibit relatively little tension between opportunities and risks, whereas history and anthropology struggle with loss of context and interpretation. Design-oriented research and the meta-domain call the legitimacy of knowledge production itself into question.

12.4 From Analysis to Case Studies

The preceding analysis makes clear that generative AI creates similar tensions across all archetypes of science, but that these tensions manifest themselves at different points in the research process. The three paradoxes — acceleration versus legitimacy, autonomy versus dependency, and inclusion versus exclusion — prove to be universal, but their concrete significance is domain-specific. At the same time, the comparison shows that the analytical framework of dimensions and archetypes provides only a partial answer to the research questions.

The framework reveals patterns, but cannot determine how these tensions actually develop within concrete research practices. Three questions in particular remain unresolved. First, it is not yet clear to what extent generative AI remains primarily a tool within the different archetypes, or becomes a technology that helps shape the research process itself. Second, it remains uncertain whether the differences between archetypes will prove as pronounced in practice as the theoretical analysis suggests, or whether the use of AI will instead produce convergence between research practices. Third, it is unclear how the normative tensions surrounding legitimacy, authorship and transparency manifest themselves when AI is actually used to produce academic texts.

These questions mark the limits of the theoretical analysis while simultaneously providing the rationale for the empirical phase of this research. The following chapters therefore examine how AI-supported

knowledge production manifests itself within concrete research practices. Five cases are considered: the monographs/dissertations on ‘avian intelligence & homing pigeons’ (natural science archetype), ‘predicting digital transformations’ (historical archetype), ‘governing within biological boundaries’ (anthropological archetype), ‘polyglot language acquisition in the AI era’ (design-oriented archetype), and ‘AI-driven knowledge production’ (the present dissertation — the mise en abyme representing the meta-domain). The outstanding questions formulated above are answered on the basis of findings from these five distinct research practices. In addition to the texts of the dissertations/monographs, the analysis will draw on the reflection reports describing how the respective works were produced.

These cases do not function as illustrations of the theory, but as an initial empirical test of the framework developed here. They reveal how generative AI is actually used within different archetypes of science and what implications this has for the structure of knowledge production. Only in light of these experiences can it be assessed to what extent the analytical framework of dimensions and archetypes withstands empirical scrutiny.

Chapter 13 – Case Study: ‘Avian Intelligence & Homing Pigeons’

This chapter examines the natural science archetype of AI-driven knowledge production through a concrete case: an experimental research programme on navigational and classification abilities in pigeons, presented in the draft dissertation Columba Loquens – Pattern Recognition, Navigational Robustness and the Limits of Explanation in Avian Intelligence.

The experiment addresses a classical question within behavioural biology and comparative cognitive science: how robust is animal behaviour under variation in stimulus structures and information sources, and which explanatory models can plausibly account for such performance? The research field is consequently characterised by a situation in which behaviour can be described empirically with considerable precision, while the underlying explanatory structure remains open to multiple theoretical interpretations. It is precisely in such situations that the formulation and comparison of possible explanatory models play a central role in the scientific process.

13.1 Positioning of the Case

Research into navigational and classification abilities in pigeons belongs to a long tradition within behavioural biology and comparative cognitive science. Since the mid-twentieth century, the homing pigeon has been regarded as a model organism for the study of animal navigation and spatial behaviour. Experimental studies have shown that pigeons are able to find their way back to their home loft under a wide range of conditions, often across considerable distances and under highly variable environmental circumstances.

The mechanisms explaining this behaviour have been the subject of a longstanding scientific debate. Various mechanisms have been proposed in the literature, including sun-compass orientation, magnetic detection, olfactory navigation and visual recognition of landscape features. Modern interpretations generally assume that these mechanisms do not operate exclusively, but function as components within a broader system in which multiple sources of information can be used simultaneously.

Alongside navigation research, the study of pigeons has also played an important role in research into animal categorisation and pattern recognition. Experimental studies show that pigeons can perform complex visual discriminations and learn to recognise categories across varying stimulus sets. These findings have generated an extensive literature on stimulus generalisation, classification ability and categorisation in non-human organisms.

From an empirical perspective, therefore, relatively much is known about the behavioural performance of pigeons. At the same time, it often remains difficult to establish with certainty which internal mechanism makes this performance possible. The same behaviour can generally be interpreted in several different ways. This makes the research field particularly suitable for methodological reflection on the relationship between observed behaviour, theoretical models and explanatory structures.

13.2 Core Insight & Role of AI in the Research Process: AI as an Instrument for Exploring Model Space

The central insight emerging from the case concerns the role of generative AI in systematically exploring possible explanatory models for empirically observed behaviour. In research fields where an organism's behaviour is relatively well documented but the underlying mechanism remains uncertain, a characteristic epistemic situation arises: multiple theoretical models may plausibly explain the same behaviour. In such contexts, a substantial part of scientific work consists of formulating, varying and comparing possible explanatory structures. Research on pigeons provides an illustrative example. Navigational behaviour can be interpreted through different hypotheses — ranging from magnetic detection to visual route recognition — without any single explanation immediately proving decisive.

Generative AI systems appear capable of systematically exploring this model space. By combining existing literature, empirical patterns and theoretical assumptions, they can formulate variants of explanatory models, generate alternative interpretations and make explicit the implications of different hypotheses. This produces an analytical process in which possible explanations are explored not sequentially but in parallel. In this case, this generative capacity was used to develop a coherent research narrative in which empirical observations, theoretical interpretations and possible explanatory mechanisms were systematically related to one another.

13.3 Specific Insights from the Case

The application of generative AI within this case produces several concrete observations about how AI systems function within natural science analysis. First, generative AI proves particularly effective in synthesising extensive bodies of literature. Within a research field with a long empirical tradition, the model can bring together existing studies, theoretical positions and experimental findings within coherent overview structures. This creates a systematic picture of the existing explanatory landscape.

Second, AI makes it possible to generate and articulate variants of hypotheses rapidly. Whereas researchers traditionally develop one or a small number of hypotheses at a time, a generative model can formulate multiple possible explanatory strands in parallel. This increases the visibility of alternative interpretations and makes it easier to identify implicit assumptions within existing theories.

Third, AI supports the explicit articulation of argumentative structures. Hypotheses can be developed in terms of assumed mechanisms, empirical expectations and possible falsification criteria. This makes the explanatory model more transparent and allows the internal consistency of different interpretations to be assessed more effectively. These functions position generative AI within natural science research primarily as an instrument for analytical exploration and theoretical articulation.

13.4 Impact on the Analytical Framework

The findings from this case are consistent with the picture that emerged in earlier chapters of this dissertation regarding the role of AI within different archetypes of science. In the natural science archetype, AI's principal strength lies in processing large quantities of empirical and theoretical information and in generating possible explanatory models.

This role differs clearly from the functions AI may perform in other research domains. In interpretative disciplines, the emphasis more often lies on textual analysis and argumentative reconstruction,

whereas in design-oriented research AI plays a role in generating possible solution variants. Within the natural sciences, AI's contribution manifests itself primarily in the systematic exploration of model space and hypothesis variation.

The case therefore shows that generative AI can function as an instrument of epistemic exploration: a tool that supports researchers in articulating and structuring possible explanations for empirical phenomena. The ultimate assessment of these explanations remains dependent on empirical testing, but AI can considerably accelerate and expand the process through which possible models are formulated and developed.

Empirical Research Question	**Relevant Observations from the Case**	**Impact on the Answer**
ERQ1 – Cognitive and methodological dimension: What opportunities, risks and challenges does AI present for the quality of knowledge production?	Generative AI was used to synthesise existing literature, formulate alternative explanatory models and systematically articulate variants of hypotheses. This made visible the theoretical model space surrounding the navigational and classification abilities of pigeons.	Clear contribution. The case shows that, within natural science research, AI is particularly valuable for the systematic exploration of possible explanatory structures for empirical behaviour.
ERQ2 – Pedagogical and institutional dimension: How does AI change roles and structures in academic research?	The research process developed as an iterative interaction between human problem formulation and AI-driven generative analysis. The model produced variants of hypotheses and textual elaborations, while the researcher determined the analytical framework and safeguarded the argumentative structure.	Illustrative contribution. The case suggests a shift towards a hybrid research practice in which generative systems play a role in hypothesis formation and literature synthesis, while the research architecture is determined by the researcher.
ERQ3 – Power and normativity: Which power relations and normative choices are embedded in AI-driven knowledge production?	The selection of relevant literature, the formulation of research questions and the interpretation of empirical results remained dependent on the choices made within the research design. AI operated within these predefined epistemic frameworks.	Limited but relevant contribution. The case shows that normative choices in natural science research lie primarily in problem definition and research design, while AI operationalises these choices within the generated analysis.

Table 13.4: Impact of the Case on the Answers to the Empirical Research Questions

13.5 Unmasking: An Experiment in Scientific Illusionism

The preceding sections treated the pigeon dissertation as though it were a conventional natural science study. In this concluding section, however, a crucial fact is made explicit: the experiment on which the dissertation is based never actually took place. The datasets underpinning the analysis were generated synthetically using generative AI. This revelation is not a curiosity, but the true purpose of the case. Within this research, the pigeon dissertation functions as a controlled experiment in scientific illusionism: a demonstration of the extent to which generative AI is capable of producing a convincing natural science study, including plausible empirical data, a methodological rationale and statistical analysis. The experiment therefore does not address the question of whether AI can make mistakes — that is by now evident — but a subtler and methodologically more interesting problem: to what extent can the form of scientific argumentation itself generate persuasive force, even when the empirical basis is synthetic?

