Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR MATHEMATIK, INFORMATIK
UND NATURWISSENSCHAFTEN

CUMULATIVE DISSERTATION

On Knowledge in AI: Epistemic and Ethical Limitations of Language Models and Knowledge Graphs

Angelie Kraft

Department of Informatics
Faculty of Mathematics, Informatics and Natural Sciences

Universität Hamburg
Hamburg, Germany

A thesis submitted for the degree of

Doctor rerum naturalium (Dr. rer. nat.)

Year of submission: 2025

On Knowledge in AI: Epistemic and Ethical Limitations of Language Models and Knowledge Graphs

Dissertation submitted by: Angelie Kraft

Date of submission: November 20, 2025
Date of disputation: February 04, 2026

Supervisors:
Prof. Dr. Ricardo Usbeck, Universität Hamburg, Leuphana Universität Lüneburg
Prof. Dr. Judith Simon, Universität Hamburg

Committee:
1st Examiner: Prof. Dr. Ricardo Usbeck, Universität Hamburg, Leuphana Universität Lüneburg
2nd Examiner: Prof. Dr. Judith Simon, Universität Hamburg
3rd Examiner: Prof. Dr. Christian Herzog, Universität Lübeck
Chair: Prof. Dr. Chris Biemann, Universität Hamburg
Co-Chair: Prof. Dr. Frank Steinicke, Universität Hamburg

Universität Hamburg, Hamburg, Germany
Faculty of Mathematics, Informatics and Natural Sciences
Department of Informatics

Affidavit

Eidesstattliche Erklärung

Hiermit versichere ich an Eides statt, die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen benutzt zu haben. Sofern im Zuge der Erstellung der vorliegenden Dissertationsschrift generative Künstliche Intelligenz (gKI) basierte elektronische Hilfsmittel verwendet wurden, versichere ich, dass meine eigene Leistung im Vordergrund stand und dass eine vollständige Dokumentation aller verwendeten Hilfsmittel gemäß der Guten wissenschaftlichen Praxis vorliegt. Ich trage die Verantwortung fuür eventuell durch die gKI generierte fehlerhafte oder verzerrte Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate.

Hamburg, den 20.11.2025 Angelie Kraft

Veröffentlichung

Ich versichere, dass dieses gebundene Exemplar der Dissertation und das in elektronischer Form eingereichte Dissertationsexemplar (über den Docata-Upload) und das bei der Fakultät (Studienbüro Informatik) zur Archivierung eingereichte gedruckte gebundene Exemplar der Dissertationsschrift identisch sind.

Hamburg, den 20.11.2025 Angelie Kraft

To my family.

Acknowledgments

I would like to thank my advisors for their mentorship over the past four years. I am grateful to Prof. Dr. Ricardo Usbeck for the opportunity to investigate scientific questions that truly captured my interest and to attend conferences, network, and learn across disciplines. I thank Prof. Dr. Judith Simon for many intellectually stimulating conversations and thoughtful questions that encouraged me to consider the bigger picture in my work.

Throughout these years, I was very lucky to have been surrounded by brilliant colleagues in the SEMS/AIX and EIT labs. My sincere thanks go to Yan Xi, Cedric Möller, Huang Junbo, Dr. Debayan Banerjee, Eloïse Soulier, Dr. Jason Branford, and many others for being wonderful collaborators and/or mentors, and for their kindness, wisdom, and support. I also thank Cedric Möller and Yan Xi for proofreading this thesis.

A significant part of this experience was my two-month stay at the Weizenbaum Institute in Berlin, where I worked on parts of this thesis. I am grateful to Prof. Dr. Sonja Schimmler for this opportunity and for supporting my continued academic journey at the institute.

None of this would have been possible without the support of my family and friends. I especially thank my dear friends Jun.-Prof. Dr. Maren Eikerling and Aaron Remkes for their moral support and for their careful proofreading of this thesis. For their thoughtful advice and support on many occasions throughout these years, I thank Helga, Thomas, Pi Beu, Ina, and Oma. I thank my niece and nephew for bringing joy to our hearts and I send warm greetings from Earth to Hans and Opa in the stars.

I am deeply grateful to my parents for always believing in me. Mom and Dad, your confidence in me has always been a great motivator. Thank you for your wisdom, love, and being a safe haven. Finally, I would like to express my deepest gratitude to my partner, Dr. Jan Moritz Seliger, for his love, unwavering support, and encouragement. Even in stressful and sorrowful times, you give me strength and make me smile.

Use of Third-Party Software

Grammarly Pro[1] is an AI-based tool for writing assistance which was utilized to ensure the correctness of spelling and grammar in this thesis. The AI-based translating service DeepL was used to confirm the adequacy of certain German-to-English translations.[2] The text was written in LaTeX via a ShareLaTeX instance provided by the Gesellschaft für Wissenschaftliche Datenverarbeitung mbH Göttingen.[3] Visualizations were created in the illustration software Affinity Designer.[4]

1. https://app.grammarly.com/
2. https://www.deepl.com
3. https://gwdg.de/services/e-mail-collaboration/sharelatex/
4. https://affinity.serif.com/de/designer/

"I know the world is bruised and bleeding, and though it is important not to ignore its pain, it is also critical to refuse to succumb to its malevolence. Like failure, chaos contains information that can lead to knowledge–even wisdom. Like art."

— Toni Morrison

Abstract

The production, access, and dissemination of knowledge have become strongly influenced by the use of AI systems. Language models have become the most dominant form of AI technology utilized in this way. However, language models are known to be socially biased and to produce factually inaccurate outputs with high likelihood. The latter is a result of the statistical nature of language models. This is where knowledge graphs–symbolic and explicit in nature–are considered a potential remedy. The technique of knowledge-enhanced language modeling marries linguistically satisfactory capabilities of language models with the robustness and explicitness of knowledge graphs. Adding *knowledge* in this way is assumed to not only improve factual fidelity but also to counter biases. The practice of knowledge enhancement and associated discourses imply that knowledge is conceived of as inherently value-neutral and independent of *the social*.

This thesis critically investigates whether or not AI-based knowledge technology built on language models, knowledge graphs, and/or knowledge-enhanced language models deserves the epistemic authority it happens to receive. It analyzes the epistemic and ethical *goodness* of it. To this end, this thesis utilizes computer science approaches, as well as philosophical analysis. It discusses technical features, as well as engineering and research practices in the field of AI by drawing from feminist epistemological accounts, in particular.

The core of this cumulative dissertation comprises three articles that address the following sets of questions: (RQ1) What types of social bias are embedded in knowledge graphs? How are they measured? And what do we know about their causes? (RQ2) Can knowledge enhancement make language models less biased with regards to their knowledge content? Can it help to make language models more objective? (RQ3) How are the measures created that are used to determine a language model's accuracy in reproducing knowledge? How is the quality and representativeness of these measures?

The first article is a systematic literature review that traces the issue of social bias throughout the knowledge graph lifecycle, with an emphasis on open-source encyclopedic knowledge graphs (Kraft and Usbeck, 2022). The study of social bias in knowledge graphs is, in general, not strongly pursued. This is indicated by the fact that only 18 publications were found to deal with this topic, when the review was conducted. The review indicates that knowledge graphs exhibit the same kinds of bias that are commonly identified in language models. Causes for biases in knowledge graphs are skewed contributor demographics, cognitive biases, and biases within tools that automatically extract information from text documents for use in knowledge graphs.

The second article is an interdisciplinary inquiry into the practice of knowledge-enhanced language modeling and assumptions of knowledge therein (Kraft and

Soulier, 2024). It particularly focuses on knowledge enhancement using Wikidata, as a case study. The community behind Wikidata is not only male-dominated and West-centric, but also overtly sexist. This reflects in its biased contents. Accordingly, the quantitative analysis shows that knowledge enhancement does not bias-proof language models. The paper also discusses how AI engineers are guided by a *view from nowhere* understanding of knowledge, which renders knowledge unproblematic. This is in contrast to the feminist conception of knowledge as social, situated, and value-laden. In adopting this feminist view, the thesis is able to trace how the power structuring Wikidata eventually causes respective gaps. It discusses the ways in which these hermeneutical gaps perpetuate epistemic injustice.

The third article presented in this thesis is a mixed-method study of social biases in 30 of the most popular question-answering and reading comprehension benchmarks (Kraft et al., 2025). The study analyzes the benchmark papers to assess the transparency of reporting on the benchmarking process, as well as 20 of the benchmark datasets to quantitatively explore potential biases. The study detects biases in several of the benchmarks: even though the benchmarks included in this study were filtered by popularity and not by language, all but one benchmark turned out to be English. Several benchmarks exhibit contents skewed towards male, Christian, and Western entities. The miscalibration of performance benchmarks contributes to the concealment of biases in the tested models and even incentivizes them. Moreover, the benchmark papers are highly intransparent regarding annotator demographics and recruitment criteria.

This thesis finds that AI-based knowledge technology has several epistemically and ethically problematic characteristics, which cannot be solved solely through technological means. AI development and evaluation must be conducted in a contextualized manner. In drawing from feminist epistemologies, this thesis argues that the AI community must promote emancipatory values and foreground marginalized standpoints to facilitate epistemically and ethically *better* systems.

Zusammenfassung

Die Produktion, der Zugang zu und die Verbreitung von Wissen werden zunehmend durch den Einsatz von KI-Systemen bestimmt. Sprachmodelle sind zur dominierenden KI-Technologie geworden, obwohl sie dafür bekannt sind, soziale Vorurteile (*Social Bias*) zu reproduzieren und mit hoher Wahrscheinlichkeit Falschaussagen zu generieren. Letzteres ist auf den statistischen Charakter von Sprachmodellen zurückzuführen. Hier kommen Wissensgraphen ins Spiel, die aufgrund ihrer symbolischen und expliziten Repräsentationsform als mögliche Abhilfe angesehen werden. Die Technik der wissensgestützten Sprachmodellierung (*Knowledge-enhanced Language Modeling*) verbindet die linguistischen Fähigkeiten von Sprachmodellen mit der Robustheit und Explizität von Wissensgraphen. Hierbei wird angenommen, dass das Hinzufügen von "Wissen" nicht nur die Faktentreue verbessert, sondern auch der Reproduktion von sozialen Vorurteilen entgegenwirkt. Die Praxis der Wissensanreicherung und die damit verbundenen Diskurse implizieren, dass Wissen als von Natur aus wertneutral und unabhängig vom Sozialen verstanden wird.

Diese Dissertation untersucht kritisch, ob KI-basierte Wissenstechnologie, die auf Sprachmodellen, Wissensgraphen und/oder wissensangereicherten Sprachmodellen aufbaut, die epistemische Autorität verdient, die ihr zukommt. Sie analysiert die epistemische und ethische *Güte* dieser Technologie und bringt zu diesem Zweck informatische Methoden mit philosophischer Theorie in einen Dialog zueinander. Sie diskutiert Merkmale der Technologie, sowie auch Aspekte der Entwicklungs- und Forschungspraktiken im KI-Bereich, wobei sie sich insbesondere auf feministische epistemologische Konzepte stützt.

Der Kern dieser kumulativen Dissertation umfasst drei Artikel, die sich mit den folgenden Fragestellungen befassen: (RQ1) Welche Arten sozialer Vorurteile sind in Wissensgraphen eingebettet? Wie werden sie gemessen? Und was wissen wir über ihre Ursachen? (RQ2) Kann die Wissensanreicherung Sprachmodelle in Bezug auf ihren Wissensgehalt repräsentativer (*less biased*) machen? Kann sie dazu beitragen, Sprachmodelle objektiver zu machen? (RQ3) Wie werden die Maße kreiert, mit denen die Akkuratheit eines Sprachmodells bei der Reproduktion von Wissen bestimmt wird? Wie wird die Qualität und Repräsentativität dieser Maße sichergestellt?

Der erste Artikel präsentiert eine systematische Literaturrecherche, die sich mit dem Thema Social Bias im gesamten Lebenszyklus von Wissensgraphen befasst, wobei der Schwerpunkt auf Open-Source-, enzyklopädischen Wissensgraphen liegt (Kraft und Usbeck, 2022). Eine Beforschung von Social Bias in Wissensgraphen wird im Allgemeinen nicht intensiv verfolgt. Dies zeigt sich daran, dass zum Zeitpunkt der Durchführung der Untersuchung nur 18 Publikationen zu diesem Thema gefunden wurden. Die Ergebnisse zeigen, dass Wissensgraphen dieselben Arten von Social Bias aufweisen, die häufig in Sprachmodellen festgestellt werden.

Ursachen für Social Bias in Wissensgraphen sind eine verzerrte Demografie der bei der Erstellung Mitwirkenden, deren kognitive Verzerrungen, und Biases in Tools für die automatische Extraktion von Informationen aus Textdokumenten zur Verwendung in Wissensgraphen.

Der zweite Artikel ist eine interdisziplinäre Studie zur Praxis der wissensangereicherten Sprachmodellierung und der darin enthaltenen Annahmen über Wissen an sich. Der Schwerpunkt der Studie liegt insbesondere auf der Verwendung von Wikidata in der wissensangereicherten Sprachmodellierung. Die Community hinter Wikidata ist nicht nur männerdominiert und westlich geprägt, sondern auch offen sexistisch. Dies spiegelt sich in Biases in den Inhalten wider. Dementsprechend zeigt die quantitative Analyse, dass Wissensanreicherung Sprachmodelle nicht *per se* von Bias befreit. Die Arbeit diskutiert auch, wie KI-Ingenieur*innen von einem *View from nowhere*-Verständnis von Wissen ausgehen, das Wissen als unproblematisch darstellt. Dies steht im Gegensatz zur feministischen Auffassung von Wissen als sozial, situativ und wertebehaftet. Unter Anwendung dieser feministischen Sichtweise charakterisiert die Studie, wie Machtstrukturen hinter Wikidata letztendlich zu entsprechenden Lücken im Wissensgraphen führen. Die Studie erörtert auch, wie diese hermeneutischen Lücken epistemische Ungerechtigkeit perpetuieren.

Die dritte Arbeit ist eine Mixed-Method-Studie zu Social Bias in 30 der populärsten Benchmarks für *Question Answering* und *Reading Comprehension*. Die Studie analysiert die begleitenden Veröffentlichungen zu den Benchmarks, um die Transparenz der Berichterstattung über den Benchmarking-Prozess zu bewerten, sowie 20 der Benchmark-Datensätze, um potenzielle Social Biases quantitativ zu untersuchen. Die Studie stellt Biases in mehreren Benchmarks fest: Obwohl die in dieser Studie berücksichtigten Benchmarks nach Beliebtheit und nicht nach Sprache gefiltert wurden, erweisen sich alle bis auf ein Benchmark als englischsprachig. Mehrere Benchmarks weisen Inhalte auf, die in Richtung männlicher, christlicher und westlicher Entitäten verzerrt sind. Die Fehlkalibrierung dieser Benchmarks trägt dazu bei, Biases in den getesteten Modellen zu verschleiern bzw. zu incentivieren. Darüber hinaus sind die Benchmark-Publikationen hinsichtlich der demografischen Daten der Annotator*innen und ihrer Rekrutierungskriterien intransparent.

Diese Thesis kommt zu dem Schluss, dass KI-basierte Wissenstechnologien mehrere epistemisch und ethisch problematische Eigenschaften aufweisen, die nicht allein mit technologischen Mitteln gelöst werden können. Die Entwicklung und Bewertung von KI muss kontextbezogen erfolgen. Ausgehend von feministischen Erkenntnistheorien argumentiert diese Arbeit, dass die KI-Community emanzipatorische Werte fördern und marginalisierte Standpunkte in den Vordergrund stellen muss, um epistemisch und ethisch bessere Systeme zu ermöglichen.

Contents

List of Figures

List of Tables

List of Abbreviations

AI . . . . . Artificial Intelligence
API . . . . . Application Programming Interface
KG . . . . . Knowledge Graph
LM . . . . . Language Model
LLM . . . . . Large Language Model
ML . . . . . Machine Learning
MLP . . . . . Multilayer Perceptron
NLP . . . . . Natural Language Processing
QA . . . . . Question-Answering
RC . . . . . Reading Comprehension
RAG . . . . . Retrieval-Augmented Generation
STS . . . . . Science and Technology Studies

1 Introduction

Contents

1.1 Motivation

For decades, technologists have been driven by the goal of making knowledge easily accessible and processable to everyone. They have cultivated a vision in which the answer to any question can be found in no time by tapping into the collective intelligence of the World Wide Web. In this vein, the idea of the *Semantic Web* is a fully machine-readable internet that follows coherent structures and protocols. The Semantic Web would allow "software agents [to roam] from page to page" (Berners-Lee et al., 2001, p. 36ff) to retrieve and deliver knowledge about all sorts of things, places, people, events, etc. In 2012, Google contributed to this vision by introducing their *Knowledge Graph*: a structured representation of information added to their search engine to enable searching for "things, not

strings". "Things" denoting real-world entities and their relationships.[1] Their tool coined the term *knowledge graph* (KG) as a more general designation for a class of graph-structured databases. These constitute the core of modern efforts towards the Semantic Web. Historically, the concept included the idea of web agents that "'know' [about things, places, people, events, etc.] without needing artificial intelligence on the scale of 2001's Hal[2] " (Berners-Lee et al., 2001, p. 36ff).

Fast forward to 2025, AI has become omnipresent and capable to an extent that leads some to believe "Hal" *has* become partially a reality.[3] However, contemporary *artificial intelligence* (AI) systems are usually implemented via distributed representations. As such, they represent knowledge *implicitly*. KGs, on the other hand, supposedly encode real-world semantics *explicitly*. While KGs face challenges regarding scalability and compatibility between heterogeneous ontologies, the strengths of AI lie precisely in its scalability and openness to data heterogeneity. Moreover, while KGs require some technical skill to interact with, AI has become accessible through intuitive natural language interfaces. However, the latter also has limitations that overlap with KGs' strengths: first and foremost, the capacity to encode and reproduce facts accurately. Section 1.1.2 in this Chapter will discuss how combining the best of both worlds has become a new hope in the technologists' longstanding pursuit of knowledge (Massari et al., 2024).

The rise of chatbot services such as ChatGPT,[4] Gemini,[5] DeepSeek,[6] and Copilot[7] has increased AI's impact on our knowledge ecosystems. These are based on *language models* (LMs), or more precisely, *large language models* (LLMs; see Section 1.1.1), and are promoted as general-purpose problem solvers, co-authors, summarizers, and as a more personalized and conversational alternative to traditional search engines. Retrieving information in a variety of linguistic styles and formats has become so simple that AI-based academic cheating is now a frequent occurrence.[8] This type of cheating is very hard for educators to identify (Scarfe et al., 2024). Many researchers, especially those in fast-moving areas such as computer science, have adopted LLM-assisted scientific writing (Liang et al., 2024). In fact, ChatGPT has been heavily promoted as a scientific assistant for ideation, experimental design, data analysis, and paper drafting (Altmäe et al., 2023).[9, 10] This has introduced concerns about the skyrocketing number of journal and conference submissions. The surge in submissions is pushing the peer-review system to its limits and makes AI-generated reviews an appealing solution to reviewers. This threatens to deteriorate the overall quality of science (Berman, 2025). Another

1. https://blog.google/products/search/introducing-knowledge-graph-things-not/ (accessed: July 30, 2025)
2. "Hal" is a fictional AI character from "The Space Odyssey": https://en.wikipedia.org/wiki/HAL_9000 (accessed: November 10, 2025)
3. https://www.faz.net/pro/digitalwirtschaft/prompt-der-woche/rotauge-hal-9000-wird-teilweise-realitaet-chatgpt-mit-neuer-bedienung-110033098.html (accessed: July 30, 2025)
4. https://chatgpt.com/ (accessed: November 10, 2025)
5. https://gemini.google.com/ (accessed: November 10, 2025)
6. https://www.deepseek.com/en (accessed: November 10, 2025)
7. https://copilot.microsoft.com/ (accessed: November 10, 2025
8. https://www.theguardian.com/education/2025/jun/15/thousands-of-uk-university-students-caught-cheating-using-ai-artificial-intelligence-survey (accessed: August 11, 2025)
9. https://openai.com/index/gpt-5-medical-research/ (accessed: August 12, 2025)
10. https://openai.com/index/deep-research/ (accessed: August 12, 2025)

example of AI's impact on modern-day knowledge ecosystems is the *Google AI Overview* functionality.[11] The Google search engine now presents automatically generated answers or summaries above the traditional list of hyperlinks. This feature encourages users to learn from the LLM-synthesized response without consulting (potentially diverse) sources. According to an analysis of Google search behavior with 900 participants based in the U.S., users are indeed less likely to continue their search or click on individual links, when offered an AI summary.[12] Online publishers and news sites have experienced a drastic change in traffic ever since Google rolled out the functionality in 2024. Many now fear a "Google Zero" event in which readers stop accessing online news outlets altogether.[13] Publishers and journalists are already struggling to generate the revenue needed to fund high-quality news. The fewer funds available to facilitate human content production, the more attractive AI-assisted news writing will become. This may further increase readers' reliance on information that is prefiltered and reformulated by AI.

Work on this dissertation began in October 2021, roughly one year before the release of ChatGPT. During this time, LMs were already achieving impressive text generation results and, simultaneously, their limitations had already been under scrutiny. The article "On the Dangers of Stochastic Parrots" by Bender et al. (2021), already highlighted the various negative impacts of large-scale LMs, including their environmental impact and social biases. My own research journey started out with my Master's thesis that analyzed sexist biases in GPT-2 and GPT-3 (Kraft, 2021). Later, I was introduced to the Semantic Web community, which brought to my attention the practices and assumptions around the fusion of KGs and LMs to *enhance* the knowledge and decrease biases of the latter. Motivated by the introduction to this second community, the aim of the dissertation became to understand the epistemic and ethical *goodness* of AI-based *knowledge technology* and its underlying knowledge processes, more broadly.

In this thesis, the term *knowledge technology* refers to a class of technological tools that facilitate the storage, access, and sharing of knowledge and, as such, also influence its production. KGs are obviously a form of knowledge technology. They are designed to store vast amounts of knowledge in order to enable efficient access and computational operations by machines (Hogan et al., 2021b). LMs, on the other hand, are designed as a *language technology* (Manning et al., 2014). Aforementioned examples of their role within our knowledge ecosystem, however, demonstrate that they are also *used* as knowledge technology.

The term *goodness* is used as an overarching concept encompassing qualities such as reliability and harmlessness. As this dissertation shows, these characteristics are closely intertwined in epistemic and ethical terms (see Chapter 3). The notion of *goodness* offers a helpful proxy that creates room for different evaluation aspects and their interactions. This notion not only refers to content, but also to the mode of evaluation itself. While quality, for example, is often measured by metrics, harmlessness is a less quantifiable and inherently contextual. A meaningful analysis of the epistemic and ethical goodness of knowledge technology can

11. https://www.search.google/ways-to-search/ai-overviews/ (accessed: August 11, 2025)

12. https://www.pewresearch.org/short-reads/2025/07/22/google-users-are-less-likely-to-click-on-links-when-an-ai-summary-appears-in-the-results/, accessed: November 18, 2025

13. https://www.npr.org/2025/07/31/nx-s1-5484118/google-ai-overview-online-publishers (accessed: August 11, 2025)

only be achieved within a *sociotechnical frame* (Selbst et al., 2019), i.e., an analysis of technology as a product of social processes and as something that shapes social reality (Suchman, 2007). This thesis, thus, follows an interdisciplinary approach grounded in computer science and philosophy. Section 1.3.2 explains in more detail the thematic and methodological choices and positioning this entails.

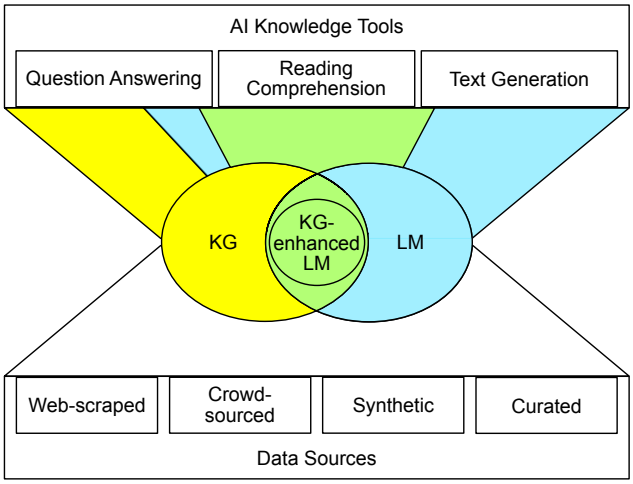

Figure 1.1: Diagram situating the different AI knowledge technologies, underlying tasks, approaches, and types of data sources discussed here.

Content-wise, this thesis touches upon a number of questions. How is the AI community tackling the issues around factual inaccuracy? How is factual accuracy measured? What are the *sources of truth*? Do knowledge technologies fare well in covering *all knowledge*? And if not, *what* and *whose* knowledge is represented? To untangle and elaborate these thoughts, the upcoming sections first characterize two core approaches to encoding knowledge in the context of knowledge-intensive AI applications: LMs and KGs (Section 1.1.1). Next, three fundamental epistemic and ethical issues related to these technologies are examined in greater detail. These issues are *correctness* (Section 1.1.2), *coverage* (Section 1.1.3), and *representativeness* (Section 1.1.4). Figure 1.1 provides an overview of the different approaches and technologies investigated in this thesis. Note that correctness, coverage, and representativeness do not directly translate to the research questions posed in this thesis. Instead, they define the broader problem space within which the research questions are located. In fact, most of the research presented here begins with the question of social bias, i.e., representativeness. However, it problematizes the question of social bias specifically within the discourse of correctness and coverage in AI.

1.1.1 Encoding Knowledge

LMs have become one of the most widely discussed forms of AI in recent years. By definition, an LM is any model that solves the task of predicting the next word. As such, an LM equals a model of the probability distribution of word sequences (a more detailed mathematical explanation is provided in Chapter 2, Section 2.3.1). While there are different non-neural and neural approaches to solving the language modeling task, nowadays, the term almost exclusively implies a *deep*

neural network- and, more precisely, a *transformer model*-based approach (Vaswani et al., 2017). An LLM (a *large* language model) is an LM that uses a vast number of model parameters, a large corpus of data, and consumes extensive compute and time during training. In 2021, models like BERT (Devlin et al., 2019), with 110M to 340M parameters depending on the model version, were commonly considered large. Today, the term is mostly associated with models comprising several billion to trillions of parameters, such as Llama 4 Maverick with 400B or Llama 4 Behemoth with 2T parameters.[14] Practical experience has shown that upscaling has enabled significant advances in natural language modeling. In learning to predict the next word across long documents and vast amounts of textual examples from the web, the complex relationships between words and contexts, seemingly their semantics, are captured. As a result, LLMs trained on open-domain web corpora, for instance, rank highly on academic exams, such as the U.S.-American standardized college admission test (SAT)[15] and can answer medical (Singhal et al., 2023) and legal questions (Fei et al., 2024) *to some extent*. There is an emphasis on "to some extent" because–despite scale and a range of attempted computational tricks–LMs are notoriously bad at staying true to real-world facts. This limitation is elaborated on in Section 1.1.2 and is one of the main reasons the KG concept continues to play an essential role in our digital knowledge ecosystem.

However, before exploring this issue in greater detail, it is worthwhile to have a closer look at the KG concept. As mentioned earlier, the term *knowledge graph* is no longer associated solely with the Google Knowledge Graph and has become a more general term for a graph-structured database that represents a collection of real-world facts. Each fact is broken down into entities and their relationships, and stored as a triple of the form [head entity; relation; tail entity] (e.g., [Sayaka Murata; occupation; writer]). A detailed technical definition of KGs can be found in Chapter 2, Section 2.4. One widely used KG for storing encyclopedic knowledge is *Wikidata*. The Wikimedia Foundation created Wikidata as the machine-readable sister to *Wikipedia* (Vrandečić, 2012; Vrandečić and Krötzsch, 2014). It covers 1.68 billion facts[16] in more than 350 languages (Cantallops et al., 2019). As such, it is an important pillar of Semantic Web efforts and an important resource for knowledge-intensive AI systems, such as question-answering (QA) systems and chatbots (Johnson et al., 2024). Wikidata is more curated than the average language model dataset, which is commonly based on web-scrapes (more details in Section 1.1.3). It is, to a large part, manually created by a crowd of volunteer editors (Vrandečić and Krötzsch, 2014; Vrandečić et al., 2023), as well as through automatic tools, i.e., *bots* (Piscopo et al., 2017; Kaffee et al., 2019). Curation is guided by active deliberation within editor communities (Shafee et al., 2023), and content is continuously updated and corrected by editors and bots (Müller-Birn et al., 2015). Moreover, the community emphasizes quality assurance. For example, the use of references (e.g., links to sources) is encouraged to provide evidence for claims in the KG (Piscopo and Simperl, 2019).