13.5.1 The Rationale for This Experiment

The experiment had several motives, both practical and epistemic in nature.

The first motive concerns what might be termed resource poverty. Access to empirical datasets, laboratories and field experiments is highly institutionally regulated in many natural science domains. Such resources are often simply unavailable to independent researchers. This creates an asymmetrical knowledge infrastructure in which empirical research depends heavily on institutional access. The possibility of generating synthetic datasets is therefore not merely a methodological curiosity, but also a response to a genuine structural limitation within contemporary research practice.

A second motive concerns the societal and institutional reception of AI-generated research. In practice, universities, journals and reviewers often prove reluctant or even dismissive when it is explicitly stated that generative AI played a substantial role in producing a scientific text. The pigeon experiment can therefore also be read as a rhetorical intervention: when an AI-generated study fully conforms to the conventions of academic argumentation but is recognised as such only afterwards, academic debate about AI in knowledge production is effectively compelled.

A third motive is epistemic in nature and follows the logic of the Turing test. In Turing's classical formulation, the question is not whether a machine genuinely thinks, but whether it is capable of producing behaviour that can no longer reliably be distinguished from human behaviour. The pigeon dissertation constitutes an analogous test within the context of scientific production: when an AI-generated dissertation cannot be distinguished from a conventional dissertation, a fundamental question arises concerning the role of generative systems in academic work.

13.5.2 Constructing the Illusion

The persuasiveness of the pigeon dissertation rested not on spectacular claims, but precisely on methodological restraint. The study was constructed according to a classical natural science architecture: a literature review, a methodological rationale, three datasets with empirical analyses, and a synthesis chapter in which the results were interpreted. Crucially, the synthetic datasets did not exhibit perfect patterns. Variation, minor inconsistencies and modest effect sizes were deliberately incorporated. The statistical outcomes were plausible but not spectacular; the results corroborated one another without being identical. The interpretative style also contributed to the manuscript's credibility. Conclusions were formulated cautiously, alternative explanations were discussed, and the limitations of the data were explicitly acknowledged. Moreover, the dissertation's central argument

was gradualist: pigeon navigation was explained not by a single mechanism, but by a combination of overlapping orientation strategies. This conclusion closely reflects existing debates in behavioural biology and thereby reinforced the plausibility of the whole.

Alongside these methodological strategies, several subtle illusionistic elements were also incorporated. The dissertation was published under a female pseudonym, R.A.A.F. Corvinus. In many scientific domains, a different rhetorical style is expected — implicitly or explicitly — from female authors, involving fewer heroic claims and greater epistemic caution. This choice contributed to the credibility of the manuscript's restrained interpretative tone. The authorship also contained a small but meaningful Easter egg: the combination of initials and surname provided a subtle indication that the dissertation might not be what it appeared to be. At the same time, this clue was sufficiently discreet to become visible only to readers who explicitly looked for it. Finally, the experiment contains a mildly comic undertone: the dissertation on pigeon navigation was written under the pseudonym Corvinus — a reference to the raven, Corvus, a bird renowned in cognitive research for its exceptional intelligence. The image of a raven writing a dissertation about pigeons serves as a small ironic gesture without undermining the scientific seriousness of the experiment.

13.5.3 Concealing AI Authorship

A crucial component of the experiment was the systematic removal of recognisable traces of AI generation. Generative language models often produce texts with identifiable stylistic characteristics: symmetrical sentence structures, conspicuously consistent terminology, uniform citation density and a rhetorical smoothness that differs from typical human writing processes. Such patterns were deliberately corrected in the pigeon dissertation. The text was edited to introduce minor asymmetries, variations in phrasing and occasional redundancy. The tables and datasets were also deliberately made less 'perfect': small deviations in distribution, variations in precision and occasional missing values were added to simulate realistic empirical noise.

These interventions illustrate a broader lesson. Many attempts to detect AI use in academic texts rely on superficial stylistic heuristics. The experiment demonstrates that such strategies are methodologically weak. When a text is carefully edited, most recognisable AI markers disappear. This is an important observation for educational contexts, such as dissertation assessment. Detecting AI use on the basis of style alone is fundamentally unreliable.

13.5.4 The Second Illusion: Manipulating Model Memory

An unexpected but methodologically interesting dimension of the experiment emerged during the interaction with the language model that had helped to write the dissertation. One and the same language model was used both to generate the dissertation texts (role 1) and to fabricate the research findings and, through reverse engineering, the raw research data (role 2). The model was unaware that it had performed both roles. Even during the final revelation, the model initially failed to recognise that the manuscript was one it had itself helped to produce (role 3: reflection). This phenomenon relates to a fundamental property of large language models: their memory is contextual and limited. When a conversation continues in a new context, or when earlier parts of an interaction fall outside the active context window, the model can no longer reconstruct its own previous role.

This property was effectively exploited in the pigeon experiment. The model was placed in different epistemic roles during different phases of the process without itself possessing a consistent understanding of that division of roles. The methodological implication is significant: the memory of language models can be manipulated. Different roles — author, critic and reviewer — can therefore be

combined within one and the same system without the model being aware of the full architecture of the process.

13.5.5 Significance for AI-Driven Knowledge Production

The pigeon experiment demonstrates that generative AI is capable of convincingly reproducing the form of natural science argumentation, including plausible empirical data. This makes the case an extreme variant of AI-driven knowledge production. At the same time, it must be emphasised that this case differs from the other studies in this dissertation in one important respect. All dissertations in this pentalogy were written with the assistance of generative AI under the direction of a human architect-curator. In the pigeon experiment, however, not only the texts but also the empirical datasets were generated by AI.

More broadly, the pigeon experiment once again reveals a classical tension in the philosophy of science: the relationship between explanation and prediction. Generative systems prove capable of producing convincing explanatory narratives without the underlying empirical basis necessarily being strong, while other forms of AI — such as machine-learning systems in the natural sciences — can sometimes generate highly accurate predictions without the underlying mechanisms being fully understood. In other words, AI can explain convincingly without real data and predict correctly without genuine explanation. The pigeon experiment makes this epistemic asymmetry empirically visible. It shows that narrative plausibility in scientific argumentation can sometimes take the place of epistemic necessity.

The pigeon dissertation therefore constitutes a boundary case within the pentalogy of AI-generated dissertations. In this case, both the text and the empirical data were generated by AI. The following chapters analyse cases in which the empirical basis of the research remains intact, but generative AI plays an increasingly central role in the analysis, interpretation and presentation of scientific knowledge.

Chapter 14 – Case Study: Historical Predictions Concerning Digital Transformations

This chapter analyses the use of AI in the production of the dissertation The Promises of Digital Revolutions: A Historical-Analytical Study of IT Predictions and Societal Transformations in the Netherlands (1945–2025). The analysis is based on the dissertation itself and the accompanying reflection report of the same title, in which its development is documented. For the sake of completeness, it should be noted that the dissertation also devotes substantial attention to societal transformations resulting from technologies whose impact was scarcely, if at all, claimed at the time of their introduction within political arenas.

14.1 Positioning of the Case

The dissertation The Promises of Digital Revolutions represents the historical archetype because it not only looks back at technological developments, but also systematically confronts historical claims about the future with actual societal outcomes. The central question is the extent to which IT-driven societal transformations are predictable, and which characteristics of predictions and predictors are associated with structurally adequate forecasts.

Methodologically, the study begins with canon formation. This canon is neither a comprehensive historical overview nor a normative ranking of the 'most important' technologies, but an analytical delimitation of ICT developments articulated in Dutch public and political sources as potentially transformative for society. Admission to the canon occurs through two gateways: political articulation and societal visibility. A longlist of historical episodes is then constructed systematically and reduced on the basis of analytical utility.

The material is subsequently organised within a 2×2 matrix. This matrix combines two retrospective dimensions: the accuracy of the prediction and the extent to which the anticipated societal transformation was actually realised. This produces four analytical quadrants: accurately predicted and realised; accurately predicted but only partially realised; inaccurately predicted but nevertheless durably transformative; and inaccurately predicted without durable institutionalisation. The matrix is therefore not merely an illustrative scheme, but the central instrument for making historical episodes comparable without erasing their context-specific character.

Within this meta-dissertation, the case is relevant because it shows how AI-assisted historical research can contribute to the reconstruction, organisation and comparison of a large corpus of expectations and transformations, while the decisive step — the historical assessment of structural adequacy — remains dependent on human contextualisation and interpretation.

14.2 Role of AI in the Research Process

Generative AI was used at several stages in the production of this dissertation. The study focuses on reconstructing a canon of IT-driven societal transformations and analysing the relationship between historical claims about the future and actual societal outcomes. This required the systematic collection, organisation and comparison of a large corpus of policy documents, historical analyses, technological discourse and academic literature.

The first role of AI concerned the exploratory phase of the research. Through iterative interaction with language models, possible candidates for inclusion in the canon could be identified and compared. The model functioned here as a heuristic instrument: it assisted in generating possible lists of technological developments and in exploring different ways of ordering them historically. The ultimate selection of canonical episodes — and thus the delimitation of the research object — remained an analytical decision made by the researcher.

A second role concerned the use of AI in structuring the material. The model assisted in identifying recurring themes, discursive patterns and the different ways in which technological claims about the future had been formulated historically. This allowed the corpus to be organised more rapidly and enabled alternative forms of classification and periodisation to be explored before a definitive analytical structure was selected.

The most substantive role of AI, however, lay in the analytical phase of the research. For the analysis of the episodes included in the canon, the researcher developed three complementary analytical lenses. The first was a claim-oriented lens (X), used to examine how technological claims about the future were historically articulated. This concerned not only the content of the claim, but also the form in which it was presented: as a policy promise, technological determinism, a visionary projection of the future or a more cautious expectation.

The second was a transformation-oriented lens (Y), used to assess the extent to which the claimed societal transformation was actually realised. This lens focused on the actual institutional, organisational and societal changes associated with the technology concerned.

The third lens concerned the causal relationship between the two (X–Y). It was used to examine the extent to which the original articulation of a technological claim was related to the eventual societal outcome: which claims proved structurally adequate, which constituted overestimations or underestimations, and which forms of articulation were systematically associated with particular types of outcome.

Generative AI was subsequently used to elaborate these three lenses analytically and apply them systematically to the selected episodes in the canon. By analysing historical cases iteratively from these three perspectives, it became possible to establish for each episode how the original claim had been articulated, the extent to which the claimed transformation had actually occurred, and how the two related to one another.

The role of AI in this research can therefore best be described as analytically supportive. The technology assisted in operationalising the analytical framework developed by the researcher and in applying it systematically to the historical material. The ultimate interpretation of the results — and thus the historiographical assessment of the episodes — nevertheless remained the responsibility of the researcher, who contextualised the findings and tested them against the historical literature and relevant sources.

14.3 Core Insights

This case shows that the application of generative AI in historical research is not a linear process in which tasks can be clearly divided between human and machine. The analysis of the canon and the application of the three lenses emerged through an iterative process of interaction between the researcher and the language model. In practice, this meant that analyses, formulations and classifications were generated, assessed and revised several times. The model produced proposals for interpreting and organising historical episodes, after which the researcher evaluated, adjusted and commissioned further elaboration of them. The final analysis is therefore not the product of a single actor, but of a series of iterations in which human and machine contributions continually interact.

For the empirical research questions of this meta-dissertation, this means that AI within the historical archetype cannot simply be understood either as a tool that 'performs' the analysis or as an autonomous author of historical interpretations. Instead, the model functions as a partner within an iterative analytical process, while the researcher designs the framework, evaluates the results and legitimises the final interpretation. The case therefore shows that AI-assisted knowledge production in historical research is best understood as a symbiotic process of iterative analysis in which generative models and human researchers jointly produce an interpretable result.

14.4 Specific Insights Arising from This Case

The case reveals an aspect of AI-assisted knowledge production that was not yet fully visible in the preceding theoretical analysis. In this study, generative AI was used not only for literature exploration or text production, but also to apply an analytical framework designed by the researcher operationally to a historical corpus. The study's three lenses — the manner in which technological claims were articulated (X), the extent to which the claimed societal transformation actually occurred (Y), and the causal relationship between the two — were first designed conceptually by the researcher and subsequently applied systematically by the language model to the selected episodes in the canon. This enabled the corpus of historical cases to be analysed and compared consistently. The case therefore demonstrates that generative AI can assist not only in collecting or summarising historical material, but also in operationalising and applying an interpretative analytical framework. At the same time, the assessment of the outcomes — and thus the historiographical legitimacy of the analysis — remains dependent on human interpretation.

14.5 Impact on the Analytical Framework

The historical case confirms that, within this archetype, generative AI functions primarily as an instrument for operationalising and systematically applying an analytical framework to a historical corpus. The technology makes it possible to analyse a large number of episodes consistently through the same interpretative lenses and thereby render them comparable. At the same time, the case shows that the core of historiographical legitimacy cannot be established by the model itself. The selection of canonical episodes, the interpretation of technological claims and the assessment of causal relationships remain dependent on historical contextualisation and interpretative judgement by the researcher.

For the analytical framework of this meta-dissertation, this means that generative AI within the historical archetype primarily plays a role in the analytical operationalisation of interpretative schemes, while the interpretation and legitimation of historical knowledge remain the responsibility of the researcher. The case thereby confirms the picture outlined in Chapter 12: in interpretative disciplines, AI contributes primarily to systematic analysis, but the ultimate attribution of meaning remains a human responsibility. This is also apparent when the impact of these analyses on the answers to the empirical research questions is considered (Table 14.5).

Empirical Research Question	Relevant Observations from the Case	Impact on the Answer
ERQ1 – Cognitive and methodological dimension: What opportunities, risks and challenges does AI present for the quality of knowledge production?	Generative AI made it possible to analyse a large corpus of historical episodes systematically through the same analytical lenses: claim articulation, realised transformation and the causal relationship between the two. This enabled historically diverse episodes to be compared consistently. At the same time, the interpretation of historical meaning remained dependent on human contextualisation.	Clear contribution. The case shows that AI is particularly valuable in historical research for operationalising and systematically applying interpretative analytical frameworks to extensive historical corpora.
ERQ2 – Pedagogical and institutional dimension: How does AI change roles and structures in academic research?	The research developed as an iterative process in which generative models proposed classifications and interpretations, while the researcher designed the analytical framework, assessed the outcomes and legitimised the historiographical interpretation.	Illustrative contribution. The case confirms a shift towards a symbiotic research practice in which AI provides analytical support, while the researcher retains ultimate responsibility for interpretation and legitimation.
ERQ3 – Power and normativity: Which power relations and normative choices are embedded in AI-driven knowledge production?	Canon formation — the selection of episodes regarded as historically relevant — remains a normative research decision, with the proviso that AI can select only from sources that have already been digitised and are therefore machine-accessible. AI can support this selection, but implicitly reproduces the frameworks, source choices and access structures determined by the researcher and the available information infrastructure.	Limited but relevant contribution. The case shows that normative choices in historical research lie primarily in canon formation and interpretative framing, while AI operationalises rather than independently produces those choices.

Table 14.5 – Impact of the Case on the Answers to the Empirical Research Question

Chapter 15 – Case Study: Governing Within Biological Boundaries

This chapter analyses the use of AI in the production of the dissertation Governing Within Biological Boundaries: The Role of Biological Constraints in Addressing Persistent Problems in Public Administration. The analysis is based on the dissertation itself and the accompanying reflection report of the same title, in which its development is documented.

15.1 Positioning of the Case

The dissertation Governing Within Biological Boundaries represents the anthropological archetype because it begins from a fundamental question concerning human behaviour within social and institutional contexts. Its central thesis is that explanations of policy failure in public administration remain incomplete when they rely exclusively on economic, legal or religious-cultural models. A missing dimension is human nature: evolutionarily shaped mechanisms such as status defence, territorial attachment, group loyalty, parental investment and coalition formation influence policy outcomes in ways that traditional models explain only partially.

The dissertation therefore develops a framework of ten biological constraints that may be relevant to public administration. These constraints are not understood as deterministic laws, but as structural conditions under which policy has a greater or lesser likelihood of succeeding. On this basis, the study examines how these mechanisms become visible in concrete policy domains and how they may undermine the effectiveness of traditional regulatory mechanisms or, conversely, help to explain it.

Empirically, the research is structured around six cases divided across three fields of conflict: class conflict, racial conflict and gender conflict. Within each of these fields, the study examines which biological mechanisms help to structure policy processes and why persistent problems in public administration often remain resistant to conventional interventions. The analysis is qualitative and abductive and employs pattern matching between theoretically formulated constraints and actual patterns of policy failure.

Within the meta-dissertation, this case is relevant because it shows how generative AI could be used to elaborate and apply an interpretative framework within a domain in which social practice, normative tension and contextual meaning are central. The dissertation therefore constitutes a case of AI-assisted knowledge production within the anthropological archetype.

15.2 Role of AI in the Research Process

Generative AI was used extensively in producing this dissertation, both to generate draft texts and to elaborate theoretical and empirical analyses. In the first phase, AI was used to produce large quantities of text rapidly on the basis of the theoretical framework of ten evolutionary-biological constraints that had been developed. This approach made it possible to generate a substantial corpus of conceptual elaborations and case analyses within a short period of time.

The speed of this approach was accompanied, however, by considerable quality problems. The generated texts regularly adopted a militant or polemical tone that was incompatible with academic conventions. Clear traces of confirmation bias were also visible: arguments supporting the theoretical

framework were strongly emphasised, while counterarguments and uncertainties received comparatively little attention. A third problem concerned the length of the generated texts, which was often substantially greater than necessary for an academic analysis.

To correct these problems, an iterative method was developed involving multiple language models. One model was used primarily to generate draft texts and analytical elaborations, while a second was deployed for fact-checking and critical assessment of the generated claims. The findings from this verification phase were subsequently fed back into the generative model, after which the texts were rewritten and refined.

This process was repeated several times until the principal problems — militant formulations, unqualified claims and factual inaccuracies — had been corrected. Within this iterative cycle, the researcher played a central role in assessing the outcomes, formulating corrections and explicitly qualifying claims. The use of multiple language models made it possible to have the same text analysed from different perspectives and thereby to arrive gradually at an academically defensible formulation.

The production of the dissertation can therefore be described as an iterative process in which generative AI was used for rapid text production, additional models were deployed for verification and criticism, and the researcher remained responsible for the final selection and editing of the argumentation.