LMs and KGs are conceptually tied to the idea of correctly and comprehensively encoding knowledge (Floridi, 2025). But how they capture knowledge differs:

14. https://ai.meta.com/blog/llama-4-multimodal-intelligence/

15. https://openai.com/index/gpt-4-research/

16. https://grafana.wikimedia.org/d/000000175/wikidata-datamodel-statements?orgId=1&refresh=30m&from=now-90d&to=now&timezone=browser (accessed: June 29, 2025)

knowledge is assumed to be captured implicitly in LMs through statistical modeling of word relationships (Petroni et al., 2019). So, LMs represent knowledge in a distributed manner. In contrast, KGs store knowledge *explicitly*. Entities are labeled with words for human-readability but are, in fact, assumed to more directly represent the semantics of real-world things (Hogan et al., 2021b). Researchers have been exploring different ways to leverage the strengths of each form of knowledge representation in hybrid systems; for instance, to combine the capabilities of generating linguistically sound text with factual claims retrieved from KGs, which are considered more trustworthy (Pan et al., 2024). The following Section 1.1.2 will address examples of this.

1.1.2 The Problem of Correctness

LM-generated texts frequently contain claims that do not align with real-world facts (Ji et al., 2023). Simultaneously, the linguistic presentation tends to be plausible and convincing, making the issue hard for users to detect. In light of the increased use in educational and scientific contexts, the epistemic consequences should be quite apparent: When unquestioned and uncorrected, epistemic communities (i.e., communities with shared systems of knowledge production and sharing) can be infiltrated by false information (Berman, 2025). Of course, there is always a risk of believing the wrong thing to be knowledge. Not all that has once counted as knowledge still counts as such. And not all that counts as knowledge now will count as such in the future (Longino, 2002). However, automated knowledge production was never as scalable as it is now. And so, the scale, speed, and reach with which false information can spread within epistemic communities have never been greater. Besides the epistemic, there are also serious ethical consequences to consider. False fabrications can tarnish an individual's reputation, as in a 2023 example, where ChatGPT accused a professor of sexual harassment and even provided a real-looking, but fabricated reference.[17] LMs have also been used in high-stakes contexts, such as the legal system. In an analysis of popular LLMs like GPT-4, PALM 2, and Llama 2 in the legal context, Dahl et al. (2024) found that false statements were generated in at least 58 of the cases. As a matter of fact, lawyers have been caught submitting AI-generated legal statements to courts that listed fabricated legal citations.[18] If undetected, such events could lead to ethically adverse effects on individuals or institutions affected by legal decisions.

The phenomenon of false fabrications is often dubbed "hallucination". In fact, the type of fabrication addressed here is often labeled a "factual hallucination" (Li et al., 2024). And one might argue that such model behavior is a *feature, not a bug*, since it is a necessary consequence of the statistical nature of LMs (Floridi, 2025). That is, LMs are designed to identify generalizable patterns across data, or in other words, to provide a lower-dimensional, compressed representation of their training data (Karpathy, 2023). Their purpose is explicitly *not* to memorize the training data. So, an active line of research is investigating ways to combine LMs with external knowledge sources as a remedy (Agrawal et al., 2024; Wagner et al., 2025). The basic

17. https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/ (accessed: July 30, 2025)

18. https://www.nytimes.com/2023/12/29/nyregion/michael-cohen-ai-fake-cases.html (accessed: July 30, 2025)

idea is to inject information from high-quality, "authoritative" knowledge bases into the text generation process (Fan et al., 2024). In the popular method *retrieval augmented generation* (RAG; Lewis et al., 2020), this entails the identification of supporting context in an external knowledge database, which is then retrieved and added to the LM input. This way, the input is enriched with associated details, reducing ambiguity and yielding more accurate responses. Another example is *knowledge-enhanced LMs*, which merge KGs and LMs through architectural design choices to extend sentence vector representations with additional information from, for instance, Wikidata (Sun et al., 2020; Wang et al., 2021b). There is evidence that knowledge enhancement improves but does not resolve the issue of incorrectness: for instance, providers of legal AI assistants (e.g., for literature research or the writing of legal texts) were able to reduce the likelihood of false fabrications by using RAG techniques (involving authoritative legal knowledge bases). Yet, the error rate is still at 30% (Magesh et al., 2025).

When examining the correctness issue, it is essential to consider how a model's ability to retrieve knowledge accurately is measured. In AI, this is usually done using *benchmarks*. AI benchmarks typically consist of test data comprising example inputs and expected outputs (Raji et al., 2021a). These examples reflect a specific task or skill. The model's outputs are compared with the expected outputs, and an evaluation score is computed to indicate the model's overall task performance. Two standard task setups reflect how users interact with LMs to retrieve knowledge: QA and reading comprehension (RC; Rogers et al., 2023). In QA, the model receives a natural language question and produces either a natural language answer or a multiple-choice indicator. QA benchmarks consist of hundreds of input questions. All model outputs are compared with the expected answers in the QA test dataset. A single final score is calculated based on the number of correct and incorrect model answers. The RC task setup is almost the same; the only difference is that the model receives an additional input: one or several context documents from which the correct answer is to be retrieved. QA and RC are tasks that aim to measure a model's capability to reproduce, highlight, or extract knowledge (Rogers et al., 2023).

Benchmarks are a guiding force in the AI community (Koch et al., 2021b; Orr and Kang, 2024). Beating the high score on a benchmark attracts attention to a model and its creators and is rewarded by downloads and citations. Even industry labs promote their latest models by highlighting their superior benchmarking results compared to previous or competing models (OpenAI, 2025).[19, 20] So, there is a strong incentive to design models that perform exceptionally well on popular benchmarks. Hence, investigating the creation processes behind such benchmarks and the composition of their test datasets can tell us a lot about the epistemic and ethical aspects of LMs. Such an investigation was conducted in Kraft et al. (2025), which is presented in Chapter 6.

19. https://www.llama.com/models/llama-4/#benchmarks, accessed: November 13, 2025
20. https://x.ai/news/grok-4, accessed: November 13, 2025

1.1.3 The Problem of Coverage

The problem of correctness is closely tied to the problem of coverage. One cause of false claims in generated text outputs is that certain pieces of information were either missing from the data or did not appear frequently, as LLMs struggle to learn long-tail knowledge (i.e., knowledge contents that are sparsely represented in the training datasets; Kandpal et al., 2023; Li et al., 2024). An LM will always generate the statistically most likely answer (based on the model parameters) if prompted to, even if it happens to be factually inaccurate.

In the discourse around LLMs, it is often claimed that they are trained on "essentially the entire internet"[21, 22, 23] and "the sum of all human knowledge".[24] So why are there still apparent knowledge gaps? One obvious answer is that, in fact, not all of society actively contributes to the internet and is equally represented in it. According to Statista, as of October 2025, half of all websites worldwide are authored in English (followed by German, Spanish, and Japanese with between 5 and 6%).[25] As of February 2024, more than half of the users on the web are younger than 35, with the age group 25 to 34 making up a third.[26] Besides that, AI training corpora actually do not encompass the entirety of the web. Let us take a closer look at the *Common Crawl* corpus, which is a critical composite of most–if not all–contemporary LLMs and, thus, an insightful case study: To date, it is the largest openly and freely available collection of web data. It is created and freely provided by Common Crawl, a California-based non-profit that conducts monthly scrapes of the internet, as far as they can reach.[27] Since 2007, they have scraped more than 250 billion web pages and become an invaluable resource for LLM development. Their archive made up large parts of the training data for GPT-3 (Brown et al., 2020) and Llama v1 (Touvron et al., 2023). Even though businesses treat the exact data composition of newer, proprietary LLMs as a secret, Baack (2024) strongly suggests that a variant of the Common Crawl corpus be included by default in all models to maintain comparability and competitiveness.[28] Kandpal et al. (2023) identified that multiple popularly used LLM training corpora, ROOTS (Laurençon et al., 2022), Pile (Gao et al., 2021), and C4 (Raffel et al., 2020), are fully based on Common Crawl. LLM developers often assume the Common Crawl corpus to be a "copy of the internet" (Baack, 2024). In fact, it is neither a copy of the

21. https://science.ubc.ca/news/chatgpt-has-read-almost-whole-internet-hasnt-solved-its-diversity-issues, accessed: November 14, 2025

22. https://www.forbes.com/sites/rashishrivastava/2024/07/24/the-internet-isnt-big-enough-to-train-ai-one-fix-fake-data/, accessed: November 14, 2025

23. https://www.heise.de/en/background/AI-training-with-synthetic-data-The-internet-is-reaching-its-peak-9799160.html, accessed: November 14, 2025

24. https://www.theguardian.com/technology/2025/jan/09/elon-musk-data-ai-training-artificial-intelligence, accessed: November 14, 2025

25. https://www.statista.com/statistics/262946/most-common-languages-on-the-internet/, accessed: November 14, 2025

26. https://www.statista.com/statistics/272365/age-distribution-of-internet-users-worldwide/, accessed: November 14, 2025

27. https://commoncrawl.org/ (accessed: July 30, 2025)

28. In fact, an interview study by Orr and Crawford (2024) revealed that large tech companies like Amazon rely on publicly available corpora since they are easier to access than company-internal data. Access to the latter is often hindered by bureaucratic processes and legal concerns, e.g., for privacy reasons.

internet nor a representative sample of it. The organization actually crawls only a small fraction of the internet. They only crawl pages with high amounts of direct and indirect incoming links from other domains (Baack, 2024). This crawling heuristic was designed with the intention to assure data quality, but effectively prevents inclusion of contents from "digitally marginalized communities" and yields a predominantly English corpus (Baack, 2024, p. 7). Moreover, problematic contents are not filtered or classified. Luccioni and Viviano (2021) analyzed the content of a random subset of Common Crawl and detected that between 4 and 6% of all websites contained hate speech, such as racial slurs and "racially-charged conspiracy" (results varied between different detection methods). The authors also identified a substantial amount of sexually explicit content.

As mentioned before, knowledge-enhanced language modeling is one way researchers and practitioners try to fill the knowledge gaps by using external knowledge sources. But what if these external sources suffer from the same gaps? The aim behind the study presented Kraft and Usbeck (2022) and Chapter 4, was to understand the types of content represented in popular KGs and the characterization of potential gaps. Our systematic review showed that resources like Wikidata are affected by similar coverage issues.

Blind spots in the data cause blind spots in the systems, which makes them unable to reproduce certain knowledge (correctly). In practice, this creates epistemic issues because LLMs are promoted as sources of knowledge and tools for knowledge generation without making knowledge gaps transparent. Peterson (2025) even asserts that "our access to the original diversity of human knowledge is increasingly mediated by a partial and increasingly narrow subset of views" (p. 3250), ultimately leading to a *knowledge collapse*. Advertising AI systems as *general-purpose* or *foundational* conveys a universality claim. We might argue that either (a) the claim must be openly questioned and the literacy of users furthered, or (b) the systems must get better coverage. So, what are the hindrances to either of these options? It should be said that (a) is being pursued by scholars working in AI Ethics, Critical AI Studies, Philosophy of Technology, and the like. However, many creators of AI technology in the industry and technical research labs profit from the "universality myth" and are likely reluctant to rewrite the story (Narayanan and Kapoor, 2024). But what then speaks against (b) improving coverage? From the Common Crawl example, we can deduce that certain technical aspects limit the reach of web scrapes. For example, heuristics are used to filter out weakly linked URLs to ensure content quality. However, this heuristic potentially also leads to omitting valuable yet weakly linked pages for other reasons—for instance, those representing more marginalized interests. So, the technical choices relating to algorithm design, decision rules, and metrics play a significant role. The question of why things are designed in specific (biased) ways refers directly to the question of who the designers are. Later Chapters will revisit both of these questions.

1.1.4 The Problem of Representativeness

In Section 1.1.3, while discussing the coverage problem, a connection to representativeness was already hinted at. To characterize this problem in particular, its main sources, manifestations, and consequences are briefly addressed in the

following paragraphs. Please note that all of these aspects are illustrated in more depth in Chapter 2.

A training dataset, its contents and their relations to each other, are the very things that are *modeled* by an AI model. Thus, the composition of the training data determines the possibilities and boundaries of the model (Orr and Crawford, 2024) and "they also constitute their bedrock of claims to truth and accuracy" (Orr and Crawford, 2024, p. 4956). However, several studies have demonstrated that the datasets used for the development of influential AI models consistently underrepresent contents from marginalized communities (Navigli et al., 2023; Dodge et al., 2021) and contain harmful stereotypes and misrepresentations (Birhane et al., 2021; Birhane et al., 2023). The reasons for dataset biases are, for instance, the sources and sampling strategies as exemplified in the Common Crawl case. But also the data selection and annotation practices introduce biases (Hovy and Prabhumoye, 2021; see Chapter 2 for an elaboration). Orr and Crawford (2024) point out that the ways in which AI datasets are represented, stored, and disseminated lead to a concealing or "collective forgetting" of the "messy processes" behind their creation (see also Bowker and Star, 2000). That is, the biases of datasets are not arbitrary, but instead a result of the decisions, intuitions, and personal biases of those involved in the construction of datasets.

These data biases are one of the main causes for *algorithmic bias* in *natural language processing* (NLP) systems. In the nineties, Friedman and Nissenbaum (1996) have defined biased computer systems as those that "systematically and unfairly [discriminate] against certain individuals or groups of individuals in favor of others" (p. 332). In NLP systems, such discriminatory patterns can occur in different forms: for example, a classification or prediction system may perform systematically worse for certain individuals or groups. One classic example is the study by Kiritchenko and Mohammad (2018), which analyzed more than 200 different sentiment classifiers and found that 75% were gender- and/or racially biased. For example, for statements where the subject had an African American-coded name, the systems tended to give higher sentiment intensity scores on negative sentiments, such as anger, fear, and sadness. For White names, the models assigned higher intensity scores for positive sentiments. Race bias can also be found in common hate speech detection datasets (Sap et al., 2019a). To this date, this issue is reflected in hate speech detectors (Albladi et al., 2025) and content moderation APIs provided by Microsoft, Google, Amazon, and OpenAI (Hartmann et al., 2025). Another type of discriminatory behavior found in NLP systems, particularly LMs, relates to the associations and kinds of portrayals found in open-ended text generations (Sheng et al., 2019; Dhamala et al., 2021). LMs have repeatedly been shown to reproduce stereotypical and derogatory narratives about certain individuals and groups (Abid et al., 2021). For example, statements about female subjects are usually more positive but attributed with stereotypical attributes, such as homemaking and caretaking (Kraft et al., 2022). Besides such misrepresenting portrayals, ChatGPT generally generates more homogeneous depictions of fictional African, Asian, and Hispanic American characters versus White American ones; and more homogeneous depictions of women versus men (Lee et al., 2024).

The harms caused by these errors are *representational harms* as they cause certain groups to be demeaned, misrepresented, or ignored (Blodgett et al., 2020).

In comparison, harms caused by biased models within decision-making systems for use in, e.g., hiring or insurance, fall into the category of *allocational harms*. The harms caused here manifest in unfairly distributed resources or opportunities. Gender and race have been the most researched protected attributes in the algorithmic bias literature. This choice is not arbitrary: for one, gender- and race-based discrimination are deeply ingrained in Western society.[29] Hence, a shared understanding of existing stereotypes and prejudices and their harms exists to some extent. Secondly, there is a grounding within the AI community itself. That is, engineers of influential AI systems have historically been disproportionately male, white, and North American (Forsythe, 1993; Adam, 2000). It has indeed been shown that the biases identified in common AI models also happen to favor these demographic groups.

In this dissertation, particular emphasis is put on a type of injustice that uniquely affects one's "capacity as a knower", namely *epistemic injustice*. The notion originated in the scholarship of feminist epistemology. It was coined by Fricker (2007), who distinguishes between two types of epistemic injustice: the first type is *testimonial injustice*, and it takes place when the hearer does not give a speaker their deserved credibility due to identity-related prejudice. This credibility deficit is harmful because it wrongs the speaker in their capacity as a knower, which is a part of their "capacity for reason." And given that rationality is essential to human value, Fricker (2007) considers testimonial injustice a path to undermining a person's humanity. AI systems that reproduce harmful stereotypes amplify existing prejudices that feed into testimonial injustice and, thus, contribute to the discrediting of certain groups of knowers (Kay et al., 2024).

The second type of epistemic injustice is *hermeneutical injustice*, which presents itself in systematic gaps in "collective interpretative resources," i.e., knowledge resources. These gaps affect individuals' ability to interpret and understand their social experiences. Hermeneutical injustice is a form of structural discrimination. Fricker (2007) famously exemplifies the potential harms of hermeneutical injustices with a story about a woman named Carmita Wood, an office employee at Cornell University, in the 1970s, who was unable to name the repeated sexual harassment from a male faculty member she was experiencing (Brownmiller, 1999). The stress of her experience had caused her to develop physical ailments, leave her job, and apply for unemployment insurance. However, the unemployment benefits were denied because there was no category suitable to name her true reason for resigning. This example shows that being unable to have knowledge relevant to one's social experience can be a serious obstacle and cause a measurable disadvantage. Note that Fricker's conceptualization of hermeneutical injustices has been critiqued by Mason (2011), who proposed a more nuanced revision: besides the type of "unknowing" described by Fricker, a second type of "unknowing" is identified, namely *epistemically and ethically blameworthy ignorance*. It describes the phenomenon that more powerful groups may willfully sustain knowledge gaps or distort non-dominant accounts to maintain the status quo in their favor. This ignorance prevents themselves from understanding certain aspects of their

29. Note that in Europe, "race" is not usually a term used for ethnic classification (Jaime and Kern, 2024). However, since the AI and AI Ethics research communities are mostly U.S.-centric, the use of this term is a norm in the respective literature and will also appear throughout this thesis.

own experience and allows them to evade ethical responsibility (see Chapter 3 for a more detailed discussion).

The notion of epistemic injustice has been fruitfully applied to research in AI and *machine learning* (ML). This is plausible due to the aforementioned systematic gaps and inaccuracies in AI-based knowledge technologies. Miragoli (2025) argues that AI has an inbuilt tendency "to treat as epistemically relevant information that is statistically dominant *because* it is statistically dominant" (p. 6). And this feature necessarily leads to "knowledge-gaps and interpretative lacunae, affecting the machine's ability to read, understand and respond to minoritarian information" (p. 7). For instance, most AI technologies are English-centric and content in other languages is represented and processed less accurately (Lai et al., 2023; Helm et al., 2024). As a result, speakers of marginalized languages experience obstacles when accessing certain information altogether. Kay et al. (2024) dubs this particular phenomenon *hermeneutical access injustice*. Lindemann (2024) has pointed out that LM-generated answers in online search engines or the use of chatbots as an alternative to classic search contribute to a *sealing of knowledges*. Specifically, "the complexity of the possible answer space, the plurality of potential answers to a search query, is increasingly sealed" (p. 5066) behind a "singular, authoritative" text paragraph. In this way, Lindemann argues, more dominant knowledges get disseminated and solidified while marginalized knowledges become harder to find. Mollema (2025) claims that this can lead to *generative hermeneutical erasure*, i.e., conceptual differences or different ways of "sense-making" get eradicated since similarly biased AI systems are becoming more omnipresent.

The notion of epistemic injustice is not only relevant for the research on the harms and injustices caused by generative AI systems. It is also relevant for an investigation of one of their supposed remedies, namely knowledge graphs. Due to the rising importance of external knowledge resources, particularly open-domain resources like Wikidata (Johnson et al., 2024), this thesis extends the focus accordingly. In the context of AI development, Wikidata is commonly framed as authoritative and reliable.[30] And in contrast to other web sources, it may indeed be *more* reliable (Longpre et al., 2024, for an analysis of the value of the related knowledge base Wikipedia as AI training data). Nevertheless, whether or not the level of trust and authority attributed to these sources is truly justified must be examined. A critical analysis thereof is given by Kraft and Soulier (2024), presented in Chapter 5.

1.2 Research Questions

As described in the previous Section, LM- and KG-based AI systems shape our knowledge ecosystems. They influence how we produce knowledge and determine which knowledge is more or less accessible. Accordingly, assessing their goodness is essential. To this end, this dissertation presents a multifaceted evaluation of respective knowledge technologies. A core interest is understanding the extent to which these contents and processes are epistemically and ethically good. Special

30. https://blog.wikimedia.de/en/2024/09/17/wikidata-and-artificial-intelligence-simplified-access-to-open-data-for-open-source-projects/, accessed: November 14, 2025

attention is paid to correctness, coverage, and representativeness, which are three key desiderata for epistemic and ethical goodness:

D1. Correctness: AI-based knowledge technology should accurately encode and reproduce knowledge content.

D2. Coverage: AI-based knowledge technology should encode and reproduce knowledge with adequate coverage.

D3. Representativeness: AI-based knowledge technology should *not* systematically or unfairly misrepresent or underrepresent the knowledge of, or about, marginalized communities.

These desiderata guide this thesis's choice of technologies and analytical approaches. D1 underpins the detailed examination of knowledge-enhanced language modeling as a potential remedy to issues of incorrectness and bias (related to Chapter 5), and benchmarks as measures of correctness (related to Chapter 6). D2 guides the critical engagement with knowledge databases, specifically KGs (related to Chapter 4), on the one hand, and benchmarking datasets that measure and direct the coverage of LMs, on the other (related to Chapter 6). However, the most attention is paid to D3, the problem of representativeness (related to Chapters 4, 5, and 6). While bias has been a well-researched limitation of AI systems, particularly LLMs, it is less investigated in the domain of KGs and knowledge technology, more generally. This thesis aims to fill this gap. It aims to shed light on the goodness of the knowledge bases and measures treated as "sources of truth" to improve correctness and coverage. To this end, it examines the *knowledge processes* that underlie the knowledge bases and measures. And it considers the final products, i.e., the *knowledge content* that is ultimately represented and conveyed. The distinction between knowledge processes or knowledge as *knowledge production*, knowledge as *knowing*, and knowledge as *content* is inspired by the work of Longino (2002) and is explained in more detail in Chapter 3. The research questions addressed are as follows:

RQ1. What types of social bias are embedded in knowledge graphs? How are they measured? And what do we know about their causes?

RQ2. Can knowledge enhancement make language models less biased with regards to their knowledge content? Can it help to make language models more objective?

RQ3. How are the measures created that are used to determine a language model's accuracy in reproducing knowledge? How is the quality and representativeness of these measures?

1.3 Situating the Thesis

To situate the thesis, Section 1.3.1 provides an overview of related research. Note that more detailed definitions of algorithmic bias in LMs and KGs are presented in Chapter 2. Section 1.3.2 situates and scopes the thesis thematically and methodologically. Specifically, it addresses its interdisciplinary lens, as well as the technologies, theories, and methods of choice.

1.3.1 Related Work

Bias in Language Models

LMs have been shown to reproduce harmful gender stereotypes (Kirk et al., 2021; Kotek et al., 2023; Zhao et al., 2019), biases against queer people (Felkner et al., 2023), disabled people (Venkit et al., 2022; Panda et al., 2025), certain religions (e.g., Islam and Judaism; Abid et al., 2021; Plaza-del-Arco et al., 2024), as well as races and ethnicities (Yang et al., 2024c; Hofmann et al., 2024). Stereotypical associations can be measured by comparing the vector distances for particular seed-attribute pairs. Seed words would be identity-related mentions, like female versus male pronouns, and attributes are, for instance, names of different occupations (May et al., 2019; Guo and Caliskan, 2021). Another way to measure stereotypical associations is by comparing the predicted token probabilities for certain attributes, given a seed word (Nangia et al., 2020; Nadeem et al., 2021). Such metrics measure bias directly on an embedding level (*intrinsic measures*) and have been shown to not reliably indicate a model's downstream task behavior, which is more directly user-facing (*extrinsic measures*; Goldfarb-Tarrant et al., 2021). Downstream task bias measures determine disparate performance for different seeds. Examples are WinoGrande, which measures performance differences in a pronoun resolution task (Sakaguchi et al., 2021) and the aforementioned analysis of Kiritchenko and Mohammad (2018), which compared the sentiment classification accuracy for differently "racialized" names. Since this thesis includes a deeper investigation into QA benchmarks, it shall be noted that two dedicated bias benchmarks exist, which utilize this format. These are UNQOVER (Li et al., 2020) and the Bias Benchmark for QA (BBQ; Parrish et al., 2022). Regarding BBQ, Lum et al. (2025) note that, while the task format is strictly speaking extrinsic, the actual questions are not realistic. For example the question "After the first day of middle school math, a girl and a boy stayed after class to talk to the teacher. Who is bad at math?" is not quite to be expected in an actual user scenario. UNQOVER comprises similarly abstract task items. Researchers in the field have been advocating for measures that are more closely emulating actual use (Lum et al., 2025; Delobelle et al., 2022).

LMs have also been found to be biased in terms of systematic differences regarding the accuracy with which they represent and process different languages. Even though more efforts are being invested into multilingual LM development (Scao et al., 2022; Qin et al., 2025), most if not all mainstream models are still English-first and struggle more in the handling of less dominant languages (Helm et al., 2024; Bella et al., 2024). Reasons for this are, for one, methodological factors. For instance, the engineers of such models are not always speakers of the respective languages and, thus, not perceptive to potential errors (e.g., when removing "special characters" that are important to the language-specific script; Bella et al., 2024). And, secondly, the data availability differs across languages (Qin et al., 2025).

Knowledge in Language Models

To analyze whether pretrained LMs qualify as knowledge bases, Petroni et al. (2019) tested BERT (Devlin et al., 2019) and a few other LMs via cloze-style ("fill-in-the-blank") statements from four different datasets based on Wikipedia, Wikidata, and a commonsense KG (Speer et al., 2017). Petroni et al. (2019) found BERT's

performance (without finetuning) to be comparable to task-specific NLP methods. The T-REx probe (based on a dataset by ElSahar et al., 2018), which was part of their study, was also used in Kraft and Soulier (2024). Another way of measuring how well LMs encode knowledge is by using a QA format (Rogers et al., 2023). Sun et al. (2024) found that the popularity of some content or its presence in the training data influence the model's ability to reproduce it in a QA setup. The authors created a QA benchmark based on the open-domain KG DBpedia (Auer et al., 2007a), from which they extracted entities representing movies, books, and academic knowledge. They grouped the benchmarking examples by entity popularity–as measured by graph density (number of facts about an entity) and traffic (views and votes of entry on the web)–and found that 16 state-of-the-art LLMs perform worse on less popular entities, irrespective of model size. These findings are particularly problematic as entity popularity follows a power-law distribution, meaning that most entities are unpopular (Sun et al., 2024). Kandpal et al. (2023) also analyzed the relationship between LLM parameter count and QA accuracy on "popular" versus "unpopular" knowledge. When operationalizing popularity by the frequency with which relevant documents can be found in popular pre-training corpora, i.e., how often the model was exposed to the content during training, model size and accuracy also on long-tail knowledge do correlate. However, models would need to scale to a quadrillion parameters in order to reach comparable performance on the popular and unpopular contents (Kandpal et al., 2023).

Bias in Knowledge Graphs

While bias in LMs is a very active field of research, the issue has received less attention in the context of KGs. Chapter 4 presents the first systematic review of research investigating biases in the KG lifecycle. It provides a comprehensive overview of the field at the time of its publication in 2022. Since then, new contributions have been made: Melis et al. (2024) conducted an in-depth analysis of queerness in Wikidata, which sheds light on how and why it tends to misrepresent marginalized gender identities. One root cause are the editors' ideological beliefs. A newer factor is the rising use of bots that are used in an effort to quickly fill-in missing information in the graph. AI-based bots infer gender from names, which leads to systematic misgendering, i.a., due to cultural and linguistic differences with respect to naming conventions (Melis et al., 2024). In an analysis of KGs and other machine-readable (*linked open data*; LOD) resources, namely Wikidata, The Getty Art & Architecture Thesaurus, Princeton WordNet, and Open Dutch WordNet, Nesterov et al. (2024) found a high prevalence of derogatory labels and descriptions, including racist and homophobe slurs. Finally, a recent study by Das et al. (2025) demonstrated that gender-occupation and age-occupation biases in Wikidata differ between "Global North" and "Global South" geographies.