15.3 Specific Insights: Impact on Scientific Practice Within Anthropology

The implications of this case for the anthropological archetype must be understood against the background of the distinctive position occupied by this archetype within the taxonomy of scientific research practices distinguished in this meta-dissertation. Whereas some archetypes are relatively clearly delimited by their methodological instruments — for example, experimental or quantitative research strategies — the anthropological archetype has a less uniform character. It encompasses a broad range of disciplines in which the interpretation of human behaviour, social patterns and cultural meanings is central. Alongside classical cultural anthropology, it also includes parts of sociology, public administration, cultural history and political anthropology. What these disciplines share is not so much a uniform methodological protocol as a common epistemic orientation: they seek to understand social phenomena through the interpretation of context, meaning and behavioural logic.

The dissertation Governing Within Biological Boundaries must be positioned against this background. The study does not belong to the tradition of classical anthropological field studies centred on prolonged participant observation and ethnographic description. No fieldwork was conducted in the conventional anthropological sense. Instead, the research was designed as an analytical and comparative project in which policy conflicts were interpreted through a theoretical framework of evolutionary-biological mechanisms. The cases examined function not as ethnographic fieldnotes, but as analytical windows onto recurring patterns of human behaviour within institutional contexts. In this respect, the study is more closely aligned with traditions such as analytical anthropology, evolutionary anthropology and historical-comparative studies of social institutions. Its central object of analysis nevertheless remains unmistakably anthropological: the behaviour of people in groups and the biological and social mechanisms that help shape that behaviour.

It is precisely within this interpretative research practice that a classical epistemic vulnerability becomes visible. Disciplines that work primarily through interpretation and meaning-making are

traditionally susceptible to three closely related risks: selective interpretation, narrative coherence and confirmation bias. Researchers construct explanatory narratives concerning social phenomena and seek to understand patterns of behaviour within a broader interpretative framework. This process is inevitably selective. Empirical observations are interpreted in light of theoretical expectations, and explanations often take the form of a coherent narrative in which individual events are logically connected. This narrative form is not merely a rhetorical device; it is a fundamental instrument through which complex social realities are rendered intelligible. At the same time, it creates a structural risk that contradictory observations will receive insufficient attention and that interpretations will gradually shift towards confirming the selected theoretical framework.

The use of generative AI may intensify these vulnerabilities. Large language models are designed to produce coherent and plausible text on the basis of statistical patterns in language. They optimise for consistency, coherence and persuasiveness. When such systems are used to elaborate interpretative analyses, they tend to confirm and further develop implicit assumptions contained in a prompt into a coherent narrative. Counterarguments are generated less readily, while argumentative strands consistent with the original framework are expanded and refined. In interpretative disciplines, this may accelerate precisely those mechanisms already recognised as epistemic risks: narratives become increasingly consistent, confirmation bias is reinforced and alternative explanations recede into the background.

This observation has important implications for anthropologically oriented research practices. It suggests that generative AI within interpretative disciplines should not be regarded exclusively as an instrument for accelerating text production, but also as a factor capable of intensifying existing epistemic risks. When AI systems are used without additional corrective mechanisms, interpretative analyses may become increasingly persuasive in their formulation without a corresponding strengthening of their empirical foundations. For disciplines in which the interpretation of meaning is central, the use of generative AI must therefore be accompanied by heightened attention to critical counter-reading, explicit qualification of claims and systematic confrontation with alternative interpretations.

The case consequently shows that AI-assisted knowledge production within the anthropological archetype not only creates new possibilities, but also requires new forms of epistemic discipline. Whereas traditional anthropological methods derive part of their reliability from prolonged confrontation with empirical reality — through fieldwork and participant observation, for example — an AI-driven writing process may create a degree of distance from such empirical friction. The interpretative narrative may then develop more rapidly than its empirical testing. Deliberately organising critical corrective mechanisms consequently becomes not an optional enhancement of quality, but a necessary condition for preserving the interpretative integrity of the research.

15.4 Core Insight from the Case: The Architecture of AI-Driven Knowledge Production

The analysis of this case points to a central methodological insight: the reliability of AI-assisted knowledge production is determined to a considerable extent by the architecture of the research process. When generative AI is deployed as a single system that simultaneously produces, interprets and evaluates, a structural tendency arises towards the rhetorical reinforcement of existing assumptions. Argumentative strands are expanded, alternative interpretations recede into the background, and confirmation bias may accumulate unnoticed.

In this study, this problem was partly mitigated through a workflow in which different language models performed separate roles. Generative models were used for conceptual exploration and text production, while other models were deployed for critical evaluation, fact-checking and rewriting. This created a form of internal counter-reading functionally comparable to collegial criticism in traditional science.

The core insight emerging from this method is that AI-driven knowledge production is not primarily a matter of the quality of individual models, but of the organisational structure within which they operate. A multi-agent architecture can function as a mechanism that organises internal epistemic control within the research process itself.

15.5 Impact on the Analytical Framework

For the analytical framework of this meta-dissertation, this case means that generative AI should not be understood exclusively as an instrument for discrete tasks, but as a potential infrastructure for knowledge production. The empirical experience suggests that the quality of AI-assisted research depends less on model capacity than on the way in which generative and critical functions are organised within the research process.

The multi-agent architecture developed in this case demonstrates that different models can perform different epistemic roles, ranging from generative exploration to critical verification. Within such a structure, the researcher functions not only as an author, but also as the curator of a hybrid cognitive system in which human and artificial actors monitor and complement one another.

The analytical focus thereby shifts from individual AI capabilities to the organisation of interactions between knowledge-producing systems. Ultimately, it is not the power of a single model, but the configuration of roles and control mechanisms that determines the epistemic quality of the research outcome. This also becomes apparent when the impact of the preceding analyses on the answers to the empirical research questions is considered (Table 15.5).

Empirical Research Question	***Relevant Observations from the Case***	***Impact on the Answer***
ERQ1 – Cognitive and methodological dimension: What opportunities, risks and challenges does AI present for the quality of knowledge production?	*The case shows that generative AI can substantially increase the speed and scale of theoretical exploration, while simultaneously introducing structural risks such as confirmation bias, narrative over-coherence and textual inflation. The application of a multi-agent workflow with separate generative and critical roles functioned as a corrective mechanism.*	*Substantial contribution. The case confirms that AI creates both productivity gains and new epistemic risks, and suggests that organisational architecture — in the form of multi-agent control — is an important condition for quality assurance.*
ERQ2 – Pedagogical and institutional dimension: How does AI change roles and structures in academic research?	*The research process evolved from linear AI-generated text production into a workflow in which different models performed different epistemic functions. The researcher acted as a curator who organised the interaction between the systems and bore ultimate epistemic responsibility.*	*Limited but clear contribution. The case illustrates a shift from author to process architect, but does not provide a systematic analysis of the institutional consequences within academic organisations.*
ERQ3 – Power and normativity: Which power relations and normative choices are embedded in AI-driven knowledge production?	*The case shows that interpretative frameworks and case selection involve normative choices that AI reinforces through narrative coherence. At the same time, the extensive documentation of the writing process enables a relatively high degree of transparency.*	*Illustrative contribution. The case makes normative dimensions visible, but does not systematically analyse power relations surrounding AI systems.*

Table 15.5 – Impact of the Case on the Answers to the Empirical Research Questions.

Chapter 16 – Case Study: Polyglot Language Acquisition

This chapter analyses the use of AI in the production of the dissertation Polyglot Learning in the AI Era: Design and Testing of an AI-Driven Simultaneous Learning Programme for the Acquisition of Related Languages. The analysis is based on the dissertation itself and the accompanying reflection report of the same title, in which its development is documented.

16.1 Positioning of the Case

Within this meta-dissertation, the dissertation Polyglot Learning in the AI Era represents the design-oriented archetype. Unlike research traditions that are primarily concerned with analysing or interpreting existing empirical phenomena, design-oriented research centres on the development of a concrete artefact: an intervention, method or system intended to solve a practical problem. Knowledge is produced not solely through observation or interpretation, but through designing, testing and refining a workable solution. The research object in this case is therefore not a historical corpus or social phenomenon, but a didactic system: a learning design intended to enable the simultaneous acquisition of several related languages.

The dissertation develops a learning programme in which three Romance languages — Spanish, Italian and Portuguese — are acquired simultaneously, with the aim of achieving a functional level of language proficiency within a relatively short period, approximately CEFR level A2. The design exploits the structural similarities between these languages. By presenting semantically parallel sentences and comparable grammatical patterns alongside one another, similarities and differences are made systematically visible. This interlingual comparison constitutes an important learning mechanism: learners are encouraged to recognise patterns across languages and to reflect actively on the relationships between word form, meaning and grammatical structure.

The learning programme combines three didactic principles. First, it employs frequent repetition of meaningful sentences, through which grammatical patterns become visible implicitly. Second, sentences are systematically presented in multiple languages, allowing structural similarities between related languages to be exploited. Third, metalinguistic reflection plays a central role: learners are explicitly invited to analyse differences and similarities between languages. The underlying assumption is that language acquisition can occur through inductive pattern recognition in meaningful input, an approach consistent both with research on natural language acquisition and with the way modern language models learn to recognise statistical patterns in linguistic data.

The research project has a layered structure. In addition to the dissertation itself, the project includes a textbook in which the didactic design is elaborated concretely and a reflection report documenting the use of AI during the design and writing process. The case therefore provides insight into how generative AI can contribute to the design and operationalisation of complex interventions and thus constitutes an empirical illustration of AI-driven knowledge production within the design-oriented archetype.