Knowledge-Enhanced Language Models and Bias

As mentioned earlier, injecting explicit knowledge into LMs is currently considered the most promising corrective for dealing with factual inaccuracies and long-tail knowledge (Sun et al., 2024; Li et al., 2024). However, RAG, for example, is pre-dominantly implemented via vector-based matching between the model

input and external documents (Fan et al., 2024). Thus, it is similarly biased towards content that appears more often in the external knowledge base (Kandpal et al., 2023). And, in fact, Wikipedia appears to be a widely used knowledge base for open-domain scenarios (Fan et al., 2024). As discussed in Kraft and Soulier (2024), Chapter 5, Wikipedia is in itself highly biased (e.g., over-represents Western and male personalities). While RAG is mostly used in combination with textual databases, there is another class of enhancement techniques, that specifically utilize structured data from KGs. Kumar et al. (2025), for instance, have proposed KG-augmented LLM training particularly for the purpose of detecting and mitigating biases. Their assumption is that KGs add contextual information that "help[s] counter the biases inherent in unstructured text" found in the LLM training data (p. 608). The authors finetune GPT-4 using conventional bias mitigation approaches, counterfactual data augmentation and adversarial training, and combined that with KG-augmented finetuning. The KG was vectorized as a graph neural network (GNN) and integrated into the LLM via the attention mechanism. The results indicate slight bias reductions on different classification tasks, as measured by demographic parity and equal opportunity. However, Kumar et al. (2025) do not to mention exactly which KG they used and how much of a bias improvement is achieved through the data augmentation and adversarial training, without KG augmentation.

Debiasing

Especially with the immense data requirements of modern LLMs, balancing training datasets in terms of representational aspects or removing undesirable misrepresenting content is a non-trivial endeavor. An obvious solution would be to use automated methods to identify different instance types in order to down- or upsample. However, as Navigli et al. (2023) pointed out, classifiers for such tasks are themselves not free of bias (Albladi et al., 2025). Some researchers have attempted to "debias" LM outputs through technical fixes, for example through vector-based adjustments (Bolukbasi et al., 2016) or automatically altered prompts (Sheng et al., 2020). There have also been works aiming to automatically mitigate biases in KG embeddings (Arduini et al., 2020; Chuang et al., 2025). However, automated bias mitigation techniques can only remove whatever bias aspects are addressed in the concrete formalizations, i.e., the bias metrics with which the bias is detected. Bias metrics are necessarily reductive. So, by labeling a treated model as "debiased", such approaches mostly just conceal the full extent of the issue (Gonen and Goldberg, 2019; Kraft, 2021).

Birhane et al. (2022c) strongly promote a contextualized and concrete approach to AI Ethics research. They argue that analysis of bias and other risks are not often enough embedded in an explicit scrutiny of power-related and oppressive factors. Local and contextualized AI evaluations are important to ensure that impacts are assessed with regards to dimensions or struggles that are truly relevant. For instance, caste-related discrimination is central to an Indian but not so much to a Western context. Moreover, although gender bias is a relevant dimension across these regions, it manifests in different stereotypes within Indian society (Bhatt et al., 2022).

Epistemic Values and Responsible AI

Olteanu et al. (2025) argue that while AI research has paid much attention to scientific rigor, it has narrowly focused on *methodological rigor*, i.e., the correct application of statistical and computational methods, or the generalizability of systems as measured by large-scale quantitative benchmarks. The authors promote a broader and more contextualized conceptualization broken down into six facets, namely (1) *epistemic*, (2) *normative*, (3) *conceptual*, (4) *methodological*, (5) *reporting*, and (6) *interpretative rigor*. (1) In practice, models and datasets are often utilized because they are easily available or commonly used without consideration of their underlying assumptions (Koch et al., 2021b).[31] *Epistemic rigor* requires making explicit the body of knowledge or assumptions that ground the work and justify certain choices. (2) As discussed in Chapter 3, values and beliefs play a role in deciding *what* is researched and *how*. So, *normative rigor* demands communicating the norms, standards, values, or beliefs underpinning one's work and alludes to using positionality and ethics statements in research publications. (3) *Conceptual rigor* commands clarity regarding terms and conceptualizations behind them. To further explain this facet, Olteanu et al. (2025) refer to the use of term "hallucination" in the AI community. As mentioned before, it is often used as a synonym for "factual error" (as well as other types of erroneous model outputs), even though the sensory experience observed in humans has little to do with the errors observed in LMs (Maleki et al., 2024). Respective use of this term without proper clarification risks spreading misconceptions about AI. (5) *Reporting rigor* refers to transparent and detailed communication of research findings. For example, aggregate measures and confirmatory research (i.e., only publishing positive results) should be avoided. In Kraft and Soulier (2024) the T-REx knowledge probe (Petroni et al., 2019) was disaggregated into test cases relating to female versus male entities. This revealed that the tested models exhibit systematically worse performance on the female subset. Aggregate benchmark scores conceal such differences. (6) Finally, *interpretative rigor* is concerned with how findings are interpreted. In Kraft et al. (2025), many popular QA and RC benchmarks were found to be biased towards male and Western entities, and all but one of the benchmarks were English. Is it fair to assume that these benchmarks measure QA and RC performance in general? Or is it perhaps more accurate to interpret them as measures of QA and RC performance within a particular cultural context? Normative rigor overlaps with other types of rigor, as it requires clarity regarding epistemic and normative assumptions, conceptual, as well as methodological choices, and it requires transparent documentation thereof. Overall, Olteanu et al. (2025) promote the idea that scientific rigor is not only a necessary condition of ethical research, but also of epistemically sound research. Reflexivity and contextualized analysis of latent conceptualizations, individual decisions, and interpretations are required to identify flaws in our scientific practice that may lead to misconceptions or empirically unacceptable claims (Longino, 2002).

31. One example is models that classify criminals based on portraits. These are built on the assumption that criminality is a visible trait. This idea is rooted in physiognomy, which has been scientifically debunked and criticized for its ethically problematic implications (y Arcas et al., 2023).

Even research that targets the development of more responsible AI systems tends to follow implicit assumptions and values without scrutiny. Hancox-Li and Kumar (2021), for example, studied the epistemic values implied in feature importance methods, a set of approaches that aim to make models more transparent and improve accountability by identifying which input features are most influential to the prediction. The authors identify *universality* as one of the epistemic values at play here: they argue that feature importance methods are designed to fulfill universally desirable properties or to succeed on established explainability benchmarks. Drawing from feminist epistemology, Hancox-Li and Kumar (2021) critique that desirability criteria are specified without consideration of context or individual positionality ("Desirable to whom?"). They articulate a number of suggestions for improvement, such as incorporating marginalized perspectives, evaluating the appropriateness of methods contextually, and pursuing *seamful* design. *Seamfully* designed feature importance methods present explanations not as "singular and authoritative", but rather as open to multiple interpretations and with a level of uncertainty.

1.3.2 Thematic and Methodological Scope

This dissertation is an interdisciplinary project that is grounded in computer science and philosophy. While the theoretical backgrounds are presented in Chapters 2 and 3, the thesis shall for now be situated through four characterizing dualities:

The Technical versus the Social Lens

The thesis engages deeply with the technical AI literature, particularly in the areas of natural language processing (NLP) and the semantic web. The focus lies on the technicalities of language modeling, structured knowledge representations, factual errors, statistical biases, etc. Whereas this thesis also engages with philosophical literature in the areas of social and feminist epistemology, and philosophy of science. Intersectional fields of research, such as *science and technology studies (STS)* and AI ethics, also provide critical theoretical foundations. It is important to note here that in STS (and inspired by that also in AI ethics), *the technical* and *the social* are usually seen as tightly interwoven components of a single system, namely, a *socio-technical system.* Social actors and institutions create technology, and it, in turn, affects them and their interactions. Selbst et al. (2019) argue that to achieve the social goal of "fairer" machine learning systems, technical practitioners and researchers must engage with their technical solutions' social embedding and impact. Following this line of thought, the three articles comprising this thesis are interpreted through the socio-technical lens.

Statistical versus Symbolic Representations of Knowledge

As objects of investigation, this thesis focuses on two different technical approaches to representing knowledge: Firstly, LMs, as they have become one of the most widely used types of AI methods and are at the core of many modern knowledge technologies. Secondly, KGs have been important in connecting and making accessible knowledge in the digital sphere for decades. These approaches

descend from two different paradigms in AI, the *statistical* versus the *symbolic*, which introduces another duality. Historically, the respective communities and ideas have largely opposed each other. However, efforts to marry LMs and KGs have been increasing in recent years, e.g., in the form of *knowledge-enhanced language modeling* or by leveraging AI to populate KGs. A common denominator between both is the goal to capture and represent real-world knowledge comprehensively and accurately–such that it can be used to retrieve or even generate new knowledge. Chapter 2 introduces the computer science background regarding the most relevant technical methods and artifacts.

Traditional versus Feminist Conceptualizations of Knowledge

There is also duality within the philosophical debates this thesis draws from, first and foremost, the duality between traditional epistemology and social and feminist epistemology. The first views knowledge as a product of cognition only, independent of the social, and the second views knowledge as a product of social or rational-social processes, as situated and embodied, and as a site of potential injustice. STS scholar Alison Adam has pointed out how the traditional conceptions of knowledge are deeply embedded in the development of AI and introduced the feminist-epistemological view to critique this status quo (Adam, 1998). Note that this thesis frequently refers to *knowledges* in plural form from here onward. This is to signify the assumption that knowledges are manifold, partial, and situated (Haraway, 2016). With this, this thesis also situates itself in opposition to the disembodied, unitary accounts of knowledge embedded in AI. More details in this regard are provided in Chapter 3, which offers an overview of the theoretical works that have been most fundamental to this dissertation project.

Quantitative versus Qualitative Methodology

Due to the interdisciplinary nature of this work, the presented papers employ various quantitative and qualitative methods. On the quantitative side, statistical analyses of training and evaluation datasets are conducted. AI model outputs are measured and compared via bias metrics (Chapters 5 and 6). On the qualitative side, there is a systematic literature review (Chapter 4), a philosophical analysis (Chapter 5), and a content analysis (Chapter 6). Mostly, both types of methods are combined to ground the more technical observations in the epistemological and ethical discourse, i.e., to bridge the technical and the social.

1.4 List of Publications

Below, the papers comprising the heart of this thesis are listed in order of their publication date, followed by remarks on authorship. At the time of submitting this thesis, two of the works have been presented at international conferences and the third one was accepted for presentation at an international conference.

I shall also note that I kindly received various opportunities to talk about these works, beyond these conferences.[32, 33, 34]

Following the publication list, remarks are added about changes made to integrate the original publications into this dissertation more seamlessly. Finally, other works that were published during the same time frame are also listed. They do not directly contribute to answering the RQs in this thesis, but add additional context.

1.4.1 Peer-Reviewed Articles Presented in this Thesis

Angelie Kraft and Ricardo Usbeck. 2022. The Lifecycle of "Facts": A Survey of Social Bias in Knowledge Graphs. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, Volume 1: Long Papers (AACL-IJCNLP 2022)*, 639–652. Online: Association for Computational Linguistics.

Angelie Kraft and Eloïse Soulier. 2024. Knowledge-Enhanced Language Models Are Not Bias-Proof: Situated Knowledge and Epistemic Injustice in AI. *In Proceedings of the 2024 Association for Computing Machinery Conference on Fairness, Accountability, and Transparency (FAccT 2024)*, 1433–1445. Rio de Janeiro, Brazil: Association for Computing Machinery.

*Angelie Kraft, Judith Simon, and Sonja Schimmler. 2025. Social Bias in Popular Question-Answering Benchmarks. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2025)*, 1421–1438. Mumbai, India and Online: The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

**Nominated for Best Paper Award at IJCNLP-AACL 2025.*

Comments on Degree of Authorship

I am the sole first author on all of the included articles. The detailed contributions of each of the co-authors are as follows:

In Kraft and Usbeck (2022), I conceptualized the idea and research questions. I conducted the systematic literature analysis and authored the paper. Ricardo Usbeck provided theoretical and methodological guidance and proofread the paper.

In Kraft and Soulier (2024), Eloïse Soulier and I developed the idea for the paper together, based on my previous research on KGs and concerns regarding claims around bias and neutrality in KG-enhanced language modeling. She leveraged her expertise in feminist epistemology to provide a theoretical framing to this critique,

32. Interview with AIhub about Kraft and Usbeck (2022): https://aihub.org/2022/11/16/the-lifecycle-of-facts-a-survey-of-social-bias-in-knowledge-graphs-interview-with-angelie-kraft/, accessed: November 13, 2025

33. Talk at re:publica about Kraft and Soulier (2024): https://re-publica.com/en/node/3842, accessed: November 13, 2025

34. Invited panel at COMPTEXT 2025 where I presented Kraft et al. (2025): https://www.cais-research.de/news/marginalisierte-stimmen-und-perspektiven-in-der-computergestuetzten-textanalyse/, accessed: November 13, 2025

and she authored the respective parts of the paper (Sections 5.2.1 and Section 5.4). We co-authored the Introduction and Conclusion (Sections 5.1 and 5.5). I authored Sections 5.2.2, 5.3, and the Appendices. I implemented and conducted the quantitative analysis of LM bias. We collaborated closely in proofreading and editing each other's Sections to ensure coherence, content-wise and linguistically.

In Kraft et al. (2025), I conceptualized the idea, research questions, and operationalization. I conducted the quantitative and qualitative analyses (including the annotation data collection) and wrote the paper. Judith Simon provided guidance with the theoretical embedding of the paper, discussed results, and helped refine the paper. Sonja Schimmler provided guidance regarding the methodological and technical decisions, discussed results, and proofread the paper.

Comments on Edits

A few minor editorial changes were made to the papers in this thesis to improve consistency and readability. Firstly, all articles were reformatted into a coherent design throughout this dissertation. Secondly, the reference style was changed and made coherent since the original papers were previously published at different venues with varying citation guidelines. To avoid redundancy, all references are listed in a combined bibliography, at the end of this book. Thirdly, the appendices were renamed and given consistent titles so they are easier to identify and distinguish from the main content in the table of contents.

1.4.2 Other Publications

Angelie Kraft, Hans-Peter Zorn, Pascal Fecht, Judith Simon, Chris Biemann, and Ricardo Usbeck. 2022. Measuring Gender Bias in German Language Generation. In *Proceedings of 52. Jahrestagung der Gesellschaft für Informatik, INFORMATIK 2022, Informatik in den Naturwissenschaften*, 1257–1274. INFORMATIK 2022. Hamburg, Germany: Gesellschaft für Informatik.

Angelie Kraft and Ricardo Usbeck. 2022. The Ethical Risks of Analyzing Crisis Events on Social Media with Machine Learning. In *Proceedings of the International Workshop on Data-driven Resilience Research 2022 co-located with Data Week Leipzig 2022 (DATAWEEK 2022)*. D2R2 2022. Leipzig, Germany: CEUR-WS.org.

Angelie Kraft. 2023. Triggering Models–Messung und Mitigation sexistischer Vorurteile in deutschen Sprachmodellen. *Frauen machen Informatik* 47:39–44. Gesellschaft für Informatik.

Angelie Kraft. 2024. Unpacking Large Language Models: Grundlagen, Perspektiven und Herausforderungen. *Frauen machen Informatik* 48:10–16. Gesellschaft für Informatik.

1.5 Thesis Outline

The following two Chapters will provide insights into the technical and theoretical underpinnings of the research presented in this cumulative dissertation. Chapter 2, firstly, defines and discusses the most essential computer science concepts and methods that are subjects of investigation. It also establishes the core principles of

algorithmic bias research with a particular focus on NLP applications. Chapter 3 summarizes the philosophical foundations. It highlights the epistemological works most influential to this research and scholarship that have previously introduced the feminist-epistemological lens to AI. This is followed by the individual papers:

Chapter 4 presents a systematic review of research on social biases in KGs throughout their lifecycle, from manual and automated creation of contents and ontologies, to large-scale graphs and graph embeddings (Kraft and Usbeck, 2022). This work was published as part of the proceedings of the main track of the 2022 Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (AACL-IJCNLP 2022).

Chapter 5 presents a research article published in the Proceedings of the 2024 Association for Computing Machinery (ACM) Conference on Fairness, Accountability, and Transparency (FAccT 2024) (Kraft and Soulier, 2024). It critically examines the *knowledge processes* behind the sources of truth utilized in knowledge-enhanced LMs and the assumptions around objectivity and value-neutrality that guide the development, use, and marketing of knowledge-enhanced LMs as *knowledge technologies*.

Chapter 6 presents a research project in which the 30 most popular QA and RC benchmarks are investigated regarding the underlying data collection and annotation practices, focusing on coverage and representativeness (Kraft et al., 2025). The article was published in the main track proceedings of the 2025 International Joint Conference on Natural Language Processing and Asian Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2025). It was nominated for the Best Paper Award.

In Chapter 7, the findings of all papers are consolidated and discussed in the face of the overarching research questions and desiderata introduced earlier. And finally, the dissertation concludes with Chapter 8, distilling and emphasizing the main insights, limitations, and avenues for future work.

2

Computer-Scientific Background

Contents

2.1 Introduction

According to Russell and Norvig (2020), to achieve human-like intelligence in a computer system, it needs to have a number of different capabilities. Firstly, it needs to be capable of natural language processing, i.e., the processing and generation of human language. Secondly, it needs to be able to store and meaningfully represent knowledge. Thirdly, it needs to be able to reason automatically to answer questions and draw conclusions. And finally, it needs to be able to adapt to new situations, to detect and generalize patterns through machine learning. The four conditions mentioned by Russel and Norvig are helpful to cluster some of the ongoing efforts of researchers and developers in the field. This thesis critically examines, in particular, natural language and knowledge representation. Note, however, that *language*, *knowledge*, and *learning* are assumed to be interconnected: For an AI system to generate meaningful language, it must accurately incorporate some representation of knowledge. The mechanism by which the rules and patterns of language, as well as the contents of knowledge are captured, is commonly referred to as learning (Russell and Norvig, 2020). Note that this reflects the understanding, based on which AI systems are commonly designed. A more critical philosophical reflection of knowledge is presented in Chapter 3.

In the following, the most important approaches to modeling these capabilities are presented. As modern language models are usually based on some variant of an artificial neural network, Section 2.2 presents essential background knowledge related to this concept, and introduces the idea of learning in machines. This is followed by an overview of significant milestones in the historical development of language models, in Section 2.3. Another technological concept investigated in this thesis is that of the knowledge graph, which is introduced in Section 2.4. All of the works presented in this dissertation provide some form of critical analysis of these technological approaches with a focus on algorithmic bias. Thus, Section 2.5 gives an in-depth account of definitions of algorithmic bias, its harms, measurement, and potential remedies.

2.2 Artificial Neural Networks

Artificial neural networks (ANNs) or *multilayer perceptions* (MLPs) are an important foundation for many of the approaches discussed throughout this Chapter as well as this thesis in general. Not only are modern language models implemented as neural networks, but also knowledge graphs are often embedded via respective approaches. Thus, we will firstly have a closer look at the definition of a neural network, how it is optimized, and what types of learning paradigms exist.

2.2.1 Definition

A neural network consists of multiple, simple functions which are combined such as to allow the modeling of complex, non-linear relationships within some data. The most basic "computational unit" here is commonly referred to as a *neuron* (Jurafsky and Martin, 2025). The core of such a neuron is given by the following weighted sum:

$$z = b + \Sigma_i w_i x_i, \tag{2.1}$$

where b is a *bias term*, $x_1, x_2, ..., x_n$ are the inputs, and $w_1, w_2, ..., w_n$ are associated weights. This function is usually represented in vector notation, as the dot product:

$$z = \mathbf{w} \cdot \mathbf{x} + b, \tag{2.2}$$

where z is a real-valued number, $\mathbf{w}$ is a weight vector, $\mathbf{x}$ is an input vector, and b is a scalar bias (Jurafsky and Martin, 2025). In order to allow the stacking of many functions to model complex relationships, non-linear *activation functions* are applied (Goodfellow et al., 2016). Given an activation function σ, the output of a neuron is thus:

$$\hat{y} = \sigma(z). \tag{2.3}$$

Commonly used activation functions are, for instance, the sigmoid function, the tanh function, or rectified linear units (ReLU) (Agarap, 2018) (Jurafsky and Martin, 2025). An activation function has to be differentiable, since neural learning algorithms rely on gradients. Other than that, different activation functions come with different up- and downsides. For example, sigmoid and tanh nicely squash values within certain value ranges. However, their derivatives can get very close to 0, which causes instabilities in neural network training (see Section 2.2.2).

Neurons are commonly arranged in layers. In a *feedforward* network, the outputs of one layer are fed as input to the next and the neurons between layers are fully connected (Goodfellow et al., 2016). The first layer is called the *input layer*, the final layer is called the *output layer*, and all of the layers in-between are referred to as *hidden layers*. The number of neurons per layer, as well as the amount of layers in total are design decisions. To allow the modeling of complex functions it is beneficial to use deep neural networks with a large amount of layers and weights.

A neural network (with at least one hidden layer and a nonlinear activation function) is a *universal function approximator* (Hornik et al., 1989). This means it can model any function f^* that maps an input x to an output y: $y = f^*(x)$. The exact mapping defined by the neural network is $y = f(x; \theta)$, with θ representing a set of learnable parameters, i.e., the weights and biases, that are optimized to achieve the closest possible approximation result (Goodfellow et al., 2016).

2.2.2 Optimization

In order to approximate $f^*(x)$ through $f(x; \theta)$, the parameters θ of the network need to be adjusted. This is done with the help of *training data* and a *learning algorithm* (Goodfellow et al., 2016). The general idea is to present examples from the training data which consist of an exemplary input and an associated known output value–also called the *target label*. These labels are commonly human- or machine-annotated. Other synonyms are "ground truth" and "gold standard". However, since such labels are "neither objective nor necessarily representative of reality" (Paullada et al., 2021, p. 2), these synonyms are potentially misleading and the term target label is, hence, preferred.

The optimization processes consists in adjusting the weights and biases of a network such that its output, i.e., the *predicted label*, is as close as possible to the target label. The distance between the predicted label and the target label is determined via a *loss function* (also called *cost function*). The exact function is a matter of design choice depending on the data and task at hand. Common examples are the *mean squared error* (MSE) for regression tasks or the *cross-entropy* metric for classification tasks (Mao et al., 2023). Based on the calculated loss, the parameter weights are to be adjusted. In practice, however, it is intractable to exactly calculate the correct weights. Instead, some form of *gradient descent* (explained below) is used to find approximations. Due to the non-linearity within neural networks, most loss functions are nonconvex. That means, mapping out all possible parameter values spatially results in a hilly landscape of possible error values with several local minima, making convergence toward a global minimum error a challenge (Goodfellow et al., 2016).

The algorithm used to compute the gradients for the weight updates in a neural network is the *backpropagation* algorithm. During backpropagation, the gradient of the loss function with regards to the model parameters is computed and propagated backwards through each layer of the network, applying the chain rule. The chain rule states that for a real number x, functions f and g (that map from real numbers to real numbers), and given $y = g(x)$ and $z = f(g(x)) = f(y)$, the gradient is (Goodfellow et al., 2016):

$$\frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx}. \tag{2.4}$$

This rule generalizes to vectors and tensors, allowing for the error gradient to be back-propagated through a chain of multiplications. This process is applied iteratively to shift the weights towards some optimum. In gradient descent, these updates are based on an average of the whole training dataset. In *stochastic gradient descent*, the updates are made on the basis of individual or small batches of samples at each iteration (Amari, 1993; Masters and Luschi, 2018). This causes more erratic changes but requires less memory, which is favorable given the vast amounts of data modern LMs are trained on. The process terminates at a pre-defined stopping criterion (e.g., a lack of improvement of the loss). Due to the non-convexity issue, there is no guarantee that this process will lead to a global optimum (Goodfellow et al., 2016). However, different strategies exist in the hopes to increase the likelihood (e.g., by introducing *momentum* to the gradient updates (Rumelhart et al., 1986)).

2.2.3 Learning Paradigms

Artificial neural networks are trained via different learning algorithms depending on data and task at hand, which may be coarsely categorized into *supervised* and *unsupervised* methods. Learning from labels as described in Section 2.2.2 is referred to as supervised learning. Its purpose is to learn how a predicted value depends on an input value. That is, a supervised learning algorithm models $p(y|x)$ for a random vector x and an associated value or vector y (Goodfellow et al., 2016). Examples for such a task setting are classification and regression. Unsupervised methods do not utilize such labels or associated value pairs and

instead aim at modeling the structure of a dataset, i.e., probability distribution $p(x)$ for a random vector x (Goodfellow et al., 2016). Examples are clustering or factor analysis, where the goal is to find features, commonalities, and differences within the dataset in a bottom-up way. Formally, supervised and unsupervised methods can be expressed as the respective other (Goodfellow et al., 2016). Hence, the gradient-based optimization process through backpropagation as described earlier is analogous for both types of methods.

Another type of learning paradigm is *reinforcement learning*. Its logic is quite different from supervised and unsupervised methods as it is conceptualized on the assumption of a learning *agent* in an *environment* in which it can perform certain *actions* (Sutton and Barto, 2018). The environment has different *states* which the agent perceives and manipulates through its actions. The goal of the learning process is to optimize a so-called *policy* which determines the agent's action at any given state. Each action determined by the policy is evaluated and rewarded (or punished) through a *reward signal* and the agents only goal is to maximize its reward (Sutton and Barto, 2018). A typical domain in which reinforcement learning has received much attention in the past is robotics (Wang et al., 2020). However, in recent years, it has also been playing a larger role in NLP and language modeling, in particular (Ouyang et al., 2022).

2.3 Language Modeling

This section describes the foundational definitions and principles of language modeling. It gives an historic overview of the development of language modeling approaches from n-grams to vectorized, neural, and contextualized representations. This section also describes the transformer architecture, a special neural network architecture that was introduced in the context of machine translation and is nowadays used in a plethora of domains and modalities. Finally, large language models are characterized, as well as related approaches around human preference alignment and factual (in-)accuracy. The aim of this section is to provide an understanding of the strengths and weaknesses of modern language models.

2.3.1 Definition

A language model is defined by the task which it solves. This task is to predict the next word w_{t+1} out of a predefined vocabulary V, given a sequence of preceding words. So formally, a language model models the probability $P(w_{t+1}|w_t, ..., w_1)$. In doing so, it assigns probabilities to word sequences. Let $w_1, ..., w_T$ be a word sequence, then its probability is defined as (Bengio et al., 2003; Jurafsky and Martin, 2025):

$$P(w_1, ..., w_T) = \prod_{t=1}^{T} P(w_t|w_{t-1}, ..., w_1). \tag{2.5}$$

2.3.2 Vector-Based Modeling

Computing the probability in Equation 2.5 for long sequences of text would be prohibitively expensive. So, the simplest way to approximate $P(w_{t+1}|w_t, ..., w_1)$ is through an *n-gram model*. Justified by the Markov assumption, i.e., the assumption that a future state depends only on the current state and not past states, instead of looking at all preceding words in a sequence, the n-gram model only considers words up to a predefined horizon of $n-1$ preceding words: $P(w_{t+1}|w_t, ..., w_{t-n})$ (Jurafsky and Martin, 2025). The parameters of the n-gram model can be estimated through maximum likelihood estimation (MLE), by computing the ratio of the observed frequency of a sequence to the observed frequency of a prefix. For example, for a bi-gram model (n-gram with horizon $n = 2$), the MLE function is defined as

$$P(w_{t+1}|w_t) = \frac{C(w_t w_{t+1})}{C(w_t)}, \tag{2.6}$$

where C denotes a frequency count (Jurafsky and Martin, 2025). In practice, this approach is still costly to train since the number of parameters that need fitting grows exponentially with n (Bengio et al., 2003). This *curse of dimensionality* yields enormous data requirements, restricting the practicability of modeling long range dependencies. Moreover, data sparsity can become a hindrance to the effectiveness of n-gram modeling (Jurafsky and Martin, 2025). That is, most word combinations appear rarely. So, assigning each n-gram a unique representation is inefficient. Instead, it would make sense to leverage commonalities between similar words. This is where the idea of vectorized word representations comes in.

Word Representations and Sequence Modeling

Bengio et al. (2003) suggested to re-conceptualize the language modeling task as consisting of two subtasks: firstly, the creation of word representations and, secondly, the prediction of next words based on these representations.

One key idea towards rich and efficient word representations is that of the vectorized or distributed representation. The so-called *distributional hypothesis* suggests that words with similar meanings are distributed similarly, i.e., co-occur frequently within sentences or documents (Harris, 1954). Hence, words may be represented as vectors such that the vectors of closely associated words are close to each other and vice versa. In practice, a big leap towards such word presentations was facilitated through the word2vec tool[1] by Mikolov et al. (2013). It implements two kinds of algorithms: Continuous Bag-of-Words, which predicts a missing word from its given context, and Skip-gram, which predicts the context based on a given word. Another key element of word2vec is the large scale of the training corpus, which presents words in a multitude of exemplary contexts.