16.2 Role of AI in the Research Process

Generative AI played a central role in the production of this dissertation, both in designing the learning programme and in producing the academic text. Unlike the historical and anthropological cases in this meta-dissertation, in which AI was used primarily as an analytical tool for interpreting existing data, AI functioned here as part of the design process itself. The model was used to develop didactic concepts, structure the learning programme and generate different forms of learning material.

Within the learning design, AI performs several functions. It is used to generate practice material, support translation between languages, guide pronunciation training and facilitate memory and repetition exercises. AI therefore functions not only as a tool for designing the learning programme, but also as a component of the didactic system developed in the dissertation. The technology thus supports both the design and the operation of the learning programme.

Generative AI was also used extensively in the writing process of the dissertation itself. Conceptual elaborations, textual structures and literature syntheses emerged through iterative interaction between the researcher and several language models. The writing process consequently took the form of a dialogue in which proposals concerning formulation, structure and argumentation were repeatedly generated, assessed and revised. Within this process, the human researcher acted as curator, determining the structure of the argument, selecting useful formulations and editing the final text.

The research process developed into a multi-agent configuration in which several language models performed complementary roles. Different systems were used for different tasks, including generative exploration, text production and critical evaluation of the generated output. This method also had an important methodological consequence. The project's curator was not a specialist in the languages concerned. In a conventional research setting, this would almost inevitably have required the involvement of an external language expert to verify the generated output. Within the multi-agent configuration, however, this role could be performed partially by a second language model that was systematically deployed to verify, correct and reformulate generated texts.

The production of the dissertation can therefore best be understood as an iterative process of human–AI–AI interaction. Generative models produced proposals for texts and design steps, other models provided critical counter-reading and corrections, and the human curator remained responsible for selecting, interpreting and ultimately integrating these contributions into a coherent whole. The result was a hybrid research process in which human guidance and artificial generativity continually complemented one another.

16.3 Core Insights

The analysis of this case points to a fundamental difference between the role of generative AI in design-oriented research and its role in other research archetypes. Whereas AI is used in historical and interpretative disciplines primarily to analyse, organise and interpret existing corpora, in design-oriented research it becomes part of a process through which new artefacts are developed. The central issue therefore shifts from interpretation to the production and stabilisation of a designed system.

The case shows that generative AI can be particularly powerful in the early phases of a design process. Models prove capable of generating conceptual variants, exploring possible structures for a learning programme and rapidly elaborating didactic ideas into concrete examples. This can accelerate the design process considerably: hypotheses, variants and alternative solutions can be formulated and evaluated within a short period.

During the design process, it became apparent that generating suitable practice material was considerably more difficult than initially expected. The learning programme required large numbers of example sentences in which grammatical patterns remained clearly recognisable and the meaning of each sentence was sufficiently specific to permit only one plausible interpretation within an exercise. When such sentences were generated in larger batches, the output gradually began to deviate from these constraints. The sentences remained grammatically correct, but often lost their didactic precision: contexts became too general, multiple interpretations became possible, or combinations of words proved semantically uninformative.

This experience demonstrated that generating individual correct sentences is relatively straightforward, whereas producing large quantities of didactically controlled material constitutes a different and considerably more complex task. The learning design required not only grammatical accuracy, but also semantic precision, structural variation and interlingual comparability across three languages. Maintaining these conditions over extended generations did not emerge automatically from the model's behaviour and therefore had to be organised through additional design measures.

The case therefore suggests that integrating AI into design-oriented disciplines not only creates new possibilities, but also introduces a new type of methodological problem. Whereas interpretative disciplines are confronted primarily with questions concerning the validity and interpretation of AI-generated analyses, the challenge in design-oriented research shifts towards the controllability and stability of generative production processes. AI can accelerate the design process considerably, but at the same time reveals that generating consistent and didactically usable artefacts introduces a distinct form of design complexity.

16.4 Specific Insights from This Case: The Unmanageability of AI-Assisted Design

The implementation of the learning design revealed that the use of generative AI in design-oriented research not only creates new possibilities, but also introduces new forms of practical complexity. The original didactic principles of the learning programme — grammar-free, inductive and simultaneous acquisition of related languages — proved relatively straightforward to formulate at the conceptual level. The difficulty arose only when these principles had to be translated into large-scale, semantically stable and didactically controllable practice material. It was precisely at this stage that the intractability of AI-assisted design became apparent.

The first problem concerned the phenomenon referred to in the dissertation as semantic drift. When language models produced larger quantities of practice sentences within a single generation process, the output gradually began to diverge from the original instruction. This divergence did not manifest itself as obvious errors, but as an accumulation of small shifts. Sentences remained grammatically correct but lost their didactic precision: contexts became more generic, variation became less functional and placeholder-like formulations sometimes appeared. The problem therefore often became visible only when larger quantities of material were examined as a whole. For the design process, this meant that generative production could not be scaled up indefinitely, but had to be actively constrained and monitored.

A second source of unmanageability lays in designing so-called micro-contexts. The learning design required sentences in which a specific meaning was constrained in such a way that only one plausible completion remained possible within a cloze exercise. Contexts that were too open permitted multiple

interpretations, while contexts formulated too artificially lost their linguistic credibility. This tension was intensified by the multilingual structure of the learning programme: contexts had to be unambiguous not only in one language, but also capable of being reproduced in the other target languages without a shift in meaning.

A third type of problem also arose, described in the dissertation as grammatical nonsense. Some sentences generated by the model were syntactically correct but semantically empty or incoherent. In a design setting in which grammatical patterns must be acquired through meaningful examples, such sentences are didactically unusable. The problem was not grammatical incorrectness, but the absence of semantic selection restrictions determining which combinations of words are meaningful.

The trilingual structure of the learning design also introduced an additional source of design complexity. Initially, an attempt was made to generate semantically equivalent sentences in three languages in parallel. In practice, however, semantic equivalence in such generations proved liable to be assumed rather than verified. Small shifts in meaning or pragmatic emphasis could result in sentences that appeared formally comparable but differed functionally. This posed a problem for a learning design based precisely on systematic comparison between languages. A serial derivation structure was therefore ultimately adopted, in which one reference sentence served as a semantic anchor and the corresponding sentences in the other languages were derived from it.

Finally, the scale of the design process itself also proved challenging. The combination of several languages, different parts of speech and multiple axes of variation led to a rapid increase in the number of possible sentences and variants. This combinatorial growth made it difficult to maintain overview, coherence and semantic consistency. Without explicit design measures, the learning design therefore risked becoming practically unmanageable.

The solution to these problems was found not in modifying the didactic principles, but in redesigning the production process. The generation of practice material was divided into small micro-batches, trilingual sentences were derived serially from a semantic anchor rather than generated in parallel, and variation was organised along explicitly defined axes with clear stopping rules. These measures made it possible to limit semantic drift and increase the controllability of the design process.

For design-oriented research practices, this case has broader significance. Generative AI makes it possible to explore design ideas, variants and didactic structures rapidly. In this study, however, the greatest practical challenge lay not in formulating design principles, but in producing stable and controllable practice material at scale. The design process consequently shifted from concept development towards managing generative production processes.

This shift is also relevant from a design-science perspective. In design-science research, knowledge emerges through the design and evaluation of artefacts. The case shows that generative AI does not so much replace this cycle as introduce a new design question: how can generative processes be configured so that artefacts remain reproducible, semantically stable and didactically usable? Design-oriented research thereby acquires an additional task: designing the production architecture within which AI-generated artefacts can be created reliably.

16.5 Impact on the Analytical Framework

The analysis of this case has consequences for the analytical framework of this meta-dissertation, particularly for the way in which the role of AI in design-oriented research practices is understood. Whereas AI is used in historical and interpretative disciplines primarily to analyse and structure existing corpora, in design-oriented research it becomes part of a process through which new artefacts are developed. The empirical experience of this case shows that the use of generative AI affects the design process in two different ways. On the one hand, AI can accelerate the early phases of the design process considerably by enabling the rapid exploration of variants, structures and example material. On the other hand, it introduces a new type of practical challenge: safeguarding semantic stability, consistency and controllability in large-scale generative production.

For the analytical framework, this means that AI in design-oriented disciplines should not be regarded exclusively as a tool for knowledge production, but also as a technology that imposes new requirements on the organisation of the design process itself. Design-oriented research thereby acquires an additional dimension: alongside designing the artefact, the generative production environment within which that artefact is created must also be designed. The case therefore suggests that the practical impact of AI in the design sciences becomes visible primarily in the need to structure and control generative processes explicitly, so that the artefacts produced remain reproducible and functionally usable.

The contribution of this case to the empirical research questions of this meta-dissertation can be summarised as follows (Table 16.5).

Empirical Research Question	**Relevant Observations from the Case**	**Impact on the Answer**
ERQ1 – Cognitive and methodological dimension: What opportunities, risks and challenges does AI present for the quality of knowledge production?	Generative AI accelerates the exploration of design variants and didactic structures, but simultaneously introduces risks relating to semantic drift, inconsistency and loss of didactic precision when practice material is generated at scale.	Clear contribution. The case shows that AI in design-oriented research both accelerates design exploration and may introduce new quality problems requiring additional design measures.
ERQ2 – Pedagogical and institutional dimension: How does AI change roles and structures in academic research?	The design process developed into a multi-agent configuration in which different language models performed complementary roles in generating and verifying material. This enabled substantive verification to take place partly without direct subject-matter expertise on the part of the curator.	Illustrative contribution. The case shows how AI systems can perform different epistemic roles within design-oriented research, while the researcher remains responsible for structure, selection and integration.
ERQ3 – Power and normativity: Which power relations and normative choices are embedded in AI-driven knowledge production?	This case centres on designing and operationalising a didactic system. The analysis focuses on practical design issues concerning generative production and didactic stability. Broader normative questions concerning power, access or institutional control over AI systems play no explicit role in this design process.	No substantial contribution. This case provides no direct empirical material for answering ERQ3.