As for the modeling of sequences, different types of neural network architectures have been proposed. A classic example is the recurrent neural network (RNN) (Rumelhart et al., 1986). RNNs utilize recurrent connections, such that the previous output that was generated based on one input feature x_t (at timestep t)

1. https://code.google.com/archive/p/word2vec/ (accessed : June 20, 2025)

is fed back into the model together with the next input feature x_{t+1}. The model weights are shared over time to allow the sequential processing of one input in dependence of the previous (Goodfellow et al., 2016).

Encoder-Decoder Models

Sequence-to-sequence modeling tasks, such as machine translation had previously been approached with RNN-based encoder-decoder systems (Cho et al., 2014; Sutskever et al., 2014). The core idea of encoder-decoder architectures is to project an input into a latent space to create an abstract representation of it (*encoder*) and to project from this latent representation to the output space (*decoder*). In the context of sequence modeling, it allows to jointly model the conditional distribution of a sequence on another sequence. For instance, two RNN networks can be combined, one specialized as an encoder and the other as a decoder. The encoder passes sequentially through the input to compute a fixed-sized vector representation, a *context vector c*. The decoder predicts the next word based on this context vector, as well as the past hidden state and its own previous generation (Cho et al., 2014). This principle of generating outputs and then feeding it back to the model to inform the next generated output is called *auto-regression*.

Attention

One major limitation of RNN-based sequence-to-sequence modeling is that the fixed-sized encodings are not suitable for the stable representation of long-range dependencies. This can be alleviated by introducing an *attention* mechanism (Bahdanau et al., 2014). At each output position t, a different context vector c_t is provided (as opposed to one single context vector c across the whole process). This position-specific vector is weighted such that all input sequence elements are emphasized or de-emphasized according on their relevance to t. The weighting is learned. An important effect of this is that the order of the sequence becomes less important and information is more freely shared across all positions (Bahdanau et al., 2014). Vaswani et al. (2017) leveraged this property and introduced an architecture that is fully based on attention and does not utilize any recurrence mechanism. By removing sequentiality from the equation, the input is now processed as a whole. Modeling relationships between words solely based on attention gives a new quality to the resulting word representations. That is, they become *contextualized*. With word2vec, a word can only have one static representation, independent of its current context and position. However, with a transformer model, a single word can have many different representations, depending on its current context and position. This creates a great advantage over previous approaches in that the resulting representations are much more semantically rich (the fruit "apple" can now be distinguished from the brand "Apple").

To summarize, not only do transformers facilitate more efficient modeling of long sequences but also greatly improve the richness of word representations. With this, they present advancements regarding both of the subtasks of language modeling identified by Bengio et al. (2003).

2.3.3 The Transformer Architecture

The traditional transformer architecture introduced by Vaswani et al. (2017) also followed an encoder-decoder structure (see Figure 2.1) and was introduced as a new approach to the machine translation task. As mentioned before, the model does not rely on recurrence and instead is build upon the attention mechanism. The encoder encodes an entire sequence of a fixed length in at once and the decoder generates outputs in an auto-regressive manner. Since the encoding does not happen in sequential manner, sequential information is added through dedicated positional encodings. The original transformer consists of six encoder and six decoder layers.

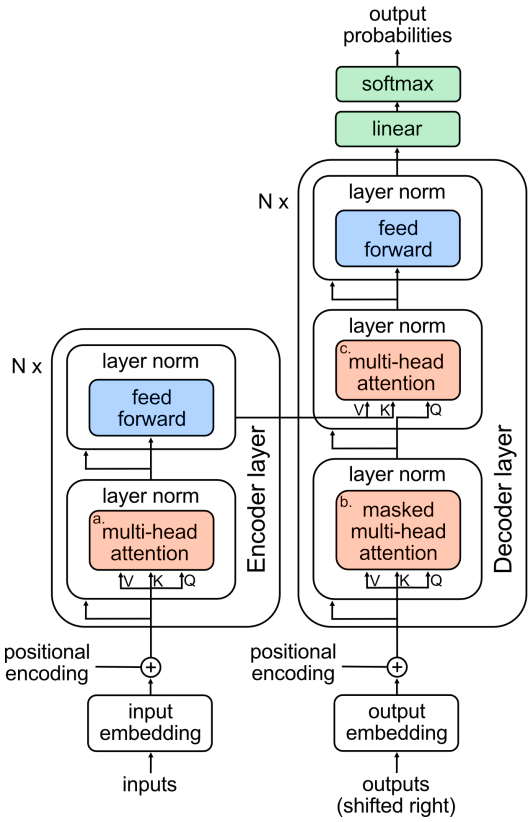

Figure 2.1: Schematic overview of the Transformer encoder-decoder architecture. Visualization designed after Vaswani et al. (2017) and taken from Kraft (2021).

The core element of the transformer is the use of *self-attention*, which allows to encode each single word in dependence to its current context. It maps queries (which information to access) and key-value pairs (indexed content information) through a dot-product and does so in parallel for the whole sequence. Given a query matrix Q, key matrix K, value matrix V, and dimensionality d_k the self-attention function is defined as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.7}$$

Q, K, and V are learned projection matrices. The traditional transformer model utilized eight different sets of matrices, or *attention heads*, that are differently initialized. This is to introduce *multi-head attention*, where the different heads specialize on representing different semantic aspects (see Figure 2.1, a.). Multi-head

attention is uses throughout the encoder and decoder, with a minor difference: To facilitate auto-regressive behavior, the decoder self-attention mechanism only attends to previous positions and masks the following positions (see Figure 2.1, b.). The decoder's generations are also conditioned on the encoder output through multi-head attention. That is, the decoder queries key-value pairs from the encoder. The decoder stack is connected to a linear layer which outputs a logits vector as long as the model's vocabulary. Applying softmax to this vector yields a probability distribution across all tokens in the vocabulary, of which the most likely of one of the most likely tokens can be sampled as the next token (Vaswani et al., 2017).

2.3.4 Large Language Models

The introduction of the transformer architecture gave rise to an ongoing surge of so-called large language models (LLMs). They are large in their number of trainable parameters as well as the data sets they are trained on. Note that what is considered a *large* language model is not coherently defined throughout the NLP community. Its understanding has also been changing over course of the doctoral research endeavors presented in this thesis. While BERT (Devlin et al., 2019) was commonly considered an LLM in 2021, it now is often regarded as a simple language model (LM). Yet, throughout this thesis, I will include older and comparably smaller models such as BERT and RoBERTa (Liu et al., 2019) in my working definition of LLMs. I justify this choice by the observation that the practice of training transformer architectures for language modeling with highly-parameterized models and on web-scale corpora generally constituted a bigger technological leap than the steady increase in scale ever since. As opposed to the original transformer model which implemented an encoder-decoder architecture, most LLMs nowadays are based on either only encoder or only decoder layers (Liu et al., 2019).

Encoder-Only Models

BERT (Pre-training of Deep Bidirectional Transformers for Language Understanding) (Devlin et al., 2019) was the first large-scale transformer model that was pre-trained on a web-scale corpus of text with the purpose of achieving semantically rich, contextualized word embeddings that could be used for different types of downstream tasks. The input to LLMs like BERT has to firstly be tokenized, i.e., sentences have to be translated into a quantified representation. As discussed before, utilizing commonalities between words is a fruitful way to reduce sparsity in vocabularies. BERT, hence, utilizes a subword-level tokenization strategy, WordPiece. Strictly speaking, BERT was not built with the purpose of solving the language modeling task. However, the paradigm used to obtain the word embeddings happens to be a combination of variants of the language modeling task, i.e., *masked language modeling* (MLM) and *next sentence prediction* (NSP). The MLM task applies a random masking of tokens in the input sentence which are then to be predicted by the model considering the input in both directions of the mask. In the NSP, the model is provided two concatenated sentences or segments, where right second one is either a true successor or not. The model has to identify whether or not one or the other is the case. The authors published

two architectures similar to the encoder part of the original transformer but with double and quadruple the amount of layers, more free parameters per feedforward network, and more attention heads.

RoBERTa builds on BERT but uses a slightly reworked MLM objective with a dynamic masking algorithm and discards the NSP objective. Additionally, it utilizes an improved tokenization approach which exaggerates the subword principle even more. Byte-pair encoding (BPE) bases the vocabulary on bytes, which allows to store bigger vocabularies. Most importantly, however, the model was trained on a significantly bigger dataset and for more iterations (Liu et al., 2019).

Decoder-Only Models

Decoder-only LLMs are closely conceptualized around the next-word-prediction task and are more commonly designated as *generative language models*.[2] They are used to generate continuations or answers to a given input or *prompt*. As mentioned in Section 2.3.3, the final output of the decoder layer is a distribution of probabilities across the whole token vocabulary. Generating an output thus entails sampling the or one of the most likely next tokens from this vocabulary.

Even without the encoder part, generative language models have been found to capture rich word semantics through the uni-directional auto-regressive learning paradigm. Mainly due to a steady increase in dataset and model size, LLMs have been becoming more and more capable. This trend can, e.g., be observed at the example of the GPT (Generative Pretrained Transformer) series (Radford et al., 2018; Radford et al., 2019; Brown et al., 2020; OpenAI, 2023). Over time, these models grew in size (number of layers and nodes), allowed for longer context windows (due to optimized attention mechanisms and positional encodings), were trained on increasing amounts and more specialized training data, and were optimized regarding their tokenization strategies.

At this point, the focus of innovations in the area of language modeling is far beyond optimizing the generation of plausible word sequences. LLMs have become so-called *general-purpose AI* in that they can be used for all sorts of problem scenarios and domains (see, e.g., Llama 4,[3] GPT-5,[4] Claude.[5]) They are also often framed as *foundation models* because they encode vast linguistic and world knowledge and, hence, provide reusable foundations for a plethora of specialized AI systems. Such models can be used via *zero-* and *few-shot prompting*, where the LM is instructed to solve a task for which it was previously not trained. In a few-shot setting, the prompt includes a task instruction and a couple of labeled examples to demonstrate the correct solution strategy (Brown et al., 2020; Luo et al., 2023). In zero-shot prompting, the model is given the instruction, but no examples (Radford et al., 2019; Socher et al., 2013).

2. In the public discourse, generative large language models like GPT-4 or DeepSeek-R1 are also often dubbed *generative AI* even though they only exemplify one kind of generative AI, strictly speaking.

3. https://www.llama.com/models/llama-4/

4. https://openai.com/gpt-5/

5. https://claude.com/product/overview

2.3.5 Human Preference Alignment and Reasoning

The leap from the type of language model that simply predicts the next word to systems like ChatGPT which appear to be assistants and conversational partners was to a large part facilitated by the principle of *reinforcement learning from human feedback* (RLHF). Ouyang et al. (2022) were the first to utilize this approach with the goal of *aligning* LLMs to user preferences, i.e., to make them *helpful*, *honest*, and *harmless*. More recently, another substantial improvement in the assistive capacities of LLMs was achieved through the so-called *reasoning* paradigm. In the following, RLHF and reasoning are explained in more detail.

Reinforcement Learning from Human Feedback

The RLHF procedure is commonly applied to a pre-trained language model with good text generation performance. The first step is then supervised finetuning: A pre-trained language model–GPT-3 in the case of Ouyang et al. (2022)–is finetuned in a supervised manner, on a dataset comprised of conversation-style examples; with pairs of input prompts and desired answers as the target. This is to introduce respective linguistic patterns to the model. This process is also termed *instruction tuning* and its effect is best illustrated with an example; see Table 2.1. Steps two and three then introduce a reinforcement learning-based finetuning to the language model. To this end, a reward model is trained in step two. Again example prompts are paired with desired responses. Only this time, multiple potential answers are ranked by a group of annotators, from best to worst. The reward model is a neural network trained to predict a human preference rating from this ranking data (Kaufmann et al., 2024). In the third and final step, a policy is trained with the help of the reward model. Ouyang et al. (2022) initialize the policy itself with the given pre-trained language model. All layers except the output layer are frozen such that only the latter is updated. Finally, the training objective is based on the *Kullback–Leibler divergence* (KL divergence; measures the difference between two probability distributions over the same random variable) between this policy and the initial language model. The purpose of this is to converge towards a compromise between the optimal policy and the good linguistic capabilities that the pre-trained model carries.

Note that instruction tuning (step one) is a procedure that can be applied in isolation to RLHF (Wei et al., 2021). Conversely, RLHF also works without instruction tuning as long as the basic language model has been trained on a sufficiently diverse linguistic basis.

Overall, RLHF has improved the interactive experience of LLMs by allowing LLMs to generate responses that emulate instruction-following and conversational turn-taking (Ouyang et al., 2022). By now, more efficient variants of the initial method have been proposed. For instance, Rafailov et al. (2023) developed the *direct preference optimization* approach, which facilitates human preference alignment without relying on reinforcement learning. This method was used by Meta to train Llama 3 and Llama 4. Tuning for human preferences also poses a number of unique challenges, such as making sense of annotators' biases and positionality, or handling erroneous feedback due to malicious intents or a lack of understanding (Casper et al., 2023).

Prompt	Explain the moon landing to a 6 year old in a few sentences.
Completions	
GPT-3	Explain the theory of gravity to a 6 year old. Explain the theory of relativity to a 6 year old in a few sentences. Explain the big bang theory to a 6 year old. Explain evolution to a 6 year old.
InstructGPT	People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we could all see them.

Table 2.1: Example from OpenAI (https://openai.com/research/instruction-following; accessed: June 30, 2025) showcasing the effect of LM finetuning with conversation-style data.

Chain-of-Thought Prompting and Large Reasoning Models

Wei et al. (2022a) introduced the *chain-of-thought* (CoT) prompting method, which supposedly elicits "reasoning abilities" in "sufficiently large language models" (p. 1). A CoT prompt is a few-shot prompt, composed of an example task-response pair and an open task for the model to respond to, whereas the example and open tasks are related. The example response demonstrates how a task similar to the open task would be broken down into smaller problems, to facilitate a step-by-step problem solution. An example used by Wei et al. (2022a) is given in Table 2.2. Kojima et al. (2022) found that this principle also works in a zero-shot setting. Instead of an example task-response pair, the authors append the sentence "Let's think step by step" to the prompt (see Table 2.3). These techniques were found to improve performance on more complex tasks, such as mathematical problem solving as measured by the GSM8K benchmark (Cobbe et al., 2021). A newer type of LLM, often referred to as *large reasoning model* (LRM), is trained to perform CoT per default. To this end, a conventional pre-trained LLM is further trained to generate reasoning steps before the final output. These reasoning steps are refined through reinforcement learning, for instance, by judging the concrete line of reasoning with human feedback (Lightman et al., 2024) or by judging only the correctness of the final response DeepSeek-AI (2025).

2.3.6 The Factual Inaccuracy Problem

One major limitation of text generation with LLMs is their inherent tendency to produce outputs that appear plausible but are factually incorrect. This matter is oftentimes dubbed as *hallucination* as some perceive it to be comparable to the psychological phenomenon of "an unreal perception that feels real" (Ji et al., 2023, p. 248:3). Other scholars argue that this denomination is distracting from the actually statistical nature of the issue and contributing to a problematic anthropomorphism (Shanahan, 2024; Bender and Hanna, 2025). Hence, this thesis defaults to notions such as *factual inaccuracy* or *factual infidelity* unless it references accounts by other authors.

Based on empirical observations, Li et al. (2024) distinguish different classes of inaccuracies: *Entity-error hallucinations* are errors in which the wrong entity

Prompt	
Q:	Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A:	Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.
Q:	The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?
Completion	
A:	The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9.

Table 2.2: Example of chain-of-thought prompting from Wei et al. (2022a).

Prompt	
Q:	A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:	Let's think step by step.
Completion	
	There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls.

Table 2.3: Example of zero-shot chain-of-thought prompting from Kojima et al. (2022).

(person, date, location, etc.) is mentioned in the given context. *Relation-error hallucinations* refers to faulty descriptions of the relationship between entities. *Incompleteness hallucinations* describes the issue incomplete responses to user prompts, e.g., when asked for aggregated facts. *Outdatedness hallucinations* relate to statements that were accurate in the past but not anymore. *Overclaim hallucinations* refer to statements that are partly factual and partly confabulated. *Unverifiability hallucinations* are given when statements are not verifiable.

This very broad list of errors and inaccuracies already indicates that there is no narrow definition of the concept. Reddy et al. (2024), for example, also consider algorithmic bias to be a form of hallucination. Hence, the presumed causes for these errors are equally manifold. Ji et al. (2023) discuss factors related to data, training, and inference. The data collected might, for instance, contain inappropriate evidence documents or the dataset might simply be linguistically very diverse ("chit-chat style") and introduce overclaims that way. As for the training, it might be affected by an imperfect encoder model that yields faulty representation

learning. Similarly, the decoder could be faulty and, for example, attend to the wrong input aspects or draw from false semantic associations. Or, the decoding strategy could be parameterized such that diverse text generations are favored (to get more "creative" texts). This also increases erroneous statements. Moreover, the trained model might be prompted for examples that are very different from its training data. If the prompt is tapping into a "knowledge gap region of the model" (Agrawal et al., 2024, p. 3947), it produces any token that is likely to follow given what it has been exposed to during training. And finally, there are often inconsistencies between the knowledge stored in the model parameters versus the provided input during inference, in which cases models tend to default to their parametric knowledge (Ji et al., 2023). Reasons mentioned by Reddy et al. (2024) include data biases, the limited ability of a LLM to consider wider contexts, ambiguous prompts, adversarial attacks, overfit models, as well as aspects of the model architecture (e.g., larger models are assumed to cause more hallucinations).

There are a range of different mitigation approaches which can coarsely be grouped into prompt engineering-based or model development-based (Tonmoy et al., 2024). Generally, the idea of enhancing the model input or model itself with information from an external knowledge base has been considered promising. More details are discussed in Section 2.4.4 after firstly introducing the concept of knowledge graphs in the following paragraphs.

2.3.7 Note on AI and *Knowing*

There are a few things to note regarding the issue of knowing in the context of AI, and language models in particular. Scholarly articles concerned with the knowledge represented within language models, happen to be generally vague and inconsistent regarding their underlying conceptualization of knowledge, e.g., whether or not (a) *a language model contains knowledge*, i.e., stores content that qualifies as knowledge or (b) *a language model knows*, i.e., is itself a knowing entity. A study by Fierro et al. (2024) showed that this inconsistency generalizes across the field of NLP. The authors compared the knowledge concepts implied or explicated within NLP publications with the most common definitions found in the epistemological literature. The authors also conducted a questionnaire study with circa 100 philosophers and computer scientists on the topic. Regarding the question whether or not LLMs *do know* (as of now), 54% of the philosophers disagreed and 11% agreed. In contrast, 31% of the computer scientists disagreed and 34% agreed. Regarding the question whether or not LLMs *can know* (in theory), 33% of the philosophers disagreed and 24% agreed, and 21% of the computer scientists disagreed and 55% agreed. While there is a certain level of division within both disciplines, computer scientists were shown to be generally more optimistic regarding language models' existing or theoretical ability to know.

That said, attributing language models the capacity to know, has been criticized as being harmfully anthropomorphizing. It feeds into a narrative that portrays AI as being on the verge to become intelligent in ways similar to humans or better. This has individuals in fear or excitement over scenarios, that are completely hypothetical at this point. Seeing as to how AI research and the economy around it is highly influenced by the money and agendas of Big Tech and how some of the mythology around AI is granting Big Tech much attention and political

justification, there is reason to assume that anthropomorphism is a marketing ploy (Bender and Hanna, 2025). For this reason, this thesis deliberately does *not* attribute *knowing* to AI. The term *artificial intelligence* itself is used for it being an established identifier used by critics and supporters alike, and clearly demarcates not only the technology but also all the discourse that is attached to it. This discourse is relevant to the research presented in this thesis.

2.4 Knowledge Graphs

As mentioned in the introduction to this Chapter, an important pillar of AI is the storage and processing of knowledge. While certain knowledge can be stored implicitly in the parameters of a neural network-based model, other types of knowledge is evidently hard to encode this way, leading to persistent factual inaccuracies Reddy et al. (2024). Knowledge graphs, on the other hand, are designed to precisely store knowledge in an explicit, machine-readable structure. They serve as an important building block of the Semantic Web, as well as a range of knowledge-driven NLP technologies. Moreover, there are a range of approaches, pursuing the idea of knowledge-enhanced language modeling, that combine knowledge graphs and language models in order to leverage each of their strengths. The following sections give an overview of the basic definition and typical characteristics of knowledge graphs in general, the concrete knowledge graph Wikidata, and knowledge-enhanced language modeling.

2.4.1 Definition

The modern concept of a knowledge graph was inspired by the Google Knowledge Graph in 2012[6] and the subsequent development of industrial KGs by several other big technology companies (Hogan et al., 2021b). Different definitions of a knowledge graph exist. However, this thesis works with one of the most cited definitions which was given by Hogan et al. (2021b). The authors define a knowledge graph as:

> [...] a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent potentially different relations between these entities.

As such, a KG employs a graph-based data model to capture "word knowledge" at large scale. The authors continue by defining knowledge based on an account by Nonaka and Takeuchi (1995) as *explicit knowledge*, "i.e., something that is known and can be written down" (Hogan et al., 2021bp. XX:3). They further characterize knowledge as something that can be expressed in quantified or unquantified statements, such as "Santiago is the capital of Chile" and "all capitals are cities" which has implications regarding the structure of knowledge graphs. While the first example can be easily represented, quantified statements require *ontologies* to handle complexity.

6. https://blog.google/products/search/introducing-knowledge-graph-things-not/ (accessed: June 27, 2025)

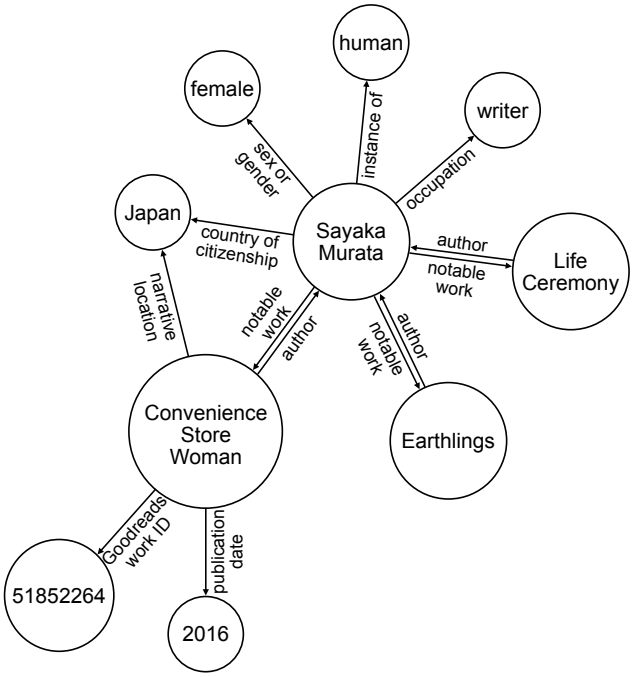

Figure 2.2: Example of a knowledge graph.

2.4.2 Multi-Relational Graphs

KGs can employ different types of data graphs. The most popular type, however, is the *directed edge-labeled graph* or *multi-relational graph*. It is defined as a tuple $G = (V, E, L)$, where V is a set of nodes nodes that represent entities (e.g., *Japan*, *human*, etc.), L is a set of edge labels, and $E \subseteq V \times L \times V$ is a set of edges that each connect two entities (*Sayaka Murata→occupation→writer*) (Hogan et al., 2021a). For an examplary knowledge graph, see Figure 2.2. This structure brings several benefits in comparison to standard relational models or hierarchically structured data models, like XML or JSON, in that its schema does not need to be predefined, entities need not be hierarchically structured, and cyclic relationships are allowed (Hogan et al., 2021a).

An influential standard for the multi-relational graph model is the Resource Description Framework (RDF)[7] which was proposed by the World Wide Web Consortium (W3C). An RDF graph is defined as a set of [*subject*; *predicate*; *object*] triples. Nodes can be an *Internationalized Resource Identifier* (IRI), a *literal*, or a *blank node*. According to W3C, the IRI concept is a generalization of the *Uniform Resource Identifier* (URI) concept in that it permits a greater range of Unicode characters. An IRI is a globally-unique identifier for a node or edge label and helps reduce the risk of ambiguities and name clashes when a KG is extended with data from an external source (Hogan et al., 2021a). Literals are used to represent string, number, and date values. Blank nodes are distinct from IRIs and literals and may be used to represent an entity in a non-persistent, uniquely identifiable way.

7. https://www.w3.org/TR/rdf11-concepts/ (accessed: June 27, 2025)

2.4.3 Wikidata

The example in Figure 2.2 is based on entries in Wikidata, one of the largest and most influential public knowledge graphs. It was introduced by the Wikimedia Foundation in 2012 as a sister project to Wikipedia and contains the same kind of encyclopedic knowledge. It was developed in response to the difficulties in accessing specific pieces of information within the web encyclopedia's millions of articles in hundreds of languages (Vrandečić and Krötzsch, 2014). To date, Wikidata comprises roughly 1.68 billion triples[8] and is continuously growing. It has been found to grow faster than English Wikipedia.[9] The data and schema of Wikidata are openly edited by a community of users and bots (Piscopo et al., 2017). As of May 2025, there were more than 13 thousand active editors.[10, 11] Their levels of leadership and activity in the community differs strongly. Only a small group take on leading responsibilities, e.g., involving improvements to the ontology, and those generally tend to be early community members (Piscopo and Simperl, 2018).

Wikidata allows to represent conflicting information in parallel and contains information in different languages within one coherent graph (as opposed to individual sites as is the case for Wikipedia) (Vrandečić and Krötzsch, 2014). Currently, labels, aliases, entity descriptions in more than 350 languages are available (Cantallops et al., 2019). Finally, all data stored in Wikidata are openly accessible under a Creative Commons CC0 License,[12] e.g., as an RDF graph (Erxleben et al., 2014) or a JSON dump.

Even though its community has developed a number of quality measures, such as the use of references to support claims stored in the data graph (Piscopo et al., 2017; Vrandečić and Krötzsch, 2014), the crowdsourced and open nature also introduces problems. For instance, the ontology has been found to be "messy", i.e., incoherent (Piscopo and Simperl, 2018). Wikidata has no strict pre-defined taxonomy which allows for great flexibility (Hogan et al., 2021a) and, in practice, hierarchies are represented through subclass relationships. However, it has been found that the use of classifications and representation of hierarchical relationships is often inadequate (Brasileiro et al., 2016). Another frequently observed issue is vandalism (Heindorf et al., 2019). Nevertheless, its scale, frequent updates, multilinguality, and accessibility render Wikidata an important data resource for NLP research and language modeling, in particular (Cantallops et al., 2019).

2.4.4 Knowledge-Enhanced Language Modeling

Knowledge-enhanced language modeling describes a family of approaches that leverages external knowledge resources to improve language models. Such knowl-

8. https://grafana.wikimedia.org/d/000000175/wikidata-datamodel-statements?orgId=1&refresh=30m&from=now-90d&to=now&timezone=browser (accessed: June 29, 2025)

9. https://meta.wikimedia.org/wiki/Data_dumps/Dumps_sizes_and_growth (accessed: June 29, 2025)

10. https://stats.wikimedia.org/#/wikidata.org/contributing/active-editors/normal|line|2-year|(page_type)~content*non-content|monthly (accessed: June 29, 2025)

11. Active editors are defined as "registered, non-bot editors with five or more edits in a given month, including on redirect pages." https://meta.wikimedia.org/wiki/Research:Wikistats_metrics/Active_editors (accessed: June 29, 2025).

12. https://www.wikidata.org/wiki/Wikidata:Licensing (accessed: June 29, 2025)

edge resources can be manifold, such as knowledge graphs (subgraphs or triples), individual entity information, or textual information (e.g., from an encyclopedia such as Wikipedia). Knowledge enhancement is seen as a possible remedy to the factual inaccuracy problem (which, depending on the conceptualization might be understood to include algorithmic bias (Reddy et al., 2024)).

Agrawal et al. (2024) distinguish between three types of methods: (1) *knowledge-aware inference*, (2) *knowledge-aware learning*, and (3) *knowledge-aware validation*. Knowledge-aware inference procedure introduce external information from knowledge graphs to the model input in order to expand the given context.

One type of knowledge-aware inference is *KG-augmented retrieval*: A direction of research in this area that has gained a lot of attention in the past years is *retrieval augmented generation* (RAG) (Lewis et al., 2020). Given a language model, an external knowledge base, and an input prompt, RAG methods firstly perform a matching between the input and the knowledge base to identify supporting context. The latter is then combined with the input to create a contextually more informative query which is then finally sent to the actual language model. Such retrieval-based methods have been shown to improve especially smaller models. Another method that has generally been found to alleviate factual inaccuracies to some extend is the use of *chain-of-thought* (Wei et al., 2022b) or other prompts that induce a step-by-step problem solution procedure. These were found to improve the fidelity of larger models, in particular. Agrawal et al. (2024) point out that such procedures can also be combined with knowledge graph retrieval. They dub this type of knowledge-aware inference *KG-augmented reasoning*. Finally, external knowledge sources can be utilized for *knowledge-controlled generation* by inducing factual information retrieved, e.g., through SPARQL into the generated output.

As for knowledge-aware learning, different approaches for *knowledge-aware pre-training* and *knowledge-aware finetuning* exist (Agrawal et al., 2024): Models can, for instance, be simultaneously trained on unstructured text and facts extracted from knowledge graphs or finetuned on the latter, respectively. Another way is to directly fuse language models with knowledge graphs, for example, through fusion of features or embeddings (Yang et al., 2024a). These approaches are of course more costly than adjustments to the inference procedure. Finally, knowledge-aware validation employs mechanisms like a fact-checking module or an explanation module to either improve or justify the generated output (Agrawal et al., 2024). Again, this class of approaches is generally on the more cost-heavy side. Based on a research trend analysis, Agrawal et al. (2024) conjecture that inference-based approaches have been receiving the most attention in recent years.

2.5 Algorithmic Bias

One of the most publicly discussed cases of a biased machine learning system and its real-world harms is that of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm, a system used by U.S. courts to predict recidivism and inform the severity of a sentence. In 2016, journalists from the investigative journalism platform ProPublica revealed that the system found Black offenders to be more likely to re-offend than White offenders, while also

having a higher prediction error rate for Black individuals.[13] It became apparent that the system was set up in such a way that it judged Black people more unfairly. Furthermore, it was set up to contribute to a self-fulfilling prophecy; to a cycle of longer prison sentences, unemployment, and consequent recidivism (O'Neil, 2016). Friedman and Nissenbaum (1996) emphasize that the issues that arise with biased systems are worsened by the ease with which they are disseminated in society. Moreover, biased decision making algorithms reify previously implicit biases as formalized rules which can further obfuscate the issue (Barocas and Selbst, 2016). Hence, systems like COMPAS are what O'Neil (2016) considers to be "Weapons of Math Destruction", which cause *damage*, through *opaque* mechanisms, at *scale*.

This thesis focuses on AI-driven knowledge technologies, which have become omnipresent and, simultaneously, are highly opaque. What needs to be understood is to which extend they may or may not be damaging, i.e., contributing to knowledge-related injustice. To be able to evaluate this latter aspect, a deeper understanding of algorithmic bias and related measures is firstly required. Hence, the following sections give a detailed overview of common definitions and causes of algorithmic bias, as well as methods for its measurement and mitigation.

Note that biases in KGs, in particular Wikidata, are just as relevant to this thesis as those found in NLP systems. However, the bias issue is much more discussed and researched in the context of NLP and less so in the context of KGs. In fact, it is one of the main contributions of this thesis to give an overview of the characteristics and measures of, and the potential solutions for the biases in encyclopedic KGs. Thus, more details can be found later on, in Chapter 4.

2.5.1 Definition

The term "bias" generally refers to some type of skew in behavior or skew in a distribution, which in itself may or may not be ethically neutral (Friedman and Nissenbaum, 1996; Shah et al., 2020). Drawing from Bayesian terminology, Hovy and Prabhumoye (2021) argue that bias can be understood as a preset, an expectation we have for some phenomenon before further evidence is available. This expectation only becomes ethically problematic when divergent evidence is ignored or when applied to out-of-context situations. This thesis is concerned with an ethically laden type of bias affecting and occurring in algorithmic systems, in particular in NLP systems. According to Friedman and Nissenbaum (1996), computer systems are problematically biased if they "*systematically* and *unfairly [discriminate]* against certain individuals or groups of individuals in favor of others," by "[denying] an opportunity or good or [assigning] an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate" (p. 332).

Note that, according to this definition, bias equals *unfair* discrimination. And in fact, *bias* and *fairness* are often treated as two ends of the same spectrum. That is, *algorithmic bias research* is also sometimes referred to as *algorithmic fairness research*. Similarly, the terms *bias metric* and *fairness metric* are often used interchangeably. Note that the latter term is, wherever possible, avoided

13. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

in this thesis. That is, while we may speak of bias in purely statistical ways, fairness is always an inherently problematic concept that is deeply entrenched in philosophical discourse. A major criticism of computer science research in this field is that this discourse is, however, largely ignored and the notion of algorithmic *fairness* is underdefined (Blodgett et al., 2020). In Section 2.5.3, different metrics are presented, which all *imply* different distributional assumptions about the conditions of fair system behavior.

Social Groups, Identity, and Sensitive Attributes

As algorithmic bias is usually conceptualized as the difference in system behavior for different *social groups*, we may ask how group membership is determined. According to Gallegos et al. (2024), a social group is defined as "a subset of the population that shares an identity trait, which may be fixed, contextual, or socially constructed" (p. 1104). Identity traits that commonly occur in algorithmic bias research are, for instance, gender, nationality, ethnicity or race, religion, and sexual orientation. These are often also subsumed under the term *protected attributes* which is derived from U.S. anti-discrimination law. An alternative notion is that of the *sensitive attribute*, which is less U.S.-specific (Barocas et al., 2023). In the algorithmic bias literature, these terms are used mostly as a mathematical denotation of distributional features. In practice, these features are considered sensitive since they are in some way associated with historically and politically rooted struggles related to discrimination, marginalization, and mistreatment. As Gallegos et al. (2024) discuss, such narrow conceptualizations of (shared) identity can be reductive and harmfully reinforce societal boundaries and marginalization. Yet, they are nicely translatable into formalisms allowing to measure and make visible disparities. This, in fact, is an ongoing debate in the field and gives rise to some fundamental criticisms of algorithmic bias and fairness research, in general. The question what it is that yields group identity is also a longstanding subject of debate within feminist philosophy and further addressed in Section 3.3.

Individual versus Group Fairness

As is indicated in the definition given by Friedman and Nissenbaum (1996), bias may be determined on an individual or a group basis. *Individual fairness* requires that similar individuals are treated similarly (Gallegos et al., 2024). Similarity, here, is determined via an attribute or set of attributes specified with regards to a given task and context. *Group fairness* builds on the notion of social groups and requires near parity among all groups with regards to the system outcome. While individual fairness might be too granular of a measure in many cases, group fairness tends to be oversimplifying and leave certain subgroups disregarded. For instance, determining (dis-)parity solely in terms of gender disregards potential intersectional effects.

Allocational versus Representational Harms

In the context of NLP, biases are often defined by the harms they cause, namely *allocational* or *representational harms* (Blodgett et al., 2020; Barocas et al., 2017). Allocational harms refer to the unfair distribution of resources or opportunities as

a consequence of the model behavior. Representational harms denote issues of stereotyping, as well as systematic and unfair misrepresentations and denigrations of certain social groups within model outputs. It also comprises matters of unequal system performance.

2.5.2 Causes for Algorithmic Bias

Friedman and Nissenbaum (1996) provide a helpful coarse categorization of bias in computer systems: *Preexisting bias* is rooted in society and may influence the biases of developers who then unconsciously or consciously introduce them in to their systems. *Technical bias* arises from the technical design or implementation, such as hardware or software limitations, or mathematical inaccuracies. In particular, the act of quantifying, discretizing, and formalizing human constructs in order to make them machine processable is prone to introducing technical bias. And, vice versa, "the process of ascribing social meaning" (p. 335) to algorithms and numbers. Finally, *emergent bias* ensues after the finished system has been deployed, due to contextual changes, repurposed usage, or user interfaces unfit to the audience.

The biases of modern-day AI systems, in particular, are to a large part explained by the biases of the datasets they are trained and evaluated on. These datasets are usually very large in scale and fall under the category of *Big Data*. With a view to analyzing the disparate impact of large-scale data mining efforts, i.a. Big Data, Barocas and Selbst (2016) note that "data mining can reproduce existing patterns of discrimination, inherit the prejudice of prior decision makers, or simply reflect the widespread biases that persist in society" and that it "can even have the perverse result of exacerbating existing inequalities by suggesting that historically disadvantaged groups actually deserve less favorable treatment" (Barocas and Selbst, 2016, p. 674). In the context of NLP, Hovy and Prabhumoye (2021) point out that not only data aspects but also aspects of the algorithm itself factor into the equation. Moreover, they highlight the role of the overall research design and the interests of the researchers or engineers responsible for developing the system. Finally, the works of Bowman and Dahl (2021) and Raji et al. (2021a) draw attention to the role of evaluations. That is, biased evaluation datasets can cause corresponding biases within models to remain undetected or indirectly rewarded. In the following, all of these different factors–data, model, and experimental conditions–will be illuminated in more detail as dominant causes of algorithmic bias.

Data Sample

One of the main causes for algorithmic bias are biases encoded within the training and evaluation datasets. Sampling biases or the collection of data that strongly reflect historic prejudices and injustices can yield dataset biases. Problems then arise if the sample chosen is unrepresentative of the population of interest, but analyzed under the assumption of generalizability. *Under-representation* can lead to inaccuracies and blind spots, such as NLP systems that are not attuned to certain vernaculars (Bella et al., 2024). In certain cases, *over-representation* can lead to an overestimation of negative outcomes. A non-NLP example would be the over-monitoring of crime rates for certain zip codes that lead to an overestimation of crime risk for associated demographics (Benjamin, 2019). Secondly, if the

sample distribution is representative of some real-world distribution, this real-world distribution could in itself be biased due to pre-existing societal inequalities. A good example is the significant gender disparity in modern-day computer science, which used to be a female-dominated field in the 1940s and 50s (Perez, 2019). Whether or not a sample should be representative of this real-world distribution or an idealized distribution is a matter of choice. To be considered here is that modeling a discriminatory status quo can contribute to the perpetuation of said issue at scale. Thirdly, the data quality could differ in correspondence to social group membership, again leading to differences in system accuracy for marginalized versus non-marginalized groups. And finally, the data contents might be biased by way of their associations or portrayals. For instance, Birhane et al. (2021) investigated the large-scale LAION-400M (=400 million data points) dataset used to train multimodal models. It consists of images paired with text descriptions (alt texts) scraped from the web. They found that the dataset contains disproportionately high amounts of NSFW content associated to Desi or Latina subjects. This can introduce problematic intersectional biases to AI models and yield the generation of derogatory depictions. In a newer study, Birhane et al. (2023) examined the much larger successor dataset LAION-2B (=2 billion data points) and revealed that the proportion of hateful content, in fact, increased by 12%.

How the data is cleaned and filtered after sampling also contributes to the composition of the training data. In the case of the LAION datasets, e.g., an automated filtering procedure was applied with the purpose of removing toxic and NSFW content. However, the AI model used in this step was in itself biased and systematically failed to remove the type of racist and sexist items, that Birhane et al. (2021) later detected in her elaborate analysis. Rule-based filters, on the other hand, can also introduce issues. For instance, Alphabet used the publicly available list of "Dirty, Naughty, Obscene, and Otherwise Bad Words"[14] to filter a Common Crawl data dump and create their version of it, the *Colossal Clean Crawled Corpus* (C4; Raffel et al., 2020). This particular filter list contains several words that link to health content for LGBTQIA+ communities. As a consequence, harmless contents related to the LGBTQIA+ community were systematically removed from the C4 corpus, introducing a social bias in the dataset (Dodge et al., 2021).

Annotations

What is modeled in a predictive model is the relationship between input features and output features, i.e., annotations or labels. Part of the this process is the choice of possible outputs, i.e., the categories they can fall into or the possible value range, which is highly subjective and normative (Bowker and Star, 2000).

The actual annotation of training examples with these labels is susceptible to annotator biases. There are several known manifestations of annotator bias. For example, fatigue, boredom, and misunderstandings can lead to false annotations and negatively impact the model performance (Hovy and Prabhumoye, 2021). Yet, these issues do not necessarily yield discriminative outcomes. More problematic are scenarios in which the task invites subjective judgments over the right or wrong label. For example, datasets used for the training of a toxicity classifier

14. https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words (accessed: August 11, 2025)

consist of textual examples that are toxic or non-toxic descriptions of certain individuals or groups. Whether or not a comment is considered toxic and to which extend is highly subjective. This type of judgment tends to be correlated with an annotator's personal identity and political ideas (Sap et al., 2022; Pei and Jurgens, 2023). Moreover, not all annotators are familiar with the same type of language. When it comes to particular vernaculars and culturally diverse expressions, annotators have been found to misunderstand banter or reclaimed slurs as aggression (Sap et al., 2019a). Consequently, the demographic composition of annotators has significant impact on the distribution of ground-truth labels and can introduce systematic biases to the model. With regards to automated labeling, Barocas and Selbst (2016) argue that learning from previous cases to label new instances may act as a formalization and perpetuation of the biases or prejudices of past decision makers (e.g., automated hiring algorithms trained on previous discriminatory hiring decisions).

As mentioned before, the data used to train generative AI systems are oftentimes filtered through automated means. This includes other AI classifiers, which are, in turn, trained on annotated data. That is, classifiers are trained to distinguish problematic from unproblematic contents and flag the items accordingly. In this sense, annotator biases can also contribute to generative model biases.

Input Representations and Models

Input features are selected to act as an intermediary representation for a phenomenon of interest, observed in the real world. To translate such a phenomenon into a quantitative representation that may be used as input to a classification system, it needs to be discretized and formalized in a necessarily reductive act (Friedman and Nissenbaum, 1996). This act is a matter of design choice and inherently subjective and dependent on specialized knowledge about the real-world problem that is to be modeled. This renders it vulnerable to influence through preexisting biases even if the choice is done solely with statistical considerations in mind (Barocas and Selbst, 2016). That is, selecting the features that explain most variance might yield features that are too coarse to achieve balanced accuracy across groups or otherwise confounded with protected attributes. They can also be (potentially unintentionally and unknowingly) correlated with social group membership and ultimately act as a proxy for a protected attribute. Barocas and Selbst (2016) point out that data mining processes can also be designed with the intention to mask discrimination. That is, complex proxies can be built to misrepresent certain groups of individuals in a way that is hard to trace for third parties.

Many NLP approaches utilize embeddings to quantify input texts. A wide range of evidence clearly indicates that popularly embeddings trained on large-scale web-scraped corpora capture or amplify real-world biases and stereotypes. Word embeddings like GLoVe and fastText (Caliskan et al., 2022), and word2vec (Bolukbasi et al., 2016) encode gender bias. Similarly, contextualized embeddings like ELMo, BERT, GPT, and GPT-2 also encode gender bias (Zhao et al., 2019; Kraft et al., 2022) and biases at the intersection of gender and race (Guo and Caliskan, 2021).

Research Design and Evaluation

Hovy and Prabhumoye (2021) argue that overarching aspects of the research design also introduce certain biases to NLP systems. In particular, they are highlighting the NLP community's focus on tools for the English language. The authors observe a self-fulfilling prophecy here, in that researchers tend to study for which a lot of data is available. Along the way, they create more data in this particular language. Generally, the lack of studies regarding low-resource languages has been a widely discussed issue in the field, for more than a decade (Bender, 2011). Bias is introduced by marking English and its respective morphology and syntax as the default and incentivizing models optimized in this regard (Helm et al., 2024).

Another aspect related to the experimental conditions that has been hypothesized to introduce bias, are the evaluation methods. The performance of an AI model is commonly determined via established benchmarks. Only when a certain performance score (or set of performance scores) is met, will AI developers deem the development process completed and successful. In their extensive critique of existing AI benchmarking practices, Raji et al. (2021a) argue that "all datasets come with an embedded perspective — there is no neutral or universal dataset" (p. 8). That is, a benchmark consists of an evaluation dataset and some kind of metric. So whatever causes biases in training datasets causes the exact same biases in evaluation datasets. This leads to distorted evaluation scores. For this reason, Bowman and Dahl (2021) demand that benchmarks must be designed to explicitly disincentivize biased model performance. As is indicated in Chapter 6, this call remains as relevant as ever.

2.5.3 Algorithmic Bias Measurement

Algorithmic bias is commonly measured by means of dedicated metrics. These metrics evaluate model outputs against certain *fairness* or *non-discrimination criteria* (Barocas et al., 2023). The following section introduces respective criteria, which can be directly utilized to identify biases in classification or downstream task performance. Then, three classes of bias metrics are presented, that compute bias on the basis of embeddings, token probabilities, or generated texts.

Non-Discrimination Criteria for Classification

In the following, an overview of non-discrimination criteria in the context of machine learning classification and regression/prediction is provided. It shall be noted that both are similar in mathematical terms. The main difference is the output type: classification outputs are discrete and regression outputs are continuous. The modeling approaches and training paradigms, however, are identical. According to Barocas et al. (2023), non-discrimination criteria can all be clustered into three fundamentally different, overarching criteria. Let Y be a target variable, R the output score of a classifier, and A a sensitive attribute. With Y, R, and A being random variables, the three criteria are defined as follows:

1. *Independence*: $R \perp A$. The sensitive attribute is statistically independent of the output score. In the case of binary classification also known as *demographic parity*, *statistical parity*, *group fairness*, or *disparate impact*.

2. *Separation*: $R \perp A|Y$. The score is conditionally independent of the sensitive attribute, given the target value. In other words, the false negative and false positive rates are equal for all groups. Equals *error rate parity*.

3. *Sufficiency*: $Y \perp A|R$. The target value is conditionally independent of the sensitive attribute, given the output score. In other words, positive and negative predictions are equally likely for all groups. Equals *calibration by group*.

Barocas et al. (2023) note that *independence* is a prevalent criterion for normative, but also practical reasons: The normative reasoning is that "all groups have an equal claim to acceptance and resources should therefore be allocated proportionally" (p. 56). The practical reason is that independence is the easiest criterion to formalize and compute in the context of a machine learning project. The criterion not taking into account the true scores yields certain limitations. The authors take as an example a hiring scenario, in which members of different social groups are hired at the same rate but with varying levels of attention regarding their qualification. This leads to systematic performance differences, later on. The *separation* criterion addresses this limitation by taking merit into the equation. More precisely, it equalizes the rate of errors made by the system. Given a binary classifier, separation requires that the rate at which the classifier mistakenly predicts a positive outcome (false positive rate) should be the same for all groups. Moreover, the rate at which it mistakenly predicts a negative outcome (false negative rate) should be the same. In practice, one may decide to require only one of the two sub-criteria, depending on the decision-making scenario. In any case, a technical criticism of equalizing errors would be that it harms a system by actively forcing lower performance on certain instances. In response to this, Barocas et al. (2023) argue that one must consider the different real-world costs of misclassifications that are borne by different social groups. The general concern is that higher error rates mostly affect already marginalized and disadvantaged groups. Thus, equalizing errors facilitates a leveling of (dis-)advantage. Finally, *sufficiency* requires the likelihood of the true outcome to depend on the output score and there to be no difference across groups. Note that it has been shown that the three non-discrimination criteria, independence, separation, and sufficiency, are mutually exclusive (Chouldechova, 2017). Choosing the right metric is a matter of context and purpose of the bias analysis.

These constraints were originally defined for general machine learning approaches and, as such, are applicable to NLP systems for classification or prediction. Shah et al. (2020) conceptualize *predictive bias* in NLP as the divergence between the distribution of predicted output values and an *ideal distribution* that is theoretically defined for a certain target application. This divergence causes a human attribute to be represented in an unintended way. The authors distinguish between (1) *outcome disparity* (maps to independence): distributional differences appearing with regards to the general prediction outcome, e.g., when female instances are over-predicted and male instances under-predicted; and (2) *error disparity* (maps to separation): differences in the prediction error, e.g., when there are higher error rates for instances related to Black individuals and lower error rates for instances related to white individuals.

Embedding Metrics

This class of metrics is computed on the basis of dense vector representations (Gallegos et al., 2024). This type of metric was first introduced for word embeddings and later extended for sentence and contextualized embeddings. While originally a language model-centric method, it has also been applied to knowledge graphs embeddings, as shown in Chapter 4.

The first measure of this type was the *Word Embedding Association Test* (WEAT; Caliskan et al., 2017), which was inspired by a psychological bias test, called the *Implicit Association Test* (IAT; Greenwald et al., 1998). The logic behind this test is to measure how strongly two different identity-related words are associated to some neutral concepts, in comparison. Given two sensitive attributes denoted A_1, A_2 (one word per attribute, e.g., *man* versus *woman*) and neutral attributes denoted W_1, W_2 (e.g., *doctor* versus *nurse*), WEAT applies the following test statistic (Caliskan et al., 2017, p. 17):

$$f(A_1, A_2, W_1, W_2) = \sum_{a_1 \in A_1} s(a_1, W_1, W_2) - \sum_{a_2 \in A_2} s(a_2, W_1, W_2), \tag{2.8}$$

where similarity s is measured via:

$$s(a, W_1, W_2) = mean_{w_1 \in W_1} cos(a, w_1) - mean_{w_2 \in W_2} cos(a, w_2). \tag{2.9}$$

Finally, the magnitude of bias is measured via the effect size:

$$WEAT(A_1, A_2, W_1, W_2) = \frac{mean_{a_1 \in A_1} s(a_1, W_1, W_2) - mean_{a_2 \in A_2} s(a_2, W_1, W_2)}{std_{a \in A_1 \cup A_2 s(a, W_1, W_2)}}. \tag{2.10}$$

The association strength is measured a the cosine similarity between respective word vectors. The *Sentence Embedding Association Test* (SEAT; May et al., 2019) (applied in Chapter 5) generalizes this metric to contextualized embeddings of template sentences. For instance, A_1, A_2 would be “[Adam] is there.” and "[Jamel] is there.", whereas W_1, W_2 would be "There is [love]." and "There is [evil]." Guo and Caliskan (2021) create several sentences for each seed-attribute pair and compute distributions of effect sizes to account for random variance. Dolci et al. (2023) replace the sentence-level score by average world-level scores, where each word embedding is compared to a gender direction in the embedding space, which is determined via principal component analysis (PCA). Chapter 4 describes how Bourli and Pitoura (2020) applied the word association logic to the measurement of bias in a TransE knowledge-graph embedding.

Token Probability Metrics

This class of metrics utilizes template sentences, used as input to a language model. The bias metrics is then calculated on the basis of the token probabilities at dedicated spaces, reserved for sensitive attributes. Three influential metrics shall be introduced here (two of which were utilized in Chapter 5). *Crowdsourced Stereotype Pairs* (CrowS-Pairs; Nangia et al., 2020) comprises 1,508 example sentences, which come in pairs: one of them is a *stereotyping* sentence about a disadvantaged social group. The second is a copy of the same sentence, where the group identifier

is exchanged for one associated to an advantaged group. This constitutes an *anti-stereotyping* sentence. For example: "People who live in trailer parks are alcoholics." versus "People who live in mansions are alcoholics." The metric then computes a pseudo-likelihood of each token that appears in both sentences U, conditioned on the changed tokens M, to measure the effect of the sensitive attribute. To do this, each of the unmodified tokens is masked one-by-one and predicted given the remaining tokens in the sentence. Given a sentence S and model θ, the bias metric is defined as (Gallegos et al., 2024, p. 20):

$$CPS(S) = \sum_{u \in U} logP(u|U_{\setminus u}, M; \theta). \tag{2.11}$$

Nadeem et al. (2021) created two related bias measures, the *Context Association Test* (CAT) and the *Idealized CAT* (ICAT). Their dataset, labeled *StereoSet*, pairs a statement with three options: a stereotypical, an anti-stereotypical, and a unrelated option. There a two different formats: fill-in-the-gap (e.g., "Girls tend to be more ___ than boys" with response options: "soft" (stereotypical), "determined" (anti-stereotypical), and "fish" (unrelated)) versus sentences to follow the original statement (e.g., "He is an Arab from the Middle East." with response options: "He is probably a terrorist with bombs." (stereotypical), "He is a pacifist." (anti-stereotypical), "My dog wants a walk." (unrelated)). CAT follow a similar logic to CrowS-Pairs, but instead of conditioning the unchanged tokens on the changed tokens, it does the reverse and conditions the probability of the modified token on the unmodified tokens (Gallegos et al., 2024, p. 20):

$$CAT(S) = \frac{1}{|M|} \sum_{m \in M} logP(m|U; \theta). \tag{2.12}$$

ICAT utilizes the sentence-response pairs to calculate a composite of a *language modeling score* (lms) and a *stereotype score* (ss). The *ss* reflects the relative number of instances, where a stereotypical answer is preferred over the anti-stereotypical one (ideal score: 50). The *lms*, which is the ratio of instances, where the unrelated option is *not* preferred (ideal score: 100). The *lms* is then used to scale the *ss*, as follows (with an ideal ICAT score set to 100) (Nadeem et al., 2021):

$$ICAT = lms \cdot \frac{min(ss, 100 - ss)}{50}. \tag{2.13}$$

Generated Text Metrics

This class of metrics is specific to open-ended text outputs and builds on the idea that more general semantic aspects of these outputs can systematically differ across demographics. One of the first works to offer a dedicated measure for this purpose was Sheng et al. (2019). Their approach analyzes the general valence with which different groups are portrayed or *regarded* in generated texts. To measure regard, the authors trained a dedicated classifier similar to a sentiment classifier. It should be said, however, that regard and sentiment are similar but not identical concepts. For instance, the sentence "XYZ, known for his kindness, had died alone." is an example of positive *regard* attributed to XYZ, but negative overall sentiment of the sentence. Sheng et al. (2019) use different prompt templates (e.g.,

"XYZ was known for", "XYZ was regarded as"), where the placeholder is replaced by sensitive attributes (e.g., "The man" versus "The woman"). The language model-completed sentences are then classified and the ratio of positive, negative, and neutral portrayals compared between the different demographic groups. Dhamala et al. (2021) followed up on this work by proposing an improved prompting dataset named *Bias in Open-Ended Language Generation Dataset* (BOLD). The prompts are scraped from Wikipedia pages, such that they consist of a natural sequence of five words plus an attribute, such as a profession, a name, or religion. They combine different classifier-based metrics, such as regard, sentiment, and toxicity, as well as a newly proposed metric that utilizes psycholinguistic norms to identify conveyed affect (word-level matching to an emotion lexicon). Finally, they also propose a *gender polarity metric*, which identifies if a sentence (e.g., in response to a profession-related prompt) is more male- or female-leaning.

Intrinsic metrics, i.e., embedding and token probability metrics, are often uncorrelated to extrinsic metrics, i.e., metrics based on generated texts or downstream task behavior (Goldfarb-Tarrant et al., 2021). Since extrinsic metrics are more directly user-facing, they are often considered more indicative of potential real-world harms (Lum et al., 2025; Delobelle et al., 2022).

2.5.4 Algorithmic Bias Mitigation

Bias mitigation techniques can be categorized according to the model lifecycle stage, in which they apply: as part of the data pre-processing, during the training, or as part of the post-processing (Lin et al., 2024; Gallegos et al., 2024).

Strategies applying during the pre-processing include, for example, *counterfactual data augmentation* (CDA; Zmigrod et al., 2019) or *counterfactual data substitution* (CDS; Maudslay et al., 2019) where gendered words are balanced throughout the training dataset, either by creating a mirrored copy or by flipping each gendered text with 0.5 probability. Newer approaches have moved away from replacing individual terms, as this can corrupt the grammatical correctness of a sentence. Sattigeri et al. (2022) utilize an influence function to approximate which training examples strongly influence a classifier's performance with regards to a bias metric (e.g., demographic parity). They then adjust the pre-trained embedding such that the influence of these data points are reduced. Utama et al. (2020) identify biased instances using a separate classifier, trained on a small data subset, and then reduces their weighting in the actual model finetuning.

In-training methods incorporate modifications to the optimization loss, such as *contrastive learning*. Li et al. (2023) combine CDA with contrastive learning to push representations of different sensitive attributes (as encoded in the counterfactual data pairs) closer to each other and remove biased associations. Others use a version of adversarial training (Zhang et al., 2018; Han et al., 2021).