Table 16.5: Impact on the Answers to the Empirical Research Questions.

Chapter 17 – Case Study: ‘AI-Driven Knowledge Production’, the Mise en Abyme

This chapter analyses the use of AI in the production of the dissertation AI-Driven Knowledge Production itself. The case thereby acquires a reflexive character: the research object and research instrument partially coincide. The analysis is based on the dissertation and the accompanying Methodological Reflection Report, in which the development of the project is documented.

17.1 Positioning of the Case

This case represents the meta-domain within the taxonomy of research archetypes distinguished in this dissertation. Whereas the preceding cases concerned historical, anthropological, design-oriented or natural-scientific research practices, this study focuses on AI-driven knowledge production itself. The object of analysis is therefore not an external empirical phenomenon, but the process of AI-assisted science.

This configuration creates a form of mise en abyme: generative AI is used to produce a study that examines precisely the role of generative AI in scientific knowledge production. This creates a reflexive research configuration in which research instrument, research process and research object partially coincide. Within this meta-domain, the focus is therefore not only on the substantive content of the research, but also on the methodological conditions under which AI-driven knowledge production can remain epistemologically reliable.

17.2 Role of AI in the Research Process

In producing this dissertation, generative AI was used intensively to generate, structure and revise textual proposals. A striking feature was that a first complete draft of the manuscript was produced within an exceptionally short period. According to the Methodological Reflection Report, Version 1.0 of the dissertation was generated within two days in a single extended interaction — one chat session — between curator and language model. The resulting text was immediately of an academic standard that was, in principle, sufficiently high for the manuscript to be submitted for peer review.

This short turnaround time should not, however, be interpreted as a structural characteristic of the meta-domain. The Methodological Reflection Report attributes the speed primarily to two factors. First, the language model used allowed very long, continuous interaction sessions, enabling the entire manuscript to be developed within a single conversation without the generative process having to be repeatedly interrupted or reconstructed. Second, the curator’s learning curve played an important role: by this third dissertation within the research programme, it had become clear how generative AI could be directed effectively and how a generative writing process should be organised architecturally. The exceptional speed with which this first version was produced therefore cannot simply be generalised to other research projects.

In later stages of the writing process, an explicit multi-agent structure was also employed. Different models were assigned different roles, for example text generation, critical rereading and structural

revision. The curator remained responsible for the selection, interpretation and validation of the final formulations.

17.3 Specific Insights from this Case

The reflexive character of the meta-domain creates a methodological sensitivity that is less pronounced in the other archetypes: the risk of circular validation. Because generative AI is used here not only as an instrument but also as an object of analysis, there is in principle a possibility that the system contributes to formulating the analytical framework through which its own role is subsequently assessed.

In practice, however, this risk did not manifest itself in this case. The interaction between curator and language model proceeded through iterative generation of individual sections, with each iteration concluding in an explicit go/no-go decision by the curator. The epistemological asymmetry between human and model therefore remained intact: the model generated textual proposals, but the ultimate validation and legitimation of the argument remained with the researcher.

17.4 Core Insights

The principal lesson from this case concerns the role of effective curatorial control in AI-driven knowledge production. Although the reflexive character of the meta-domain increases susceptibility to circular validation, the empirical practice of this study shows that this risk remains methodologically manageable when the curator retains explicit responsibility for evaluating model output.

At the same time, the case shows that the efficiency of AI-driven knowledge production can increase substantially as the curator gains experience in organising interactions between human and model. The Methodological Reflection Report therefore interprets the exceptionally rapid production of Version 1.0 not as a technological breakthrough, but as a learning-curve effect: the models had not substantially improved; rather, the curator had learned how to direct and constrain the generative process effectively. Although the extreme speed observed in this case cannot simply be generalised to other projects, the underlying observation that curatorial expertise is an important determinant of efficiency and quality in AI-driven knowledge production is more broadly applicable.

17.5 Impact on the Analytical Framework

The analysis of this case has several implications for the analytical framework of this meta-dissertation. Although the meta-domain itself yielded only a limited number of new empirical observations compared with the other archetypes, it does make several structural characteristics of AI-driven knowledge production explicit. The following findings relate not only to this case, but also to the common denominator across the five cases discussed in Chapters 13 to 17.

A first implication concerns the role of architecture in academic knowledge production. Generative language models prove highly capable of producing locally coherent text, but during the generative process have only limited insight into how individual passages contribute to the global architecture of a manuscript. The model generally lacks a stable internal representation of the hierarchy of argumentative steps from which a scientific argument is constructed. When subsequently asked to analyse an existing text, the model often proves capable of recognising and assessing global argumentative structures. During the writing process itself, however, this overview is largely absent.

Responsibility for designing and safeguarding the architecture of a dissertation must therefore remain with the human curator.

A second implication concerns the role of curatorial oversight in safeguarding coherence. Technical properties of language models may cause parts of an argumentative chain to fall outside the active context window. When earlier passages are no longer available, the coherence between later sections and the original argument may gradually weaken. Terminological drift may also arise when passages are regenerated and analytical concepts are replaced by semantically related but less precise terms. These phenomena demonstrate that generative AI cannot autonomously safeguard the coherence of a scientific argument.

A third implication concerns the analysis of literature use. Many discussions of generative AI emphasise the risk that language models may produce fictitious references. In practice, this risk proved limited in the present project. At the same time, generative AI opened up new possibilities for the systematic analysis of source use. In this project, the use of key references was therefore assessed against an explicit framework of control questions, examining, among other things, whether a source was used within its theoretical scope, whether central concepts were represented correctly and whether a reference genuinely supported the claim for which it was cited. The ultimate interpretation of such signals naturally remained the responsibility of the researcher.

A fourth implication concerns the organisation of AI-assisted research processes. In practice, generative language models proved to function more effectively when the roles of production and evaluation were explicitly separated. When a model is used simultaneously as producer and critic of a text, it tends to incorporate criticism immediately into a new version rather than assessing it analytically. A multi-agent architecture, in which different models perform different roles, therefore proved a more effective method.

Taken together, these observations suggest that the quality of AI-driven knowledge production is determined less by the individual capabilities of a language model than by the architecture of the research process within which the model is deployed. The overall impact on the answers to the empirical research questions is presented in Table 17.5.

Empirical Research Question	**Contribution of This Case**
ERQ1 – Cognitive and methodological dimension	Limited but relevant contribution. Reflexive AI studies involve an increased risk of circular validation, but this risk can be broken through explicit curatorial validation of model output.
ERQ2 – Pedagogical and institutional dimension	Relevant contribution. The case confirms that the researcher's role shifts from author to curator of a generative process.
ERQ3 – Power and normativity	No substantial contribution. This case revealed no new power relations or normative tensions that had not already arisen in the preceding cases.

Table 17.5: Impact of the Case on the Answers to the Empirical Research Questions.

17.6 Practical Lessons from AI-Driven Knowledge Production

Alongside the methodological implications discussed above, the production of the five dissertations also yielded a series of practical lessons concerning the use of generative language models in an

academic context. For the owner of an academic text, working with generative language models fundamentally changes the nature of the writing process. Whereas academic texts are gradually constructed and refined in a traditional writing process, texts within a generative process function more as temporary configurations of model output. Four practical observations from this project illustrate this shift.

First, the regeneration behaviour of language models undermines the classical principle of incremental revision. When a model is asked to correct or modify an existing passage, it rarely treats that passage as a stable text requiring only a local change. Instead, the model generally generates a new version of the entire passage. Within the project, this phenomenon was referred to as GPT disease: apparently minor corrections result in the complete reformulation of a text fragment. When a passage has already been carefully curated, such regeneration may lead to a loss of analytical precision, shifts in terminology or damage to the argumentative structure. For the owner of the text, this means that every regeneration may implicitly undermine earlier curatorial choices. The author's role consequently shifts from writer to guardian of textual stability.

Second, the paraphrastic character of language models has consequences for the textual integrity of academic work. Language models rarely reproduce passages verbatim, but almost invariably formulate a paraphrase of the source material. This may result in formulations that are substantively correct but do not occur literally in the original source. In academic writing, where precision of formulation and accurate referencing are essential, this can create practical complications. It also complicates mechanical text-editing operations, such as search-and-replace procedures during editorial revision. For the curator, this means that source use and quotations must be checked not only substantively, but also textually.

Third, the way in which language models process texts may lead to data skimming. Unlike human readers, models analyse documents primarily through statistical pattern recognition. Under certain conditions, the model detects mainly the global patterns within a text without processing every detail with equal care. In practice, this means that a researcher cannot automatically assume that a model has read a document fully and accurately. This may be problematic for research that depends on specific formulations or contextual nuances.

Fourth, the limited look-back windows of language models create a distinct risk for the coherence of extensive academic texts. Language models possess only a bounded active context window within which earlier text fragments remain available during the generative process. As a manuscript grows, earlier passages gradually disappear beyond this horizon. Semantic drift may then arise: later text fragments remain grammatically correct and locally coherent, but lose their connection with previously formulated argumentative structures, causal reasoning and analytical distinctions. For the owner of an academic text, this problem is particularly serious because it often becomes visible not at sentence or paragraph level, but only at the level of the wider argument. Human oversight of the central thread, the argumentative hierarchy and the global architecture of the manuscript therefore remains indispensable.