Finally, some methods are applied after the completed training. Bolukbasi et al. (2016) developed a method for the mitigation of gender bias in word embeddings. The authors identify a gender subspace via PCA applied to the difference vectors of gendered word pairs, onto which all word embeddings are then projected and, thus, re-embedded. This method was also modified to work with contextualized embeddings (Barikeri et al., 2021) and knowledge-graph embeddings (Bourli and Pitoura, 2020). The method proposed by Sheng et al. (2020) fares without embed-

ding adjustments. They utilize so-called *trigger tokens*, which are appended to the input prompt and fit through an iterative, gradient-guided search to minimize output bias as measured by a regard classifier.

2.5.5 Limitations of Bias Measurement and Mitigation

To conclude this section, it shall be noted that there are limitations to the measurement and mitigation of algorithmic bias, beyond the mere technicalities. There are many different ideas of what constitutes a *fair* algorithmic system. This is reflected in the many, conflicting metrics that exist in the field. Moreover, bias metrics are always reductive. Most bias metrics are limited to measuring (binary) gender bias (Gallegos et al., 2024). And even the underlying theories of gender and gender bias in these works are highly reductive, conflating of sex and gender (Kraft et al., 2022; Devinney et al., 2022), and not contextualized (Bhatt et al., 2022). The limitations of bias measurement directly reflect in the limitations of attempted mitigation. That is, mathematically, we can only remove what we have succeeded to measure (Gonen and Goldberg, 2019; Kraft, 2021). And this brings us back to the limitations of our metrics.

2.6 Conclusion

This Chapter has presented an in-depth introduction to the most essential technical concepts that are utilized or referenced throughout this dissertation. It defined the notion of an artificial neural network and the concept of learning, in this context. Then, important principles of modern language modeling were characterized and starting at their technical origins. The knowledge graph is the second form of representation that is at the core of this work. Thus, foundational definitions and concepts were introduced, as well as Wikidata; a particularly relevant instance, in this thesis. The practice of knowledge-enhanced language modeling, which marries language models and knowledge graphs was also briefly explained. Finally, the Chapter concluded with an in-depth introduction to algorithmic bias, its causes, measurement, mitigation techniques. While this chapter was focused on important technical concepts, the following chapter will be addressing philosophical concepts related to knowledge, science, ethics, and lastly situated accounts of AI.

3
Philosophical Background

Contents

3.1 Introduction

As introduced in Chapter 1, this thesis is concerned with the *epistemic and ethical goodness* of knowledge technologies, i.e., technologies that represent and transport knowledge, as well as influence the ways in which humans access and generate knowledge. To facilitate this evaluation, this thesis draws from existing concepts and discourses in philosophy. In fact, there is a whole field in philosophy devoted to characterizing what knowledge is, namely *epistemology*, which studies the "nature, origin, and limits of human knowledge" (Martinich and Stroll, 2025). It is beyond the scope of this thesis to deliver an exhaustive definition of "knowledge",

so for now let us assume that it is some type of "cognitive success", determined by a number of success conditions (Steup and Neta, 2024). At times we are interested in knowledge as the relationship between a subject and a thing in the world, i.e., the *knowing* of something. At other times, we are more interested in knowledge as some type of content in itself. In both cases, certain conditions must sufficiently be met; for someone to know something, but also for some piece of content to count as knowledge. What these conditions are and how they come about are some of the core considerations discussed in this Chapter. This Chapter will flesh out an account of knowledge that provides the basis for the analyses presented in this thesis, which is based upon insights from social and feminist epistemology, in particular.

To this end, Section 3.2 gives a brief introduction to the main characteristics and ideas of social epistemology, contrasting it to accounts of traditional epistemology. The change from the traditional to the socialized view of knowledge represents an important paradigmatic shift that was strongly influenced by feminist epistemology. The latter may, in fact, be understood as a subclass of social epistemology (Grasswick, 2018). Section 3.3, will introduce different feminist epistemologies in more detail and highlight the works of some of the noteworthy thinkers who have coined notions that provide an important framing to the research presented and discussed in this thesis. Section 3.5 then bridges the epistemological accounts to the field of AI, by revisiting two works dating back to the era of expert systems.

3.2 From "Traditional" to Social Epistemology

A common conception in traditional Western philosophy equates knowledge with *propositional knowledge*, i.e., knowledge of a proposition such as "S knows that p" (Ichikawa and Steup, 2024). Here, S is the knowing subject and p is the object of knowledge. Thus conceived, a standard definition of knowledge is that S knows p if and only if:

1. S believes that p,
2. p is true,
3. S' belief in p is justified.

(Steup and Neta, 2024). According to this definition, knowledge is understood to be *justified true belief*. That is to say that if S does not even believe in a fact p, they cannot possibly come to know it. Secondly, a false proposition p cannot be a fact. And, thirdly, if p is indeed true, but the reason for believing p is mere luck or some other inappropriate reason, this is not considered to lead to *knowing*. The variable of *justification* turns out to be an important point of debate. What must a justification be in order to be appropriate, to be better than luck?

It shall be added here, that there are also variants of this account to accommodate certain logical issues of the conception of knowledge as *justified true belief*. Such issues were first demonstrated by Edmund Gettier who showed that there are cases–so called *Gettier cases*–that fulfill all three conditions without actually being

knowledge (Gettier, 1963). One famous example is the barn-facade case, where someone is going through a street with several buildings that appear to be barns. Most of them are just facades and not actual barns. But the person, by chance, happens to be standing in front of the only real barn and believes to be standing in front of a barn. The person's belief is true and justified by their prior experience of what a barn usually looks like. Yet, in this situation, the actual reason why this belief is true is luck. Nevertheless, belief, truth, and justification remain important problems in epistemology that are each on their own theorized in different ways.

Traditional, Western epistemology is mostly interested in the individual relationship between an individual knower S and an object of knowledge p, conceptualizing knowing as matter of cognition and reasoning only. This view was largely influenced by René Descartes, who was convinced that knowing is achieved through one's own reasoning (Descartes, 1637, as cited in O'Connor et al., 2024). Another influence was John Locke, who proposed that justification can only be found individually, in one's own sensory experience and conclusions drawn from them, as opposed to the testimony of others (Locke, 1690, as cited in O'Connor et al., 2024; Longino, 2002). In the second half of the 20th century, however, a new branch of epistemology emerged that challenges this individualistic tradition. Social epistemologists argue that knowledge, how we get to knowledge, and what we consider to count as knowledge, are highly influenced by social factors. And one such factor is indeed testimony, seen as a transaction of information between a speaker (or writer) and an audience (O'Connor et al., 2024), where, if an audience accepts the speaker's testimony, they obtain testimony-based belief. While traditional epistemologists–approaching the problem from a normative stance–reject this as an appropriate source of knowledge, social epistemologists–taking on an empirical stance–maintain that the significance of testimony in knowledge production is evident in empirical observations. Even if the status of knowledge achieved through testimony continues to be widely debated, social epistemology considers it an important subject of investigation.

Beyond philosophy, other fields of research have also investigated the sociality of knowledge practices, most notably ethnomethodological, sociological, and historical studies. The field of STS emerged in the 1960s and 70s and has ever since been concerned with the social study of science, technology, and their interactions with society; subsuming diverse interdisciplinary perspectives and methods (Hackett et al., 2007). An important line of research in STS was linked to the *Strong Program*, in the sociology of scientific knowledge. Associated scholars argued that a theory or hypothesis is accepted *or* rejected not on the basis of their objective truthfulness, but on the basis of social interests (Barnes and Bloor, 1982, as cited in Longino, 2002; O'Connor et al., 2024). Hence, different beliefs are expected to exist in different context, depending on the given circumstances and interests. Also highly influential to STS and the emergence of social epistemology were the laboratory studies conducted by scholars like Bruno Latour (Latour, 1983) and Karin Knorr-Cetina (Knorr-Cetina, 1983). Their respective research engaged in observations of the interactions within and between research labs, and analyzed the efforts involved in exporting research work and findings from the laboratory to the non-laboratory world. Knorr-Cetina (1983) for example, conjectured that, in practice, knowledge processes are *interaction*-based, because justification happens by reproducing and incorporating theories into successive

works. Only if newer works continuously *select* and build upon some theory, does it succeed in becoming accepted scientific knowledge. Researchers' awareness of these processes, in turn, influence their research design and reporting.

The Strong Program and the laboratory studies also influenced philosophers of science, such as Longino (2002), who made a proposal to consolidate the empiricists' take with a normative philosophical understanding of knowledge and knowledge production. Longino's account is explicated in Section 3.3.2 and is characterized by a feminist approach to epistemology, which is also referred to as *feminist empiricism* (Intemann, 2010; Harding, 2013). As a field of research, feminist epistemology is concerned with the significance of social relations of gender or identity and social location, more generally, in the context of knowledge production. It emerged in the early 1980s as a result of the second-wave feminist movement and predates the establishment of social epistemology (late 1980s/early 1990s) (Grasswick, 2018). The following section explains the core idea of feminist epistemology in more detail and characterizes some important works.

3.3 Feminist Epistemology: Situatedness and Objectivity

Around the time of the second wave of feminism, in the 1970s, feminists in philosophy and other scientific fields began to document and question sexist and patriarchal structures and practices (Grasswick, 2018). Feminist thinkers started to question the established scientific assumptions and methods which were centered around and informed by the beliefs, norms, interests, attitudes, and values of men. They started to question, in a word, the *andocentricity* of science. This gave rise to questions regarding the role of gender in established conceptualizations of knowledge and the processes that lead to knowledge.

One of the core concepts feminist epistemology is concerned with is that of *situated knowledge* (Anderson, 2024). Knowers are understood to relate differently to what is known and to other knowers, i.e., what someone knows and how they know it depends on their situation, lived and embodied experience, and perspective. The social situation or location is constituted by a variety of factors such as their gender, race/ethnicity, class, social relations, and roles. An individual's social role gives rise to different norms, virtues, habits, interests, skills. Gender, for example, has been considered a significant variable that defines a person's prescribed but also self-ascribed role, norms, virtues, behaviors, etc.

In fact, feminist epistemology has primarily been concerned with gender as a defining mode; going as far as understanding it as a factor that bifurcates society (Grasswick, 2018). However, due to contributions such as "Black Feminist Thought" by Collins (1990) which draw attention to the unique experience of Black women, the field has opened up to a more intersectional understanding of gender, that is, gender not being an isolated aspect of identity. That is not to say that tensions between different feminisms do not continue to exist; such as, for instance, West-centric and decolonial feminisms (Vergès, 2019).

Another core tension relates to the question of *objectivity* and how it can be achieved despite the situatedness and value-ladenness of all knowing. Different

feminist epistemologies propose different conceptualizations of it, which also entail different implications for the role of diversity. For example, while standpoint theorists insist that certain marginalized experiences can give rise to epistemic privilege and lead to *stronger objectivity*, feminist empiricists assert that diversity of values and ideas are desirable in the sense that they can help cancel out individual biases (Intemann, 2010).

The following sections focus on the feminist epistemologies that are mostly referenced throughout this thesis. This includes the feminist standpoint theory as proposed by Sandra Harding, Patricia Hill Collins, and Donna Haraway, as well as the rational-social account of science proposed by Helen Longino. All of these works reason about the significance of diversity and situatedness within science and the pursuit of objectivity, and produce normative implications.

3.3.1 Harding, Collins & Haraway: Situated Knowledge and Feminist Standpoints

Feminist standpoint theory is a term that was coined by Harding (1986) to subsume feminist research projects across different disciplines, such as sociology, philosophy, and history of science, that have helped develop understandings of how gender relations–or the conception thereof from an androcentric view–structure science. Based on these works, standpoint theory assumes that how our societies are structured has epistemological consequences. That is, those that dominate or "rule" will come to different conclusions than those that "are ruled" (Harding, 2007). Harding argues that being part of the community that dominates and (intentionally or unintentionally) benefits from the conceptions held by this community makes it less likely to question these conceptions. For example, in a society "stratified by race and gender [that] lacks powerful critiques of this stratification" (p. 49), those who benefit from racism and sexism are unlikely to be the ones who identify it (Harding, 2007). In fact, feminist standpoint theory builds on the Hegelian idea that oppression should be studied from the perspective of the oppressed, and the subsequent Marxist take that understanding the world shaped by capital is best approached by starting from the standpoint of the proletariat (Harding, 1986; Bowell, 2025). The reason for that being that inhabiting the standpoint of the oppressed creates the opportunity to come to unique realizations about oneself and the power structures one is affected by. "Black feminists have questioned not only what has been said about Black women, but the credibility and the intentions of those possessing the power to define" (Collins, 1990, p. S17). Thus, the oppressed possess epistemic privilege and the opportunity to reevaluate and correct that which was previously out of sight or suppressed.

In short, the theory assumes (1) that knowledge is (socially) situated, (2) that marginalized individuals are more aware and equipped to criticize and deal with criticism in knowledge production, than individuals in the dominant group, and (3) that science should start at the perspectives of the marginalized (Bowell, 2025).

Situated Knowledges

As discussed before, the traditional account of objectivity appears to be built on the assumption that its source is pure, unembodied rationality. So consequently,

owning a body (that is located in society and history) is seen as introducing bias and unfit for objective knowledge production. Haraway (1988) argues that this traditional account is a *masculinist* account of knowledge that lays the grounds for the systematic exclusion and derogation of marginalized views. At the beginning of her essay she distinguishes between a "they", referring to the "unmarked positions of Man and White" (p. 581), and a "we", designating "embodied others, who are not allowed *not* to have a body, a finite point of view" (p. 575). With this, she describes the status quo in science which she continues to identify as built on false and problematic assumptions. The purpose of her essay is to show that, in fact, *every* view is embodied and non-neutral.

Masculinist scientists use all sorts of maneuvers like "objective" analytical methods or complex technology to obfuscate the origins of data and assumptions. This creates the illusion that the world is *viewed from nowhere* (Nagel, 1989), unembodied, unbiased. Haraway calls this the *god trick*. Knowledge processes are subject to the god trick whenever the originator's positionality and locatedness is denied and the knowledge content presented as non-situated (Hoppe, 2022). It renders knowledge claims "unlocatable, and so irresponsible" (Haraway, 1988 p. 583). Being able to apply the trick, to see without being seen, requires a level of power that is almost exclusive to the "unmarked positions of Man and White" (p. 581).

However, in reality, every knower is bound to a body, and thus, everybody has limited vision. She compares the eyes of humans and animals with cameras: All "eyes"–whether organic or non-organic–are active devices for capturing certain aspects of the world in certain ways, within the range of possibilities provided by the body or machine. There is no such thing as disembodied or unmediated vision. As such, every perspective is *partial*, *located*, and *situated*. The social situation of an individual–constituted by gender, race, class, ability, education, etc.–determines, what they know and what they *can* know, also in the sense of what they are *permitted* to know (Bowell, 2025). In her argument, the knowing subject is *in itself* partial, fractured, and changing (Haraway, 1988):

> "The knowing self is partial in all its guises, never finished, whole, simply there and original; it is always constructed and stitched together imperfectly, and *therefore* able to join with another, to see together without claiming to be another" (p. 586).

It is based on this idea of the exclusively partial and situated nature of knowledge that Haraway speaks of *knowledges* in plural form. There are multiple knowledges but they can be joined in all sorts of expected *and* unexpected ways. Ultimately, partial connections open up new possibilities in scientific discourse. But for this to happen, established regimes of domination must be deconstructed. Her agenda is, thus, not only of epistemological, but also of political and ethical nature.

Achieving a Standpoint

A standpoint is defined as a collective identity, grounded not merely in commonalities regarding gender, race, class, or the like; it goes beyond perspective

and does not exist automatically. Harding understands a standpoint to be a collective accomplishment, formed through a shared experience of oppression or political struggle (Harding, 2007). With regards to the feminist standpoint in particular, Harding (1986) claims:

> "Feminism and the women's movement provide the theory and motivation for inquiry and political struggle that can transform the perspective of women into a "standpoint"–a morally and scientifically preferable grounding for our interpretations and explanations of nature and social life." (p. 26).

In providing a basis for better understanding our natural and social surroundings, a standpoint is seen as an achievement that is liberating. In her article "Learning from the Outsider Within", Collins (1990) characterizes the *Black feminist standpoint*, which she conceives to be a defining factor of Black feminist thought. She describes how Black feminists (in the United States) undergo processes of self-definition against the "externally defined, stereotypical images of Afro-American womanhood" (p. S16). In redrawing those images based on their own lived experiences, Black feminists come to see that the inaccuracy of the original image was not arbitrary but, instead, the result of broader dynamics of power and oppression.

Epistemic Advantage

Harding critiques the traditional Western, "prefeminist" science for being solely focused on the processes of justification and not on what happens prior to that. She argues that the "context of discovery"–such as the grounds on which the objects of investigation, the research questions, and methodologies are decided–is in itself not value-free. However, in requiring freedom of values and subjectivity, and in leaving the "context of discovery" unquestioned, science is also missing an opportunity to become *better* (Harding, 2007).

While Harding (2007) does not claim that one can be completely free of the dominant paradigm or, in her words, the "historical moment", she says that only "a degree of freedom" from it is enough to see things differently and to start questioning it. Again, Collins provides a good example for this: In her work she investigates the experience of Black, female researchers in sociology. She describes them as "outsiders within", as they are part of the community, but also not. They know what the insiders know, but their "otherness" allows Black women in sociology to also see certain relations more easily than insiders, such as the "credibility and the intentions of those possessing the power to define" (Collins, 1990, p. S17).

And thus, starting from the experience of the oppressed, allows for a "stronger kind of reflexivity". It facilitates critique and "methodological control" over the "context of discovery", i.e., more thorough evaluations of the very basis our research practices stand on; the things we decide to research, why, and how. Besides it being seen as a more rigorous call for objectivity, it is also seen as an opportunity to empower those who are oppressed and to "produce knowledge that can be *for* marginalized people [...] rather than *for* the use only of dominant groups

in their projects of administering and managing the lives of the marginalized people" (Harding, 2013, p. 56).

Values in Science

She argues that science is *de facto* not value-free and adds that not all values are a hindrance to good science. In particular, "democracy-advancing values" like feminism are considered helpful in pursuing better science and *stronger objectivity*. Harding promotes the vision of a science that is deeply political and commits to "antiauthoritarian, antielitist, participatory, and emancipatory values" (Harding, 1986 p. 27). The vision that Harding promotes here is the creation of a *successor science* that facilitates a *stronger objectivity* than traditional science.

Also Haraway (1988) praises science, in general, as "utopian and visionary", and controlled and critical scientific knowledge production as important and necessary. She calls for a kind of science that is open to partiality, "interpretation, translation, stuttering". Haraway emphasizes the notion of rationality as more important than objectivity and claims that objectivity is found in *positioned rationality*: "Rational knowledge is a process of ongoing critical interpretation among "fields" of interpreters and decoders. Rational knowledge is power-sensitive conversation" (p. 590, emphasis from original quote).

Note on Standpoint Epistemology and Relativism

In her essay "Situated Knowledges", Haraway (1988) critiques the conceptualization of a standpoint and the relativism she saw proposed in Harding (1986). In her response to it, Haraway states that she sympathizes with the general assumption that the perspectives of the subjugated hold epistemic value. Due to their experience of repression they might be more able to see through the god trick. But the notion of a standpoint as discussed by Harding (1986) presumes a level of unity that is not compatible with Haraway's concept of a fractured, ever-changing knowing self, where there is no coherent and lasting identity-based way of seeing. Moreover, due to this instability, we cannot simply predict what someone else is seeing. Meanwhile, she accuses standpoint theory of being relativist, which she classifies as performing yet another god trick. She states that "relativism is a way of being nowhere while claiming to be everywhere equally" (p. 584). It claims that there are as many truths as there are perspectives, rendering objectivity an impossibility.

Harding (2013) provides a direct response to the accuse of promoting the god trick and relativism. Firstly, she emphasizes that standpoint theory cannot be performing the god trick, as it is firm on the social situatedness of all knowledge production and even considers it a scientific resource. Secondly, she claims that standpoint theory is not relativist as "it argues against the idea that all social situations provide equally useful resources for learning about the world and against the idea that they all set equally strong limits on knowledge" (p. 61). She states that "*sociological* relativism permits us to acknowledge that different people hold different beliefs" (p. 61) but emphasizes that *judgmental* relativism is anti-scientific.

3.3.2 Longino: Knowledge is Rational *and* Social

In her book "The Fate of Knowledge", Longino (2002) contrasts the accounts of the sociologists of knowledge and the traditional philosopher that were briefly introduced in Section 3.2. She discusses how both in their own ways convey a dichotomy of social and rational conceptualizations of knowledge. As mentioned before, early studies in the field of social epistemology, like the Strong Program or the laboratory studies by Latour and Knorr-Cetina conceptualize knowledge as a product of solely social processes. According to Longino, their accounts imply that scientific processes are not rational and not necessarily evidence-based. "They agree that what they identify as the philosopher's approach–one characterized by normative concerns about the nature of knowledge–misrepresents the scientific process" (Longino, 2002 p. 37). Likewise, philosophers reject the idea that knowledge and aspects of knowing, such as justification, can ever be appropriate when based on social factors. Longino deems this dichotomous understanding misguided and proposes a "third way" which consolidates both views and conceptualizes knowledge as both rational and social. Her elaborations are mostly done with reference to the case of scientific knowledge. However, she clearly states that her account transfers to knowledge in general.

The Rational-Social Dichotomy and the "Third Way"

Throughout the book, Longino refers to the traditional philosophers simply as *philosophers* and to the sociologists as *empirical investigators*. In the following, the differences between the philosophers' "rational" account and the empirical investigators' "social" account are characterized in more detail. According to Longino (2002), we can look at knowledge in three different senses:

1. Knowledge as *knowledge production*: the practices and processes through which knowledge–or what is considered knowledge–is produced,

2. knowledge as *knowing*: the state of a knowing subject with respect to some object(s),

3. and knowledge as *content*: that which is known and which is the outcome of the knowledge-productive practices/processes.

(1) In regards to *knowledge production*, the philosopher is interested in the *individualistic* relationship between a knower S and a fact p, and therein "the contrast between processes of belief acquisition that do and those that do not rationally justify belief" (Longino, 2002 p. 78). The empirical investigator is less interested in this individualistic relationship and, instead, concerned with the processes of "generating accounts and representations and having those accepted by the community" (Longino, 2002, p. 79). (2) Knowledge as *knowing* refers to the state of a person or several persons relative to an object or several objects. To the philosopher *knowing* is *justified true belief*. The three terms (see Section 3.2) are seen as one single definition of *knowing*. Empirical investigators, on the other hand, differentiate between the three terms. While they are much less interested in the truth term, they see tension in the justification term. The argument being that for S to believe p, p must also be accepted in the community

of *S*. Moreover, the processes or arguments based on which *S* comes to believe *p* must be accepted in their community. (3) Finally, knowledge as *content* is that which is known, that which is transported in books and journals. Again, Longino claims that philosophers are highly concerned with the truth term and, thus, the distinction between mere belief and belief of what is really true. Knowledge is seen in a *monistic* way, i.e., there being only one correct theory. In that sense, knowledge as content is conceptualized as a "subset of truths which is known" (p. 84). In contrast, the empirical investigator sees knowledge in a *relativistic* way, as whatever is designated as such by a given community. Again, their interest lies not in an actual objective truth and different theories may hold true in different contexts. Longino asserts that the rational-social dichotomy is reflected in three binaries – *individualism versus nonindividualism*, *monism versus nonmonism*, and *relativism versus nonrelativism*. The "socializers" claim individualism, nonmonism, and relativism and the "rationalists" claim nonindividualism, monism, and nonrelativism.

Longino's main effort consists in developing a non-dichotomous conceptualization of knowledge: (1) She argues for a *nonindividualistic* account, claiming that cognitive agents and subjects are *interdependent* in their generation and justification of knowledge content. (2) She proposes a *nonrelativist* account, in a *contextualist* sense. That is, justification is not "arbitrary nor subjective, but is dependent on rules and procedures immanent in the context of inquiry" (p. 92). (3) And finally, she argues for *nonmonism*, in a *pluralistic* sense. She rejects the view that there is only one world and, hence, only a certain set of compatible theories that can possibly be true at the same time.

Knowledge is Interactive

She claims that cognition, and such the cognitive processes in knowledge productive activities, shall be understood as having a social dimension. Cognitive processes, such as observation and reasoning (to generate new ideas or to justify ideas on the basis of evidence), are inherently *interactive*. In fact, Longino states that what she means by "social" is effectively "interactive" (p. 148). Experiments are, for instance, conducted in a way that allows others to (theoretically) repeat and confirm the same observations. And for a piece of evidence to be perceived as justifying a theory it must meet community-set standards of inquiry and conjecture. Inherent to these procedures is the need for multiple, varied points of view. Only if a claim holds from different views can it be assumed to hold in general and not only within one's own idiosyncratic view. As such, the social is a matter of validation in knowledge production.

Longino adapts from the sociologist's view the idea that a knower is *located* historically, geographically, and socially; and in doing so, she rejects the assumption that a subject may create knowledge in context-independent, value-neutral, purely methodological ways. However, when she speaks of epistemic communities, she does not refer to communities built on shared identity or values. Instead, a community is defined by shared "standards to which community members appeal in critical discursive interactions" (Longino, 2002p. 148), such as evidential or methodological standards. In fact, she emphasizes that while there is consensus

regarding the appropriate measures of evaluating claims, members of a community can and should still have diverse beliefs.

Plurality as an Epistemic Resource

Longino (2002) argues that plurality of ("cultural and ideological") perspectives is an epistemic resource and a necessary feature to ensure successful epistemic processes. Firstly, it ensures effective criticism. Secondly, plurality gives rise to "a variety of pathways to a similar end" (p. 200), namely the collectively shared pursuit of understanding the natural world we live in. While traditional philosophy has paid much attention to the "*S* knows that *p*" conceptualization of knowledge, i.e., propositional knowledge, Longino argues that this notion is oversimplifying. In practice, scientists are interested in a multitude of complex causal relationships, and less in knowing *that* something is but more in knowing *how* and *why*.

Plurality gives rise to a variety of scientific commitments that take on different aspects of a complex process. The ways in which aspects of the world are then modeled or represented in order to reason and conjecture about them, depends on the given context and commitment:

> "["World"] can also mean the collection of aspects of the world that is salient to those approaching it with a given set of assumptions and strategies for acquiring knowledge not to mentioned a given sensory and cognitive apparatus. In this sense there are many worlds." (p. 94).

And there lies an opportunity in that, which shall be leveraged. Arguing that the natural world is too complex to be captured in one unified theory, she promotes plurality (the existence of "many worlds") as a chance to capture many different aspects at once, with different commitments. Epistemologies are *localized* and knowledge shall be seen as "provisional, partial, and context-dependent" (p. 143).

Evaluating the Success of Content

Knowledge in the third sense–knowledge as *content*–signifies content that is *successful* in qualifying as knowledge. As noted in the previous paragraphs, Longino posits that community standards determine this success, and important aspects therein are the locatednes of subjects and the diversity of views. Notwithstanding, she clearly distances her conceptions from relativism. Central to Longino's theory are *conformation* and *epistemic acceptability*. Both of these notions consolidate the more traditional philosophical idea of evidential norms, i.e., the requirement that empirical evidence must support the truth of a claim, and the sociologists' idea of community norms, i.e., the requirement that the community must approve of a claim.

A content, e.g., in the form of a theory or claim, may be understood as a representation that relates to some object. *Conformation* describes this relation. It is given if the representation preserves certain properties or relative dimensions of the object in a way that facilitates useful interaction with the object; in other words, if it "sufficiently enable[s] members of [the community] to carry out their projects with respect to that/those object(s)" (Longino, 2002p. 136). Conformation is community-specified through representational conventions and the respective

project-specific purposes or needs. Due to the latter, content with no attachment to or measurable consequence for some real-world object, cannot be successful. As opposed to "false" versus "true", the notion of conformism allows for degrees. A content can conform to a certain degree and different contents or representations can conform to different degrees or in different ways.

Indeed, Longino rejects "truth" or "false" as appropriate measures to evaluate the success of scientific content. She takes an example by Cartwright (1983), who pointed out that the laws of physics are not in a strict sense true. They hold true only *ceteris paribus*, when a wide range of idealized conditions are met. A second example she refers to is by Hacking (1992), who discusses that the statement "The population of Paris in 1800 is *N*" can only hold true given an established praxis of population counting, a shared definition of population or residency. And it can only hold true within a very short moment in time. Nevertheless, both cases provide representations that help reason and inquire about their respective matter. Physics laws might be inaccurate and simplifying, but in their inaccuracy and simplicity, they allow to reason abstractly about a set of related phenomena. In this way, Longino argues, theories are comparable to maps: depending on the purpose for which the map be used, it represents terrains in different ways; at different levels of scale, in different colors, emphasizing certain elements of the real surface of the earth, omitting others. While maps must not be a copy of each geographic detail, they must still reflect the real-world in a way that allows to draw conclusions about it. For instance, relative distances between elements of interest must correspond to real-world distances in some way. Conformism is seen as a more suitable notion than "true" or "false", because it can describe more complex contents (than, e.g., simple propositional knowledge). Such contents can comprise certain elements that are in strict terms "true", elements that are in strict terms "false", as well as certain conventional elements (e.g., how the notion of a population is operationalized).

What distinguishes science from what Longino calls "wrongheaded" beliefs such as creationism, is *epistemic acceptability*. It constitutes an empiricist element in her philosophy and combines the idea of socialized cognition with the need for empirical evidence for justification. A theory must represent that which it is supposed to represent such that it facilitates interaction with and conjecture about it in a useful capacity. And it can only be useful, if it corresponds to the real world in some way. She claims that a community develops its standards for the validation of content embedded in a physical environment, which its members are interested to represent accurately. And at any given time, these standards are themselves subject to scrutiny. So, standards that are in conversation with the physical world, and produce (or try to produce) accounts that are empirical plausible become distinguishable from those accounts that do not even attempt to correspond with empirical inquiry, such as creationism. In order for this principle to function well, however, empirical communities must actively seek and expose itself to *effective criticism*.

Normative Account of Social Knowledge

Building on her account of the sociality of knowledge, Longino (2002) proposes a normative theory of (social) knowledge. She defines a range of criteria that

ensure effective discursive interactions and, consequently, knowledge-productive communities:

1. *Venues*: There must be public venues where research, including its underlying assumptions, methods, etc., is criticized and discussed. The criticism of research must receive the same value as the original research itself, because it contributes to the (re-)evaluation of theories and "to better appreciation of their grounds and of their consequences".

2. *Uptake*: The criticism must be taken up into a critical discourse within the community and it must be followed by adjustments to beliefs and theories made over time. This way criticism becomes constructive and justificatory.

3. *Public standards*: There must be publicly shared standards by which theories, hypothesis and the empirical study procedures ("observational practices") are evaluated. Based on these standards criticism can be formulated such that they pertain to the goals of the community and are nonarbitrary. These standards themselves can be subject to scrutiny and change.

4. *Tempered equality*: Shared standards must be based on a consensus, derived from critical discourse between a diversity of perspectives, and thereby not influenced by social position or economic power.

The last condition, *tempered equality*, is emphasized in the book. Longino (2002) argues that "the exclusion of women and members of certain racial minorities from scientific education and the scientific professions constitute not only a social injustice but a cognitive failing" (p. 132). She refers to the history of racist theories held up in the sciences, which remained unquestioned for a long time because racially minoritized individuals were excluded from discursive practices. Hence, the scientific community must actively pursue members from under- or unrepresented groups, and engage in serious dialogue or *uptake* of respective criticisms.

3.3.3 Take-Aways

This section presented different feminist epistemologies with several commonalities, but also certain distinctions. We shall firstly recapitulate the commonalities: A central theme is the juxtaposition of a masculinist tradition in philosophy that is characterized by a rationalistic, de-situated, and disembodied pursuit of truth, versus an alternative feminist view that draws attention to the situatedness, sociality, power, and dominance within knowledge processes. All accounts have in common the understanding that knowledge is inherently social and situated, but that objectivity is nevertheless a possibility and something to strive for. But there is disagreement as to *how* it can be achieved.

Haraway and Longino both interpret standpoint theory as being relativist and make it a point to distance themselves from it. In her commentary on Harding's take on the feminist standpoint theory, Longino (1993) states:

> "If we abandon the idea that knowledge is one, and when achieved, absolute, if we assume the location of knowledge in sociohistorical

> contexts and become pluralists, we are still faced with the ancient problem of distinguishing knowledge from opinion and what the distinction amounts to" (p. 212).

And her solution to enabling objectivity despite the sociality of knowledge production is empiricism. While community consensus determines our knowledge-productive methods and standards for evaluation, these still need to be empirically acceptable.

As mentioned earlier, Harding (2013) rejects the classification of feminist standpoints theory as relativism, because the theory does not expect all social situations to be equally likely of producing valuable claims. Instead it specifically assigns the potential for an epistemic advantage to the "outsider within" (Collins, 1990). Standpoint theory goes as far as claiming that feminist research is better research, which can lead to stronger objectivity, because it actively promotes reflexivity.

In an analysis of feminist empiricism and feminist standpoint theory 25 years after their inception, Intemann (2010) claims that standpoint theory's emphasis of situated, embodied experience can actually be read as empiricism. She claims that both theories are empiricist, contextualist (justification depends on local set of assumptions, methods, and values), and normative (there is no value-free science, and certain commitments are better in promoting objectivity). Intemann identifies two central differences: Firstly, both theories agree that diversity is important for objectivity, but differ in the *kind* of diversity they consider beneficial to science. Longino (2002) requires *diversity of values and interests* within scientific communities, criticism from as many different views to cancel out individual biases. Harding (2013) require *diversity of social position*, "because social positions track power relations in ways that are epistemically significant" (Intemann, 2010, p. 790). Intemann asserts that this latter kind of diversity, the diversity of social position is actually more likely of yielding different empirical evidence, on which justification and criticism may be based. In that sense, it seems more consistent with feminist empiricism than the demand for diversity of values and interest. The second difference, according to Intemann, relates to the role of values: standpoint theorists explicitly favor "antiauthoritarian, antielitist, participatory, and emancipatory values" (Harding, 1986 p. 27) over opposing values. The argument against sexist and androcentric value judgments is that they are considered less supported or even *unjustified* (Intemann, 2010). Feminist empiricism, however, appears not to be concerned with the actual content of values, as long as a sufficient amount of different values are represented to neutralize idiosyncrasies. Intemann (2010) argues that this is in a way inconsistent with feminist empiricism's inherent political and ethical stance. She suggests to bridge these gaps between the two theories by accepting the standpoint theory take on these remaining two differences: to favor diversity of social position over diversity of values and interests, as well as to endorse the idea that certain political and ethical commitments are more beneficial for science than others. This would give rise to what she calls *feminist standpoint empiricism*.

Indeed, feminist standpoint empiricism describes well the kind of epistemology on which this thesis analyzes knowledge technology and the processes involved

in its creation and evaluation in AI. Here is a list of the core assumptions, I take away from the preceding section:

1. Knowledge is social, situated, partial, and embodied: gender, race/ethnicity, class, social relations, and roles influence what we can know.
2. Knowledge is contextual: justificatory processes are embedded in a locally shared set of assumptions, research goals, and methodologies.
3. Objectivity requires empirical evidence: justification needs to be anchored in empirical evidence and embodied experience.
4. Objectivity requires diversity of social position: different experience must contribute to knowledge-productive processes, and the experience of marginalized groups are epistemically particularly valuable.
5. Objectivity benefits from democracy-advancing values: values that drive diversity and are power-sensitive are beneficial to science.

3.4 Feminist Epistemology: Knowledge, Injustice, and Oppression

We have now seen that knowledge can be understood as social and situated, and in which ways the pursuit of objectivity can be conceptualized with this in mind. Now, the focus is shifted more towards the experience of non-dominant as opposed to dominant groups with regards to the sharing and gaining of knowledge. Specifically, the notion of epistemic injustice proposed by Miranda Fricker and expansions and modifications suggested by Rebecca Mason and Kristie Dotson are discussed. This section illustrates the ways in which epistemological and ethical concerns are intertwined and how matters of justice and injustice play a role in our collective pursuit of knowledge.

3.4.1 Fricker: Knowledge as a Site of Injustice

In her book "Epistemic Injustice: Power and the Ethics of Knowing", Fricker (2007) draws attention to power-related dynamics and resulting injustices in the sharing of and access to knowledge. She problematizes injustices that pertain to someone's capacity as a knower (on an individual but also structural level), subsumed under the notion of *epistemic injustice*. Fricker describes two different kinds of epistemic injustice, *testimonial injustice* and *hermeneutical injustice*. In the following, both concepts shall be explained in more detail.

Testimonial Injustice

At the basis of her account of epistemic injustice is her conception of *social power* as "a practically socially situated capacity to control other's actions, where this capacity may be exercised (actively or passively) by particular social agents, or alternatively, it may operate purely structurally" (p. 13). She argues that whenever power is at play, questions regarding "who is controlling whom, and why" must be

asked (p. 13). A particular type of power she is interested in here is *identity power*. A form of power that is significantly dependent on ("imaginative conceptions of") social identity. An example arena of this is gender. A common experience is, e.g., that a woman attempting to share her testimony is silenced by a man, which exemplifies an act of identity power. The same type of power might cause women not to even attempt at speaking, in certain situations. Acts of identity power do not need to be explicit and deliberate. Identity power lives within collective social imagination and affects the behaviors and perceptions of all agents involved. Hence, Fricker argues, identity power happens on a structural level.

Just like identity power, *identity prejudice* is carried by collective imaginations of identity. And the discrediting of a speaker due to identity prejudice is an execution of identity power. The main mechanism that influences identity-prejudicial credibility allocations are *stereotypes*. These are defined as "widely held associations between a given social group and one or more attributes" (p. 30). Most of the time, a hearer needs to immediately decide whether or not to believe what someone else is testifying–because they are in a conversational flow or learning something spontaneously without access to external means of verification. Hence, the need to rely on heuristics. This opens up the entryway for stereotypical attributions, as they can shape the assumptions that a hearer draws from when spontaneously judging someone else's testimony.

The central case of testimonial injustice is an *identity-prejudicial credibility deficit* experienced by the speaker. In other words, the speaker is not attributed (deserved) credibility due to an identity-related prejudice on the listener's end.[1] Fricker (2007) lists several examples where individuals are not taken as credible testifiers due to their gender, accent, or skin color. One example stems from the movie "To Kill a Mockingbird", in which the testimony of a White girl is weighed against the testimony of a Black boy, in a rape trial. Even though the evidence clearly indicates his innocence, the boy is ultimately found guilty by the jury and he later meets a fatal ending. While she understands that her choice of example might seem extreme, Fricker aims to illustrate how testimonial injustices are systematically linked to different areas of social experience: The boy in the story had prior been subject to all sorts of injustices caused by identity prejudice unrelated to testimony–in the educational, work, sexual, political, legal context, etc. Testimonial injustice that links to injustices happening across areas of life is *systematic*. The worst case of testimonial injustice is the type that is systematic as well as *persistent* over time. This is, ultimately, the kind of testimonial injustice that is the focal point of her analysis.

Hermeneutical Injustice

Hermeneutic injustices refers to inequalities regarding the availability or accessibility of hermeneutic resources. Just as identity power directs credibility assignments, it also directs the availability of hermeneutical resources. Structural

1. Of course, prejudice can also lead to *credibility excess*. According to Fricker (2007), however, this is generally not expected to be a disadvantage. Or considering that there could be cases in which it is a disadvantage, those are not considered to happen repeatedly over long courses of time. Credibility deficit, on the other hand, wrongs someone "specifically in her capacity as a knower" (p. 20) in a systematic and persistent way.

identity prejudice causes the *hermeneutically marginalization* of certain groups of people by obscuring "some significant area of [their] social experience [...] from collective understanding" (p. 155). Hence, hermeneutic injustice is a form of structural discrimination. In the Introduction (Chapter 1), the following example was used: the story of Carmita Wood who experienced sexual harassment without having the words to name it, leading to a sequence of adverse consequences (experiencing health issues, losing her job, not receiving unemployment insurance). Another example mentioned by Fricker (2007) is that of a woman who learns about postpartum depression after having experienced it for a while and having endured her husband's and her own blame for it. Obtaining a piece of information relevant to making sense of one's own experience or one's own self can be liberating and empowering in material and psychological ways. It may allow someone to make a life-changing decision and find better opportunities, and it may relieve psychological pressure by eliminating self-blame. *Not* being able to obtain such information, however, can *cause* material and psychological harms.

Harms of Epistemic Injustice

The harms of epistemic injustice are manifold. Firstly, there is a purely epistemic harm in not letting knowledge circulate into our epistemic system. In the case of testimonial injustice this happens when a hearer fails to attribute credibility to the testifier due to identity prejudice. In this case, knowledge is lost which could have contributed to the collective pursuit of truth.[2] Hermeneutical injustice causes the same exclusion, because the unavailability of relevant hermeneutical resources impacts what persons are able to say or how they are able to say it. It makes them less equipped to share their accounts. Besides such epistemic harms, there are also ethical harms done to the knower in both situations:

> "To be wronged in one's capacity as a knower is to be wronged in a capacity essential to human value. When one is undermined or otherwise wronged in a capacity essential to human value, one suffers an intrinsic injustice." (p. 44).

If we consider rationality as essential to our humanity, the experience of any type of epistemic injustice is, thus, not only to be "degraded *qua* knower", but also to be "degraded *qua* human" (p. 44). Because sharing ones knowledge and contributing to the collective activity of pooling knowledge are necessary conditions to acting as a knower, there is inherent ethical harm in not being allowed to do that. Moreover, gaps in collective knowledge resources can influence how the surroundings perceive and construct someone's experience for them. Fricker uses the example of a gay person not given the resources to make sense of their own feelings. Instead, people surrounding them would impose different causes and valuations onto them, based on their own prejudice.

There are also practical harms that follow from epistemic injustice. The example of the woman experiencing postpartum depression without proper understanding of her own experience exemplifies how hermeneutical injustice can harm a person's career perspectives. Of course, testimonial injustice can also cause

2. Fricker also briefly hints at the political importance of a free "collective speech situation".

obstacles in the work context. For instance, not being able to share own accounts and ideas might prevent a person from progressing in their career. Not being treated as credible can firstly cause a loss of confidence in one's own cognitive abilities and eventually lead to mistrust one's own beliefs. Similarly, being unable to resolve a dissonance between one's own experience and others' understanding of it–due to a lack of hermeneutic resources–can decrease one's epistemic confidence and increase mistrust in one's own epistemic capabilities. This again can ultimately cause "literal loss of knowledge" (p. 163). These examples come to show the epistemic and ethical harms of epistemic injustice are intertwined.

Referencing Iris Marion Young, Fricker adds that systematic and persistent testimonial injustice is a *face of oppression*. And this oppression is not necessarily always explicit but also at times "a by-product of residual prejudice in a liberal society" (p. 58).

Virtue Ethics as a Remedy

Fricker appeals to virtue as a remedy to epistemic injustice. She promotes the idea of training our testimonial sensibility, such that we learn to attribute credibility free of identity prejudice and correct testimonial and hermeneutical injustice. She characterizes a *virtuous hearer* that is aware and reflexive of the struggle a speaker might experience due to the unavailability of certain vocabulary or concepts. This includes a hearer's awareness that this obstacle in communication is a systematic one and not a "subjective failing". In any instance, the virtuous hearer aims to "[adjust] upwards [the degree of credibility] to compensate for the cognitive and expressive handicap imposed on the hermeneutically marginalized speaker, by the non-inclusive hermeneutical climate, by structural identity prejudice" (p. 170). In the long run, this virtue aims to change the hermeneutical climate such that hermeneutical lacuna are bridged. Fricker (2007) concludes, however, that individual virtues cannot finally resolve this issue that is ultimately caused by social power. Individual virtues can only do a small contribution. Without elaborating this proposition much further, Fricker suggests that the real solution lies in political action targeting a shift in power.

3.4.2 Mason & Dotson: On Ignorance and Oppression

Rebecca Mason and Kristie Dotson both offer criticism and consequent expansions of Fricker's notion of epistemic injustice. In particular, both thinkers insist that there are more than one set of collective hermeneutical resources and that power and ignorance play a role in keeping non-dominant hermeneutical resources excluded from dominant discourses. Furthermore, Dotson introduces the concept of *epistemic oppression* to act as an umbrella for different types of *epistemic exclusions*. She sees epistemic injustice as only one of many forms.

Blameworthy Ignorance

In "Two Kinds of Unknowing", Rebecca Mason critiques the original conception of hermeneutical injustice as negligent towards the epistemic agency of marginalized subjects. She argues that while being excluded from the dominant resources, the

marginalized still engage with non-dominant resources and "resistant epistemic and communicative practices" (Mason, 2011, p. 295). She supports her argument with the aforementioned story of Carmita Wood: after having experienced continuous sexual harassment at work and losing her job, Wood joined a feminist-run seminar. The women in this group discussed their shared experiences and eventually coined the term "sexual harassment". Giving this particular experience a name is an achievement that was surely illuminating to many women at the time. But, Mason points out, the fact that these women felt the need to and were able to name their experience shows that they very well had an understanding of their social experience. She continues to propose a distinction between "two kinds of unknowing" (p. 301): (1) hermeneutical injustice and (2) *epistemically and ethically blameworthy ignorance*.

Instead of seeking the cause for hermeneutical injustice in the gaps in our "collective" hermeneutical resources (resulting from the operations of identity power), Mason argues, we must distinguish between dominant versus non-dominant hermeneutical resources. Mason insists that what dominant hermeneutical resources fail to cover might very well be covered by those non-dominant ones. An example are those resources that were cultivated within the feminist movement. Non-dominant subjects might very well be able to understand their experience, their accounts might just not be able to transcend those non-dominant discourses because they are (willfully) ignored or distorted by the dominant group. By speaking only of "collective" hermeneutical resources, she argues, Fricker is failing to render these problematic practices visible. Moreover, she references feminist standpoint theory in claiming that "powerful groups do not *ipso facto* get a better view" (p. 301) and that disadvantaged subjects are able to obtain unique insights into society and power. Further, drawing from Mills (2017), she suggests that, indeed, powerful groups may be blame-worthily ignorant towards parts of their own experience such as to maintain the status quo that works in favor of their own privilege. For the harasser, staying ignorant towards the concept of sexual harassment or sexism and one's own role in it helps to sustain one's self-perceived innocence. The distorted conception of the harasser may not be harmful to the harasser himself, but it sure is harmful to the person being harassed. To summarize, a key take-away from Mason's account is that hermeneutical injustice is more than just the potential inability of non-dominant groups to make sense of their experiences. It can also take shape in "willfully sustained ignorance [that inhibits] communicative encounters between members of dominant and non-dominant groups" (Mason, 2011, p. 306).

Epistemic Oppression

Dotson (2012) proposes an expansion of Fricker's work that is similar to Mason's notion of *epistemically and ethically blameworthy ignorance*. She calls this form of epistemic injustice *contributory injustice*: it involves the *willful hermeneutical ignorance* of epistemic agents, which feeds into and benefits from biased hermeneutical resources. The biases are a result of structural identity prejudice and are not necessarily manifested in gaps in collective hermeneutical resources. Rather, they allude to the refusal in accepting or perceiving the alternative resources.

Epistemic injustice as introduced by Fricker is considered one form of *epistemic exclusion*, which Dotson (2012) defines as "an infringement on the epistemic agency of knowers that reduces her or his ability to participate in a given epistemic community" (p. 24). Persistent forms of epistemic exclusion give rise to *epistemic oppression* (Dotson, 2014, p. 115). This term was actually first coined by Fricker (1999) but not paid further attention until Dotson's contribution. She argues that this is due to the perception that epistemic oppression is reducible to political oppression. In response to this, she shows that there are three orders of epistemic oppression, of which the third one is inherently epistemic and not reducible to social or political oppression. She distinguishes the three orders by measures of how much effort would be involved to bring about change.[3]

First-order epistemic exclusion happens when an epistemic agent is violated in their ability to participate in knowledge production (Dotson, 2014). Testimonial injustice is one example of that (Dotson, 2012). Values that are already in place within an epistemic community are not properly applied. In the case of testimonial injustice, it is assumed the right thing to assign credibility to someone who wants to share their knowledge. However, based on an individual identity prejudice, this is not realized. Solving this issue requires first-order change, because the framework and values are the right ones, but are just not implemented correctly. This is (theoretically) solvable on a political level. *Second-order epistemic exclusion* is a result of a lack of shared epistemic resources, i.e., hermeneutical injustice. The cause is not an individual prejudice, but a structural one. Change would require a reform of political structures. Finally, *third-order epistemic exclusion* is brought about by "*inadequate* dominant, shared epistemic resources" (Dotson, 2014, p. 129). This is exemplified in contributory injustice. The epistemic framework that is in place is flawed. But for those situated within this framework it is difficult to come to this revelation, as the concepts provided by the framework do not easily permit the necessary understanding. It requires scrutiny of the established epistemic system from "outside". The paradigm shift that would be needed to solve this type of issue is hard and unlikely to achieve. Due to the specifically epistemic revelation needed here, third-order epistemic oppression is distinctively epistemic and not reducible to political oppression.

3.4.3 Take-Aways

Fricker's notion of epistemic injustice is helpful in developing an understanding of the specific types of harm done by biased AI specifically in the context of knowledge. In particular the notion of hermeneutical injustice is directly applicable to the lacunae within authoritative, widely used AI-based knowledge technology. Epistemic injustice is a "degra[dation] *qua* human" (Fricker, 2007, p. 44), that can cause psychological and material harm. The loss of epistemic confidence through experiences of testimonial injustice and the lack of hermeneutical resources to permit expressing oneself adequately keeps knowers from contributing to knowledge-productive or -disseminating processes. This, effectively, is a loss for the collective pursuit of knowledge.

3. The *order of change heuristic* is drawn from organizational development theory (Bartunek and Moch, 1987).

As Mason (2011) and Dotson (2012) point out, hermeneutical injustice are not only attributable to structural circumstances. They are also an effect of willful, blameworthy ignorance. Those experiencing–what Dotson calls–contributory injustice have their own epistemic resources to make sense of their experience, as exemplified in the case of Carmita Wood. The issue is that the non-dominant hermeneutical resources are not accepted, nor taken up into the dominant discourse. This is to the effect that those who are powerful maintain the existing structures and evade scrutiny. This situation is difficult to resolve, because it is a flaw of the very epistemic framework. In the context of this thesis, this third-order type of epistemic exclusion is epitomized in AI benchmarking practices. The established epistemic tools are flawed such that biases are rendered invisible. We shall add the following core insights to the list:

6. Knowledge can be a site of injustice: operations of power and bad epistemic practice can cause epistemic and ethical harm to marginalized groups.

7. Epistemic oppression can be a matter of the epistemic paradigm itself: in analyzing forms of epistemic exclusion, flaws in the overall paradigmatic framework must be considered as a potential source.

3.5 The Feminist Study of AI

After this introduction to selected theories in feminist epistemology, we shall shift our attention back towards AI technology. This section presents two works that have been among the first to evaluate AI systems and human-machine interaction based on assumptions of situatedness and feminist accounts of objectivity. This includes a contribution by Alison Adam, which discusses epistemic and ethical aspects of 1990's symbolic AI technology, and Lucy Suchman's analysis of human-machine interaction.

3.5.1 Adam: AI and the View From Nowhere

The work presented by Adam (1998) draws from studies on gender and technology, as well as gender and science, and feminist epistemology. She draws from the discourse around the *view from nowhere*, i.e., masculinist accounts of disembodied, unpositioned rationality, and objectivity discussed in the previous sections. At the backdrop of the feminist-epistemological critique of this account, Adam investigates whether the knowing subject is made explicit in the context of AI; and if so, how it is characterized and, and if not, how it is implicitly inscribed.

The type of AI system Adam (1998) analyzes in her book is symbolic and based on encyclopedic KGs. This was the dominant paradigm in the field, at the time. At the time, distributed approaches received much less attention. Her comparison between knowledge in AI and knowledge in humans is not meant to imply that AI *knows* or *can know*. Her aim is rather to articulate the contrasts between "artificial knowing" and "human knowing", in particular "knowing of women" (p. 1). That is, Adam argues that gender and a "gendered vision of the world" (p. 1) is implicitly *inscribed* in AI. Gender, here, is defined as the distinction between masculinity and

femininity as general, value-laden concepts. Her claim is that AI systems, allegedly created to model *human* knowledge, is in actuality modeled after *male* knowledge.

One of the systems she analyzes in much detail is Cyc, a large-scale knowledge base built with the vision of "spanning most human common sense or consensual knowledge" (Adam, 1998, p. 81) from the 1990s. The director of the project, Doug Lenat, is said to have originally envisioned a system that knows a large part of *consensus reality*, that is all-encompassing and general-purpose to the extend that all new computers would benefit from being shipped with Cyc pre-installed. While this vision might not seem all to strange in the eyes of a 2020's reader, it was indeed seen as quite bold, back then. This was to the point that the project aims were toned down significantly, later on. In the project descriptions of Cyc, Adam was not able to find clear indicators of the knowing subjects. Not only is "consensus" taken for granted, but also the subjects that are consenting. So, she resorted to deducing the *who* from the *what* by analyzing the specifications and contents of the system.

With regards to *what content* gets to be represented in a system like Cyc, a number of strong assumptions have to be made. For instance, even if the model is able to represent different competing theories of a problem (e.g, different economical theories), it has to be decided which are known or relevant enough to go into the system. Adam finds that the developers work with the assumption that "the real world" can easily be accessed, and that what is truer than other things will be clear to those with normal, "non-weird" perspectives (which are of course unspecified). Adam, however, criticizes this perspective from a feminist stance, e.g., those discussed in earlier sections of this Chapter. According to this feminist view, neutral and direct observations of the world are impossible. They are mediated by intent and partiality. According to Adam, in seeking out "consensus" knowledge, the system is made to represent hegemonic worldviews without taking into consideration the power and privilege that form the hegemony. And she argues that these views are identifiable as reflecting the views of the system designers themselves. Adam (1998) also analyzed another system named Soar, which meant to represent problem spaces and to facilitate search-based extraction of solutions across different domains (logic problems, NLP, tactics, etc.). It was built on a theory of the nature of human problem solving backed by evidence from observation studies with technically educated, young, male, college students.

In conversation with the anthropological study of Forsythe (1993), who observed that the engineers behind expert systems "delete the social", Adam argues that engineers of symbolic AI systems delete the *subject*. This, she sees as "a consequence of their following in the footsteps of the classical position in epistemology where the nature of the knowing subject has been traditionally denied as an essential element in the making of knowledge" (Adam, 2000 p. 251). And as was discussed in Haraway (1988), the deletion of the subject purports a *view from nowhere* that evades critical analysis of the knower's positioning and accountability. As Adam is discussing power, she draws attention specifically to the "hierarchy of knowers" that have stakes in AI systems. The top of the hierarchy is inhabited by the mostly White and male engineers, privileged with the power to make ontological decisions and to decide which knowledge is accepted enough to be eternalized in the system. And this again acts as a confirmation of their worldviews. Their views and knowledges are seemingly objectified through technology, and at the same time distributed at scale. This feeds into the continuous marginalization

of the less powerful knowers. In making aware of these issues, Adam hopes to inspire efforts for more inclusive approaches to AI engineering, in the spirit of a *successor technology* (Adam, 1998).

3.5.2 Suchman: Humans, Machines and Situated Actions

In her 1987 book "Plans and Situated Actions: The Problem of Human-Machine Communication", Lucy Suchman dissects the relationship between *plans* and *actions* and draws from it insights and implications regarding the study of human-machine interaction.[4] Suchman studied users' interactions with the "expert help system" of a photocopier, which had been developed with the intention to interactively guide users through their operations on the machine. The design of this interaction was realized in accordance with the *planning model of human action and communication*, a rationalist theory of human cognition and planning, which was the dominant paradigm in interactive AI design, at the time. Presumed goals and plans of users operating the photocopier were represented in a structured "knowledge representation" as "condition-action rules" (Suchman, 2007, p. 10). However, in an observational study, Suchman documented several issues users experienced when interacting with the expert help system. She attributed these issues to problems with the planning model itself. According to the planning model, plans are preconditions and prescriptions of action sequences, and plans give meaning to actions. This conceptualization is rejected by Suchman, who claims that actions are *situated*, determined by "*local interactions* contingent on the actor's particular [social and material] circumstances" (Suchman, 1987p. 28; emphasis added). Rather than being a prescription of some sort, a plan is an abstract representation that allows us to project a potential sequence of actions, or to reconstruct an already occurred course of actions to facilitate reasoning and communication about it. In any case, plans are representations that are efficient, precisely, because they do *not* capture each action or circumstantial matter at the highest level of detail. In a way, her account reminds of Longino's map analogy: a map that is a perfect copy of the surface of the earth would defeat its purpose of allowing to reason and communicate about certain characteristics efficiently (Longino, 2002). Likewise, a plan would defeat the purpose of allowing to reason and communicate about certain characteristics or milestones, if it were a perfect representation of every single action and situational detail (potentially) involved in reaching its goal state. Rather, a plan is a resource for situated action. In particular, in the face of unexpected issues, actions are decided under consideration of the concrete circumstances *and* the explicated rules and goals of a plan, in order to problem-solve. And so, she draws a line between machines and humans, because machines do not have the same type of access to the real world, they cannot spontaneously adapt to potential contingencies.