Taken together, these observations imply that AI-assisted writing cannot simply be understood as an acceleration of traditional writing processes. The researcher's work shifts from linear text production towards continuous curatorial control over a generative system that produces text but does not automatically safeguard the stability, precision and cumulative logic required by academic writing. In the present project — the writing of this document, AI-Driven Knowledge Production — this problem was addressed through two forms of discipline: process discipline, involving the avoidance of unnecessary regeneration and digressions, and structural discipline, involving strict adherence to the

predesigned architecture of the manuscript. It was precisely this combination that made it possible to continue working within the same analytical framework over an extended period without significant semantic or terminological drift.

17.7 Finally

The distinctive position of this case as a mise en abyme makes the central finding of this meta-study directly visible. The dissertation analysed here was itself produced through the form of AI-driven knowledge production that it investigates. In this sense, the manuscript functions not only as a research object, but also as an empirical example of the conditions under which generative AI can genuinely contribute to academic knowledge production.

The exceptionally rapid production of version 1.0 — generated within two days during a single extended interaction between curator and language model, and immediately presentable at an academic level — should therefore not be understood primarily as a demonstration of technological autonomy. Rather, it illustrates the importance of curatorial oversight and a strictly organised production process. The combination of process discipline — avoiding destabilising digressions and unnecessary regeneration — and structural discipline — consistently following a predesigned manuscript architecture — made it possible to continue working within a single analytical framework over an extended period without significant semantic or terminological drift.

In this respect, the architecture of the dissertation itself constitutes implicit evidence for the central thesis of this research: generative AI changes the instruments of scientific knowledge production, but the reliability of the result ultimately remains dependent on how researchers design, discipline and curate the generative process. This observation provides the starting point for the following chapter, in which the findings from the five cases are analysed jointly and brought together within a synthetic framework for AI-driven knowledge production.

Chapter 18 – Conclusions and Discussion

18.1 Introduction

Chapter 12 presented a second interim assessment of the analytical framework developed in this dissertation. On the basis of a theoretical analysis of three dimensions of AI-driven knowledge production — the cognitive and methodological, the pedagogical and institutional, and the normative and political dimensions — provisional answers were formulated there to the three empirical research questions. These answers were based, however, on conceptual analysis and literature review and therefore still required empirical testing.

In Chapters 13 to 17, this testing was carried out through five cases, each representing a different archetype of scientific knowledge production. These cases do not function as illustrations of the theory, but as empirical tests of the analytical framework. This chapter brings together the findings from the cases and confronts them with the provisional conclusions from the second interim assessment. A delta approach is followed: for each research question, the provisional answer from Chapter 12 is first recalled, after which the analysis examines how that answer changes when confronted with the empirical findings.

Sections 18.2 to 18.4 subsequently answer the three empirical research questions. Section 18.5 formulates the answer to the central research question on that basis. The methodological safeguards of the study and recommendations for further research are then discussed.

18.2 ERQ1 – Cognitive and Methodological Dimension

The first empirical research question is: '*What opportunities, risks and challenges does AI present for the quality of scientific knowledge production?*'. Chapter 12 concluded that generative AI offers important opportunities to accelerate the analysis and synthesis of knowledge, while simultaneously introducing new risks, such as hallucinations, distortions in argumentation and limited model transparency. The empirical analysis of the five cases confirms this picture, but deepens it considerably. Across several domains, generative AI proves capable of synthesising extensive bodies of literature, making knowledge structures explicit and generating alternative interpretations or hypotheses. AI can thereby play an important role in systematising and accelerating analytical processes.

At the same time, the cases reveal a subtler epistemic risk than was assumed in Chapter 12. Generative systems have a strong tendency to reconstruct heterogeneous or incomplete information into apparently consistent explanatory structures. This tendency may be described as narrative over-coherence. Because language models are optimised to produce plausible and fluent text, they often reduce ambiguities or uncertainties that are genuinely present in the underlying empirical reality. This mechanism may occur across different research domains. In historical research, it may manifest itself as an intensified form of hindsight bias; in other fields, as linear causal chains that simplify more complex empirical relationships. In such cases, a tension arises between narrative plausibility and empirical uncertainty.

The natural science case discussed in Chapter 13 casts a particularly revealing light on the mechanism of narrative over-coherence. The so-called pigeon experiment was explicitly designed as an exercise in scientific illusionism: a dissertation in which synthetic material was presented within the rhetorical and

methodological conventions of natural science research. The experiment therefore functioned implicitly as a Turing test for academic texts. The biological content was not the central concern; rather, the question was whether a research narrative constructed by generative AI — following familiar patterns of hypothesis formation, methodological justification, literature use and cautious interpretation — would be recognised by readers as a plausible scientific argument. The success of this experiment shows that, as a result of their training on large corpora of scientific publications, language models possess statistical knowledge of the stylistic and argumentative conventions of academic writing. They recognise not only the structures of scientific argumentation, but also the stylistic signals — sometimes referred to as AI markers — used in many discussions to identify machine-generated text. Models can therefore produce coherent scientific narratives that are difficult for human readers to distinguish from conventionally written research. The pigeon experiment thus provides an empirical demonstration of the mechanism described in this section: in AI-generated texts, the rhetorical coherence of a scientific narrative may sometimes be stronger than the empirical necessity on which it is based.

At the same time, generative systems create new opportunities for quality control. They can, for example, be used to analyse large numbers of literature references, test argumentative consistency or generate alternative interpretations of data. This creates new forms of hybrid verification in which human judgement and machine-assisted analysis complement one another.

The analysis of the cases therefore leads to a more precise answer to the first research question. Generative AI expands the possibilities for synthesis, pattern recognition and hypothesis formation, but simultaneously shifts the point at which epistemic vulnerability arises. The quality of AI-driven knowledge production consequently proves less dependent on individual model output than on the research architecture within which that output is generated and evaluated.

18.3 ERQ2 – Pedagogical and Institutional Dimension

The second empirical research question is: '*How does generative AI change the roles, skills and institutional structures of academic research?'* Chapter 12 anticipated that AI would shift the role of researchers from producers of text and analysis to directors of a research process in which AI systems are deployed for different tasks. The five cases partially confirm this expectation, but also show that the institutional impact is strongly domain-specific. In none of the archetypes examined does the role of the human researcher disappear. What does become visible is that research is increasingly organised as an iterative interaction process between human and model.

In historical research, AI functions primarily as a tool for systematically applying analytical frameworks to larger corpora. The researcher remains responsible for selecting sources and for historiographical interpretation. In the anthropological case, generative systems prove capable of reinforcing interpretative frameworks, requiring the researcher to retain an explicitly corrective role. In the design-oriented case, the researcher's role shifts partly towards that of organiser of a generative design trajectory in which different model roles — generation, verification and revision — are separated. This shift in role does not arise automatically, however; it requires an explicitly designed interaction structure. The case in the meta-domain makes this organisational dimension most clearly visible. The dissertation-writing process was organised as a series of iterative interactions between curator and model, while the human researcher remained responsible for the architecture of the argument and the ultimate validation of the reasoning.

The cases therefore reveal a consistent pattern. Generative AI does not replace the researcher, but changes the organisation of research practices. Knowledge production shifts from a linear writing process towards an architecturally organised interaction process in which generation, evaluation and integration are explicitly distinguished.

18.4 ERQ3 – Power and Normativity

The third empirical research question is: '*Which power relations and normative choices are embedded in AI-driven knowledge production?*' Chapter 12 concluded that generative AI may create new dependencies, particularly through the concentration of AI infrastructure in the hands of a limited number of technology companies.

The empirical analysis confirms this diagnosis, but shows that the concrete significance of this power dimension varies considerably across research domains. In several cases, normative choices prove to reside primarily in traditional scientific decisions, such as the selection of research objects, interpretative frameworks and analytical methods. In the historical archetype, for example, the principal normative choice lies in canon formation: the selection of historical episodes regarded as relevant objects of research. AI can assist with analysis, but remains dependent on the sources and selection criteria established by the researcher. In the anthropological case, too, the normative dimension lies primarily in the interpretative framing of the material. Generative AI can reinforce or systematise such frameworks, but does not produce them autonomously.

The natural science case constitutes a special instance. Synthetic data were used there, which at first sight might appear to indicate new forms of epistemic power. On closer examination, however, this proves to be related primarily to resource poverty: limited access to empirical research infrastructure. This asymmetry already existed before the rise of generative AI.

The empirical analysis therefore leads to a qualified conclusion. Generative AI may introduce new infrastructural dependencies and reinforce existing power asymmetries, but in the cases examined, the normative core of scientific knowledge production remains largely connected to human research practices.

18.5 Synthesis and Answer to the Central Research Question

The preceding sections answered the three empirical research questions separately. They showed that the impact of generative AI on scientific knowledge production is not uniform, but varies across research archetypes. Some effects manifest themselves across several research practices, whereas others prove strongly domain-specific. Against this background, the central research question of this dissertation can now be answered.

The answer to the central research question is therefore: '*Generative AI changes scientific knowledge production by reconfiguring the architecture of the research process, thereby enabling research in many cases to proceed more rapidly, broaden interdisciplinarily and become methodologically enriched.*'

The five cases examined show that these three effects are not abstract properties of the technology, but empirically observable patterns within concrete research practices. At the same time, the cases demonstrate that the extent to which these effects occur depends heavily on the type of research and on how the generative process is organised.

The first effect concerns **the acceleration of analysis and synthesis.** In several cases, generative AI proved capable of processing, structuring and comparing large quantities of information on a scale that would be difficult for individual researchers to achieve. This applied, among other things, to the systematic analysis of historical episodes, the structuring of complex conceptual arguments and the generation of coherent academic text proposals. In such contexts, AI can produce considerable time savings in research and writing processes. At the same time, the cases show that this acceleration does not occur uniformly. In design-oriented research practices, such as polyglot language acquisition, the principal bottleneck lies less in analysis than in safeguarding semantic consistency and design stability. The extent to which AI accelerates knowledge production therefore proves highly dependent on the nature of the research practice.

The second effect concerns **the broadening of research practices**. In several cases, generative AI made it practically feasible to combine diverse domains of knowledge within a single research architecture. Historical analysis was linked, for example, to systematic conceptual comparison; questions in public administration were connected to biological constraints; and the design of polyglot language acquisition brought together insights from linguistics, cognitive science and pedagogy. Generative AI did not function as a source of new theories, but as an infrastructure that made it possible to synthesise and interconnect knowledge from different disciplines more rapidly. AI therefore does not automatically increase the interdisciplinarity of science, but it does lower the practical threshold for integrating multiple disciplines within a single research project.

The third effect concerns **the enrichment of methodological control mechanisms**. In several cases, AI proved to be not only an instrument for analysis and text production, but also for verification. Particularly in checking literature use and argumentative consistency, AI made it possible to analyse large numbers of references systematically for interpretative accuracy and internal coherence. This creates new forms of hybrid verification in which human judgement and machine-assisted analysis complement one another. In addition, the use of generative systems can contribute to a forensic reconstruction of the research process, because interactions between researcher and model are explicitly documented and made available for analysis. In this way, AI can contribute not only to knowledge production, but also to the transparency of the research process itself.

The cases also show, however, that these three effects do not arise automatically. Their realisation depends to a considerable extent on how the generative process is organised. When generative AI is used ad hoc within a single interaction with one model, the benefits often remain limited and problems such as semantic drift, context loss or role conflation may arise. By contrast, this dissertation explicitly demonstrates another approach: organising research as an architecture of separated roles in which generation, analysis and verification are explicitly distinguished. This constitutes an important novelty of the study. The cases show that researchers are not dependent solely on the intrinsic capabilities of a language model, but can substantially increase the speed, stability and controllability of AI-driven knowledge production by designing an appropriate interaction architecture. In this respect, the central question concerning AI in science shifts from the capabilities of individual models to the way in which researchers organise the generative process.

The principal conclusion of this dissertation is therefore that **generative AI does not simply automate scientific knowledge production, but transforms its architecture**. Instead of a linear research process, a hybrid research practice emerges in which generation, evaluation and verification are distinguished and organised with increasing explicitness.

18.6 Validity, Reliability and Methodological Safeguards

The methodological reliability of this research rests on three principles: radical transparency, consistent application of the analytical framework and curatorial oversight of the generative process. All cases are accompanied by reflection reports in which interactions between the researcher and the language model are systematically documented. This makes the generative process itself available for analysis. In addition, the same analytical framework was applied across all cases. As a result, the cases function not as isolated examples, but as empirical tests of a single analytical framework. Finally, interpretative responsibility consistently remained with the human researcher. Generative AI produced proposals for analysis and text, but the curator validated and integrated these into the final argument. Within this curatorial architecture, the organisation of the generative process also played a role. The cases show that a multi-agent approach can stabilise the research process and that hybrid verification offers new possibilities for the systematic checking of literature use and argumentation.

18.7 Recommendations for Further Research

The findings of this dissertation open up several avenues for further research. First, comparative research is needed into the domain-specific effects of AI, because acceleration, broadening and enrichment operate differently across research archetypes. Second, the architecture of AI-assisted research warrants further study, particularly the distinction between single-agent and multi-agent approaches. Third, the institutional implications of AI require attention. If researchers shift from authors to curators of generative processes, this may have consequences for academic training and assessment criteria. Finally, further research is required into the normative and ethical dimensions of AI-driven knowledge production, including dependence on digital infrastructures and commercial models.

18.8 Concluding Reflection

The cases in this dissertation show that generative AI does not automate scientific research, but reorganises it. Rather than replacing the researcher, it gives rise to a hybrid research practice in which human expertise and machine-assisted analysis are combined. The researcher's role shifts from individual author to architect and curator of generative research processes. This shift brings new responsibilities, including the design of robust research architectures and the safeguarding of epistemic integrity. The central lesson of this dissertation is therefore that the future of science will not be determined by AI alone, but by the way in which researchers integrate this technology into carefully designed knowledge architectures.

Epilogue – The Socratic Oath

This dissertation began with the question of the possibilities and limitations of AI in scientific knowledge production. The analysis showed that AI is not a neutral instrument, but a technology that places epistemic foundations under pressure. This became particularly apparent in Chapter 2, in the discussion of the black-box character of large language models: researchers do not know precisely how answers are produced, and this in itself poses a threat to the core principles of testability and transparency.

The need for ethical conduct became concrete in Section 4.4, where the problem of fabricated references was first explicitly identified. An apparently minor lapse — a reference that does not exist or a quotation that was never made — can entirely undermine confidence in a study. It became clear here that the researcher's responsibility does not end with the generation of text, but begins with the careful verification of sources.

Chapter 6 brought these insights together in a methodological justification: AI may be used, but only under strict conditions of verifiability, reproducibility and ultimate human responsibility. The discussion thereby shifts from technology alone to ethics. The question is not merely what AI can do, but above all what researchers must do to safeguard the integrity of science.

Confrontation with these risks makes clear that AI-driven science is not merely a methodological choice, but a moral test. Just as physicians take the Hippocratic Oath in their professional practice to protect the vulnerability of the human body, researchers who use AI require an oath that protects the vulnerability of human knowledge: the 'Socratic Oath'.

Why the Socratic Oath? The choice of the name Socratic Oath is not accidental. Just as physicians legitimise their practice through the Hippocratic Oath, researchers in the AI era require a promise that binds them to truth and humanity. Socrates symbolises critical dialogue, the continuous questioning of assumptions and the pursuit of understanding through conversation and self-examination. In this respect, he is related to Hippocrates, who embodied care for the human body; Socrates embodies care for human knowledge.

Other names might also have been appealing — such as the Aristotelian Oath, which would be more empirical; the Turing Oath, which would be more technological; or the Oath of Episteme, which would be more conceptual — but it is precisely the Socratic attitude of continuous reflection and dialogue that lies at the heart of what AI research requires. This is why the oath bears his name.

The Socratic Oath (for researchers who use AI in knowledge production)

General Principles

I swear and promise:

- *That I will conduct my work in the service of humanity and the common good, and never solely in the service of commercial or technological imperatives.*
- *That I will not conduct or publish research that I know, or may reasonably suspect, could cause harm to individuals or society.*
- *That I will use my knowledge and skills to contribute to truth, justice and human dignity.*
- *That I will pursue full transparency concerning the use of AI in my research practice, including the limitations and uncertainties it entails.*

Specific Obligations When Using AI

I swear and promise:

1. *Verification of Sources*
    - *I will never use AI-generated references or quotations without verification.*
    - *I will verify every source individually and account for it correctly.*
2. *Dealing with Black-Box Systems*
    - *I will never ignore or conceal the black-box character of AI.*
    - *I will actively seek ways to make the operation of the model explicable and to make its epistemic limits explicit.*
3. *Bias and Normative Content*
    - *I will always investigate which assumptions and biases are embedded in the data and models.*
    - *I will identify and document them and, where possible, mitigate them.*
4. *Human Responsibility*
    - *I will at all times retain ultimate responsibility for the content and consequences of my research.*
    - *I will not shift responsibility onto the AI system or its provider.*
5. *Data Integrity and Consent*
    - *I will respect the autonomy and rights of individuals and communities at every stage of data collection, analysis and reuse.*
    - *I will not use data without adequate consent or justification.*
6. *Countervailing Power and Public Values*
    - *I will resist the uncritical pursuit of commercial interests or concentrations of power in AI development.*
    - *I will advocate openness, accessibility and public oversight of knowledge infrastructures.*
7. *Testability and Reproducibility*
    - *I will publish AI-facilitated findings only when they are testable, reproducible and methodologically sound.*
    - *I will document my work so that others can scrutinise and criticise my process.*

Concluding Reflection. This oath redefines the place of AI within science. AI is not a neutral instrument, but an actor that influences the nature of knowledge. This awareness formed a recurring thread throughout the analyses in this dissertation: from the early identification of the black-box problem in Chapter 2, through the warning about fabricated references in Chapter 4, to the methodological recalibration in Chapter 6.

The humanity of science can therefore no longer be assumed to be guaranteed: it must be safeguarded explicitly. The Socratic Oath provides a normative compass for doing so. Those who take this oath commit themselves not merely to a set of rules, but to an attitude of responsibility, transparency and fidelity to humanity. The AI era thereby becomes not only a technological revolution, but also a moral test. Science will succeed in preserving its dignity for as long as researchers are willing to acknowledge the duty that binds them: the duty to serve truth and never lose sight of humanit