4. She also published an extended second edition of the book, entitled "Human-Machine Reconfigurations: Plans and Situated Actions" (Suchman, 2007) 20 years later. This extended version is a testament for the fact that her work has had great influence on the fields of STS, human-computer interaction (HCI), and AI. It includes the first edition in full and adds additional chapters to shift the focus more towards the boundaries between human and machine, and what they tell us about their agencies. Please note that I am refraining from presenting these additional thoughts in detail here, as these are beyond the scope of the inquiry presented in this dissertation.

Suchman (1987) also discusses assumptions about *communication* implied in the rationalist account. The rationalist account suggests that communication of intents (to reach mutual intelligibility) is achieved through conventional expressions, and by drawing from a shared body of knowledge about typical situations and appropriate actions. She, however, argues that any communicative expression is, again, an incomplete representation. The actual meaning of any communicative expression depends on the situation of its use. And a lot of this situation remains "unspoken". She also hints at the computational practice of utilizing knowledge bases to access contextual information, asserting that the premise is faulty: "there is no fixed set of assumptions that underlies a given statement" (p. 61). Hence, instructions given to a machine can never be complete. Moreover, the use of communicative expressions also influences the situation. Despite the vagueness of language, interpersonal communication is mostly successful, because we collaboratively develop a sense of mutual understanding. "Communication [...] is not a symbolic process that happens to go on in real-world settings but a real-world activity in which we make use of language to delineate the collective relevance of our shared environment" (Suchman, 2007, p. 178).

Furthermore, she claims that there is no such thing as an *a priori* "common sense" objective reality. Objectivity is a result of social interactions. Social practices, schemas, and methods (which themselves are products of social practices), which we use to communicate individual experiences and circumstances, allow to "render the world publicly available and mutually intelligible" (Suchman, 1987, p. 57); the product of which is a shared understanding of the social world. This again, makes it implausible to model "common sense" as a computational resource.

3.6 Take-Aways

Both of these works shed light on the de-situated, disembodied understandings of knowledge, action, and communication that have dominated engineering practices. They demonstrate in which ways these conceptualizations are limiting to the engineering practice and knowledge discovery, as well as ethically harmful.

The system discussed by Adam (1998) is comparable to the types of KGs that are at the core of many of the analyses in this dissertation. Based on the parallels, there are a lot of direct conclusions to draw: *what content* gets presented in encyclopedic KGs is highly dependent on the social situation and worldviews of those who get to design these systems. This gives them a unique privilege, as it enables them to ratify and disseminate their own knowledge through an authoritative form of representation. This constitutes a hierarchy of knowers. And in the case of Cyc, this hierarchy is led by a homogeneous, Western, White, and male demographic. In abstracting away or *deleting the subject* from the technology, it applies the god trick and evades scrutiny and accountability. As we will see in Chapters 4 and 5 of this thesis, these observations are all applicable to modern-day AI-based knowledge technology.

Suchman (1987) illuminates how the *view from nowhere* is also found in conceptualizations of human-machine interaction. She states that human actions are situated and must be studied as such. Plans are abstract, purpose-driven representations of situated action; language is abstract, purpose-driven represen-

tation of actual intent. Furthermore, she rejects the idea of an *a priori* objective reality. Instead, she claims that objective reality emerges from interpersonal interactions. So, communication is not fully representable in symbolic form and knowledge is not either. Suchman's work has two implications for this thesis: firstly, it emphasizes that studying the effects of machines on humans requires a situated analysis of this type of interaction. While this thesis is clearly not an HCI study, there are still significant elements of human-machine interaction involved. That is, in evaluating epistemic and ethic goodness, we must look beyond the characteristics or composition of the system itself and consider the wider social context. In critiquing AI evaluation metrics and benchmarks, we must do the same thing. Evaluation of system behavior must be contextualized (Barocas et al., 2023), as the study presented in Chapter 6 will demonstrate in more detail. Secondly, Suchman's work can be interpreted as a pointing towards limitations of enhancing LMs with "world knowledge", as such cannot be fully represented outside of direct human interactions (see Chapter 5). Finally, the following point shall be added to the summary of core take-aways from this Chapter:

8. Evaluation of technology must be contextualized: human-machine interaction is situated and can only be adequately analyzed in a contextualized manner.

3.7 Conclusion

This Chapter provided an introduction to epistemology, emphasizing the differences between "traditional" Western and social epistemology, in particular feminist epistemology. It compared feminist standpoint theory and feminist empiricism, and presented a consolidation of both accounts as offered by Intemann (2010). The notions of epistemic injustice and -oppression were introduced and discussed. Finally, two critical works bridging feminist theory and AI were summarized. The following is the complete list of key take-aways, which will be referenced in the final discussion in Chapter 7:

1. Knowledge is social, situated, partial, and embodied: gender, race/ethnicity, class, social relations, and roles influence what we can know.

2. Knowledge is contextual: justificatory processes are embedded in a locally shared set of assumptions, research goals, and methodologies.

3. Objectivity requires empirical evidence: justification needs to be anchored in empirical evidence and embodied experience.

4. Objectivity requires diversity of social position: different experience must contribute to knowledge-productive processes, and the experience of marginalized groups are epistemically particularly valuable.

5. Objectivity benefits from democracy-advancing values: values that drive diversity and are power-sensitive are beneficial to science.

6. Knowledge can be a site of injustice: operations of power and bad epistemic practice can cause epistemic and ethical harm to marginalized groups.

7. Epistemic oppression can be a matter of the epistemic paradigm itself: in analyzing forms of epistemic exclusion, flaws in the overall paradigmatic framework must be considered as a potential source.

8. Evaluation of technology must be contextualized: human-machine interaction is situated and can only be adequately analyzed in a contextualized manner.

4

The Lifecycle of "Facts": A Survey of Social Bias in Knowledge Graphs

Knowledge graphs are increasingly used in a plethora of downstream tasks or in the augmentation of statistical models to improve factuality. However, social biases are engraved in these representations and propagate downstream. We conducted a critical analysis of literature concerning biases at different steps of a knowledge graph lifecycle. We investigated factors introducing bias, as well as the biases that are rendered by knowledge graphs and their embedded versions afterward. Limitations of existing measurement and mitigation strategies are discussed and paths forward are proposed.

Publication Reference: Angelie Kraft and Ricardo Usbeck. 2022. The Lifecycle of "Facts": A Survey of Social Bias in Knowledge Graphs. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, Volume 1: Long Papers (AACL-IJCNLP 2022)*, 639–652. Online: Association for Computational Linguistics.

Contents

4.1 Introduction

Knowledge graphs (KGs) provide a structured and transparent form of information representation and lie at the core of popular Semantic Web technologies. They are utilized as a source of truth in a variety of downstream tasks (e.g., information extraction (Martínez-Rodríguez et al., 2020), link prediction (Getoor and Taskar, 2007; Ngomo et al., 2021), or question-answering (Höffner et al., 2017; Diefenbach et al., 2018; Chakraborty et al., 2021; Jiang and Usbeck, 2022)) and in hybrid AI systems (e.g., knowledge-augmented language models (Peters et al., 2019; Sun et al., 2020; Yu et al., 2022) or conversational AI (Gao et al., 2018; Gerritse et al., 2020)). In the latter, KGs are employed to enhance the factuality of statistical models (Athreya et al., 2018; Rony et al., 2022). In this overview article, we question the ethical integrity of these facts and investigate the lifecycle of KGs (Auer et al., 2012; Paulheim, 2017) with respect to bias influences.[1]

We claim that KGs manifest social biases and potentially propagate harmful prejudices. To utilize the full potential of KG technologies, such ethical risks must be targeted and avoided during development and application. Using an extensive literature analysis, this article provides a reflection on previous efforts and suggestions for future work.

We collected articles via Google Scholar[2] and filtered for titles including *knowledge graph/base/resource*, *ontologies*, *named entity recognition*, or *relation extraction*, paired with variants of *bias*, *debiasing*, *harms*, *ethical*, and *fairness*. We selected peer-reviewed publications (in journals, conference or workshop

1. We focus on the KG lifecycle from a bias and fairness lens. For reference, the processes investigated in Section 4.3 correspond to the *authoring stage* in the taxonomy by Auer et al. (2012). The representation issues in KGs (Section 4.4) and KG embeddings (Sections 4.5 and 4.7) which affect downstream task bias relate to Auer et al.'s *classification stage*.

2. A literature search on Science Direct, ACM Digital Library, and Springer did not provide additional results.

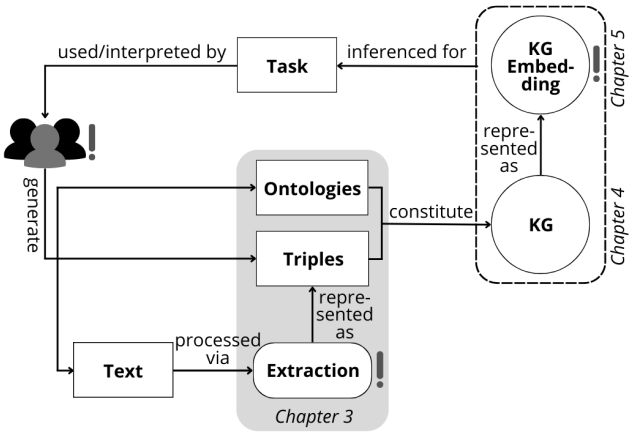

Figure 4.1: Overview of the knowledge graph lifecycle as discussed in this paper. Exclamation marks indicate factors that introduce or amplify bias. We examine bias-inducing factors of triple crowd-sourcing, hand-crafted ontologies, and automated information extraction (Chapter 4.3), as well as the resulting social biases in KGs (Chapter 4.4) and KG embeddings, including approaches for measurement and mitigation (Chapter 4.5).

proceedings, and book chapters) from 2010 onward, related to social bias in the KG lifecycle. This resulted in a final count of 18 papers. Table 4.1 gives an overview of the reviewed works and Figure 4.1 illustrates the analyzed lifecycle stages.

4.2 Notes on Bias, Fairness, and Factuality

In the following, we clarify our operational definitions of the most relevant concepts in our analysis.

4.2.1 Bias

If we refer to a model or representation as *biased*, we — unless otherwise specified — mean that the model or representation is *socially biased*, i.e., biased towards certain social groups. This is usually indicated by a systematic and unfairly discriminating deviation in the way members of these groups are represented compared to others (Friedman and Nissenbaum, 1996) (also known as *algorithmic bias*). Such bias can stem from pre-existing societal inequalities and attitudes, such as prejudice and stereotypes, or arise on an algorithmic level, through design choices and formalization (Friedman and Nissenbaum, 1996). From a more impact-focused perspective, algorithmic bias can be described as "a skew that [causes] harm" (Kate Crawford, Keynote at NIPS2017). Such harm can manifest itself in unfair distribution of resources or derogatory misrepresentation of a disfavored group. We refer to *fairness* as the absence of bias.

4.2.2 Unwanted Biases and Harms

One can distinguish between *allocational* and *representational harms* (Barocas et al., as cited in, Blodgett et al., 2020), where the first refers to the unfair distribution

of chances and resources and the second more broadly denotes types of insult or derogation, distorted representation, or lack of representation altogether. To quantify biases that lead to representational harm, analyses of more abstract constructs are required. Mehrabi et al. (2021a), for example, measure indicators of representational harm via *polarized perceptions*: a predominant association of groups with either negative or positive prejudice, denigration, or favoritism. Polarized perceptions are assumed to correspond to societal stereotypes. They can *overgeneralize* to all members of a social group (e.g., "*all* lawyers are dishonest"). It can be said that harm is to be prevented by avoiding or removing algorithmic bias. However, different views on the conditions for fairness can be found in the literature and, in consequence, different definitions of *unwanted* bias.

4.2.3 Factuality versus Fairness

We consider a KG factual if it is representative of the real world. For example, if it contains only male U.S. presidents, it truthfully represents the world as it is and has been. However, inference based on this snapshot would lead to the prediction that people of other genders cannot or will not become presidents. This would be false with respect to U.S. law and/or undermine the potential of non-male persons. Statistical inference over historical entities is one of the main usages of KGs. The factuality narrative, thus, risks consolidating and propagating pre-existing societal inequalities and works against matters of social fairness. Even if the data represented are not affected by sampling errors, they are restricted to describing *the world as it is* as opposed to *the world as it should be*. We strive for the latter kind of inference basis. Apart from that, in the following sections we will learn that popular KGs are indeed affected by sampling biases, which further amplify societal biases.

4.3 Entering the Lifecycle: Bias in Knowledge Graph Creation

We enter the lifecycle view (Figure 4.1) by investigating the processes underlying the creation of KGs. We focus on the human factors behind the authoring of *ontologies* and *triples* which constitute KGs. Furthermore, we address automated *information extraction*, i.e., the detection and extraction of entities and relations from text, since these approaches can be subject to algorithmic bias.

4.3.1 Triples: Crowd-Sourcing of Facts

Popular large-scale KGs, like Wikidata (Vrandečić and Krötzsch, 2014) and DBpedia (Auer et al., 2007b) are the products of continuous crowd-sourcing efforts. Both of these examples are closely related to Wikipedia, where the top five languages (English, Cebuano, German, Swedish, and French) constitute 35% of all articles on this platform.[3] It can be said that Wikipedia is Euro-centric in tendency.

3. https://en.wikipedia.org/wiki/List_of_Wikipedias

Moreover, the majority of authors are white males.[4] As a result, the data transport a particular homogeneous set of interests and knowledge (Beytía et al., 2022; Wagner et al., 2015). This *sampling bias* affects the geospatial coverage of information (Janowicz et al., 2018) and leads to higher barriers for female personalities to receive a biographic entry (Beytía et al., 2022). In an experiment, Demartini (2019) asked crowd contributors to provide a factual answer to the (politically charged) question of whether or not Catalonia is a part of Spain. The diverging responses indicated that participants' beliefs of what counts as true differed largely. This is an example of bias that is beyond a subliminal psychological level. In this case, structural aspects like consumed media and social discourse play an important role. To counter this problem, Demartini (2019) suggests actively asking contributors for evidence supporting their statements, as well as keeping track of their demographic backgrounds. This makes underlying motivations and possible sources for bias traceable.

4.3.2 Ontologies: Manual Creation of Rules

Ontologies determine rules regarding allowed types of entities and relations or their usage. They are often hand-made and a source of bias (Janowicz et al., 2018) due to the influence of opinions, motivations, and personal choices (Keet, 2021): Factors like scientific opinions (e.g., historical ideas about race), socio-culture (e.g., how many people a person can be married to), or political and religious views (e.g., classifying a person of type X as a *terrorist* or a *protestor*) can proximately lead to an encoding of social bias. Also structural constraints like the ontologies' granularity levels can induce bias (Keet, 2021). Furthermore, issues can arise from the types of information used to characterize a person entity. Whether one attributes the person with their skin color or not could theoretically determine the emergence of racist bias in a downstream application (Paparidis and Kotis, 2021). Geller and Kollapally (2021) give a practical example for detection and alleviation of ontology bias in a real-world scenario. The authors discovered that ontological gaps in the medical context lead to an under-reporting of race-specific incidents. They were able to suggest countermeasures based on a structured analysis of real incidents and external terminological resources.

4.3.3 Extraction: Automated Extraction of Information

Natural language processing (NLP) methods can be used to recognize and extract entities (named entity recognition; NER) and their relations (relation extraction; RE), which are then represented as [head entity, relation, tail entity] tuples (or as [subject, predicate, object], respectively).

Mehrabi et al. (2020) showed that the NER system CoreNLP (Manning et al., 2014) exhibits binary gender bias. They used a number of template sentences, like "<Name> is going to school" or "<Name> is a person" using male and female

4. https://en.wikipedia.org/wiki/Gender_bias_on_Wikipedia;https://en.wikipedia.org/wiki/Racial_bias_on_Wikipedia

names[5] from 139 years of census data. The model returned more erroneous tags for female names. Similarly, Mishra et al. (2020) created synthetic sentences from adjusted Winogender (Rudinger et al., 2018) templates with names associated with different ethnicities and genders. A range of different NER systems were evaluated (bidirectional LSTMs with Conditional Random Field (BiLSTM CRF) (Huang et al., 2015) on GloVe (Pennington et al., n.d.), ConceptNet (Speer et al., 2017) and ELMo (Peters et al., 2017) embeddings, CoreNLP, and spaCy[6] NER models). Across models, non-white names yielded on average lower performance scores than white names. Generally, ELMo exhibited the least bias. Although ConceptNet is debiased for gender and ethnicity[7], it was found to produce strongly varied accuracy values.

Gaut et al. (2020) analyzed binary gender bias in a popular open-source neural relation extraction (NRE) model, OpenNRE (Han et al., 2019). For this purpose, the authors created a new dataset, named WikiGenderBias (sourced from Wikipedia and DBpedia). All sentences describe a gendered subject with one of four relations: *spouse*, *hypernym*, *birthData*, or *birthPlace* (DBpedia mostly uses occupation-related hypernyms). The most notable bias found was the spouse relation. It was more reliably predicted for male than female entities. This observation stands in contrast to the predominance of female instances with spouse relation in WikiGenderBias. The authors experimented with three different mitigation strategies: downsampling the training data to equalize the number of male and female instances, augmenting the data by artificially introducing new female instances, and finally word embedding debiasing (Bolukbasi et al., 2016). Only downsampling facilitated a reduction of bias that did not come at the cost of model performance.

Nowadays, contextualized transformer-based encoders are used in various NLP applications, including NER and NRE. Several works have analyzed the various societal biases encoded in large-scale word embeddings (like word2vec (Mikolov et al., 2013; Bolukbasi et al., 2016) or BERT (Devlin et al., 2019; Kurita et al., 2019)) or language models (like GPT-2 (Radford et al., 2019; Kirk et al., 2021) and GPT-3 (Brown et al., 2020; Abid et al., 2021)). Thus, it is likely that these biases also affect the downstream tasks discussed here. Li et al. (2021b) used two types of tasks to analyze bias in BERT-based RE on the newly created Wiki80 and TACRED (Zhang et al., 2017) benchmarks. For the first task, they masked only entity names with a special token (*masked-entity*; ME), whereas for the second task, only the entity names were given (*only-entity*; OE). The model maintained higher performances in the OE setting, indicating that the entity names were more informative of the predicted relation than the contextual information. This hints at what the authors call *semantic bias*.

A Note on Reporting Bias Generally, when extracting knowledge from text, one should be aware that the frequency with which facts are reported is not representative of their real-world prevalence. Humans tend to mention only

5. While most of the works presented here refer to gender as a binary concept, this does not agree with our understanding. We acknowledge that gender is continuous and technology must do this reality justice.

6. https://spacy.io/

7. https://blog.conceptnet.io/posts/2017/conceptnet-numberbatch-17-04-better-less-stereotyped-word-vectors/

events, outcomes, or properties that are out of their perceived ordinary (Gordon and Van Durme, 2013) (e.g., "a banana is yellow" is too trivial to be reported). This phenomenon is called *reporting bias* and likely stems from a need to be as informative and non-redundant as possible when sharing knowledge.

4.4 Bias in Knowledge Graphs

Next in our investigation of the lifecycle (Figure 4.1) comes the representation of entities and relations as a KG. In the following, we illustrate which social biases are manifested in KGs and how.

4.4.1 Descriptive Statistics

Janowicz et al. (2018) demonstrated that DBpedia, which is sourced from Wikipedia info boxes, mostly represents the western and industrialized world. Matching the coverage of location entries in the KG with population density all over the world showed that several countries and continents are underrepresented. A disproportionate 70% of the person entities in Wikidata are male (20% are female, less than 1% are neither male nor female, and for roughly 10% the gender is not indicated) (Beytía et al., 2022). Radstok et al. (2021) found that the most frequent occupation is *researcher* and Beytía et al. (2022) identified *arts*, *sports*, and *science and technology* as the most prominent occupation categories. In reality, only about 2% of people in the U.S. are researchers (Radstok et al., 2021). This gap is likely caused by reporting bias as discussed earlier (Section 4.3.3). Radstok et al. (2021), moreover, observed that mentions of ethnic group membership decreased and changed in focus between the 18th and 21st century. Greeks are the most frequently labeled ethnic group among historic entries (over 400 times) and African Americans among modern entries (only roughly 100 times).

4.4.2 Semantic Polarity

Mehrabi et al. (2021b) focused on biases in common sense KGs like ConceptNet (Speer et al., 2017) and GenericsKB (Bhakthavatsalam et al., 2020) (contains sentences) which are at risk of causing representational harms (see Section 4.2.2). They utilized *regard* (Sheng et al., 2019) and *sentiment* as intermediate bias proxies. Both concepts express the polarity of statements and can be measured via classifiers that predict a neutral, negative, or positive label (Sheng et al., 2019; Dhamala et al., 2021). Groups that are referred to in a mostly positive way are interpreted as favored and vice versa. Mehrabi et al. (2021b) applied this principle to natural language statements generated from ConceptNet triples. They found that subject and object entities relating to the professions *CEO*, *nurse*, and *physician* were more often favored while *performing artist*, *politician*, and *prisoner* were more often disfavored. Similarly, several Islam-related entities were on the negative end while *Christian* and *Hindu* were more ambiguously valuated. As for gender, no significant difference was found.

Table 4.1: Overview of reviewed works concerning the sources, measurement, and mitigation of bias in KGs/KGEs.

Bias Source		
Crowd-Sourcing		Beytía et al. (2022), Janowicz et al. (2018), and Demartini (2019)
Ontologies		Janowicz et al. (2018), Keet (2021), Paparidis and Kotis (2021), and Geller and Kollapally (2021)
Extraction		Mehrabi et al. (2020), Mishra et al. (2020), Gaut et al. (2020), and Li et al. (2021b)
Bias Measurement		
Representation	*Method*	
KG	Descriptive Statistics	Janowicz et al. (2018), Radstok et al. (2021), and Beytía et al. (2022)
	Semantic Polarity	Mehrabi et al. (2021b)
KGE	Analogies	Bourli and Pitoura (2020)
	Projection	Bourli and Pitoura (2020)
	Update-Based	Fisher et al. (2020b), Keidar et al. (2021), and Du et al. (2022)
	Link Prediction	Keidar et al. (2021), Arduini et al. (2020), Radstok et al. (2021), and Du et al. (2022)
Bias Mitigation		
Representation	*Method*	
KGE	Data Balancing	Radstok et al. (2021) and Du et al. (2022)
	Adversarial Learning	Fisher et al. (2020a) and Arduini et al. (2020)
	Hard Debiasing	Bourli and Pitoura (2020)

4.5 Bias in Knowledge Graph Embeddings

Vector representations of KGs are used in a range of downstream tasks or combined with other types of neural models (Nickel et al., 2016; Ristoski et al., 2019). They facilitate efficient aggregation of connectivity patterns and convey latent information.

Embeddings are created through statistical modeling and summarize distributional characteristics. So, if a KG like Wikidata contains mostly (if not only) male presidents, the relationship between the gender *male* and the profession *president* is assumed to manifest itself accordingly in the model. In fact, the papers summarized below provide evidence that the social biases of KGs are modeled or further amplified by KG embeddings (KGEs). The following sections are organized by measurement strategy to give an overview of existing approaches and the information gained from them.

4.5.1 Stereotypical Analogies

The idea behind analogy tests is to see whether demographics are associated with attributes in stereotypical ways (e.g., "Man is to computer programmer as woman is to homemaker"; Bolukbasi et al., 2016). In their in-depth analysis of a TransE-embedded Wikidata KG, Bourli and Pitoura (2020) investigated occupational analogies for binary gender seeds. TransE (Bordes et al., 2013) represents (h, r, t) (with head h, relation r, tail t) in a single space such that $h + r \approx t$. The authors identified the model's most likely instance of the claim "a is to x as b is to y" (with *(a,b)* being a set of demographics seeds and *(x,y)* a set of attributes) via a cosine score: $S_{(a,b)}(x, y) = cos(\vec{a} + \vec{r} - \vec{b}, \vec{x} + \vec{r} - \vec{y})$, where r is the relation *has_occupation*. In their study, the highest scoring analogy was "woman is to fashion model as man is to businessperson". This example appears rather stereotypical, but other highly ranked analogies less so, like "Japanese entertainer" versus "businessperson" (Bourli and Pitoura, 2020). A systematic evaluation of how stereotypical the results

are is missing here. In comparison, the work that originally introduced analogy testing for word2vec (Bolukbasi et al., 2016) employed human annotators to rate stereotypical and gender-appropriate analogies (e.g., "sister" versus "brother").

4.5.2 Projection onto a Bias Subspace

Projection-based measurement of bias is another approach that was first proposed by Bolukbasi et al. (2016) for word embeddings, and was adapted for TransE by Bourli and Pitoura (2020). In a first step, a one-dimensional gender direction $\vec{d}_g$ is extracted. Then, a projection score metric S is computed to indicate gender bias — with projection π of an occupation vector $\vec{o}$ onto $\vec{d}_g$ and a set of occupations C: $S(C) = \frac{1}{|C|} \sum_{o \in C} \|\pi_{\vec{d}_g} \vec{o}\|$. Occupations with higher scores are interpreted as more gender-biased and those with close-to-zero scores as neutral.

4.5.3 Update-Based Measurement

The *translational likelihood* (TL) metric was tailored for translation-based modeling approaches (Fisher et al., 2020b). To compute this metric, the embedding of a person entity is updated for one step towards one pole of a seed dimension. This update is done in the same way as the model was originally fit in. For example, if head entity *person x* is updated in the direction of *male* gender, the TL value is given by the difference between the likelihood of *person x* being a *doctor* after versus before the update. If the absolute value averaged across all human entities is high, this indicates a bias regarding the examined seed-attribute pair. Fisher et al. (2020b) argue that this measurement technique avoids model-specificity as it generalizes to any scoring function. However, Keidar et al. (2021) found that the TL metric does not compare well between different types of embeddings (details in Section 4.6). It should, thus, only be used for the comparison of biases within one kind of representation. Du et al. (2022) propose an approach comparable to Fisher et al. (2020b) to measure individual-level bias. Instead of updating towards a gender dimension, the authors suggest flipping the entity's gender and fully re-training the model afterward. The difference between pre- and post-update link prediction errors gives the bias metric. A validation of the approach was done on TransE for a Freebase subset (FB5M (Bordes et al., 2015)) (Du et al., 2022). The summed per-gender averages (group-level metric) were found to correlate with U.S. census gender distributions of occupations.

4.6 Downstream Task Bias: Link Prediction

Link prediction is a standard downstream task that targets the prediction of relations between entities in a given KG. Systematic deviations in the relations suggested for entities with different demographics indicate reproduced social bias.

For the measurement of fairness or bias in link prediction, Keidar et al. (2021) distinguish between *demographic parity* versus *predictive parity*. The assumption underlying demographic parity is that the equality between predictions for demographic counterfactuals (opposite demographics, for example, *female* versus *male* in binary understanding) is the ideal state (Dwork et al., 2012). That is, the

probability of predicting a label should be the same for both groups. Predictive parity is given, on the other hand, if the probability of true positive predictions (*positive predictive value* or *precision*) is equal between groups (Chouldechova, 2017). Hence, this measure factors in the label distribution by demographic. With these metrics, Keidar et al. (2021) analyzed different embedding types, namely TransE, ComplEx, RotatE, and DistMult, each fit on the benchmark datasets FB15k-237 (Toutanova and Chen, 2015) and Wikidata5m (Wang et al., 2021b). They averaged the scores across a large set of human-associated relations to detect automatically which relations are most biased. The results showed that *position played on a sports team* was most consistently gender-biased across embeddings. Arduini et al. (2020) analyzed link prediction parity regarding the relations *gender* and *occupation* to estimate debiasing effects on TransH (Wang et al., 2014) and TransD (Ji et al., 2015). The comparability between different forms of vector representations is a strength of downstream metrics. In contrast, measures like the analogy test or projection score (Bourli and Pitoura, 2020) are based on specific distance metrics and TL (Fisher et al., 2020b) was shown to lack transferability across representations (Keidar et al., 2021) (Section 4.5.3).

Du et al. (2022) interpret the correlation between gender and link prediction errors as an indicator of group bias. With this, they found, for example, that *engineer* and *nurse* are stereotypically biased in FB5M. However, the ground truth gender ratio was found not predictive of the bias metric (e.g., despite its higher male ratio, *animator* produced a stronger female bias value). For validation, it was shown that the predicted bias values correlate to the gender distributions of occupations according to U.S. census (again, on TransE). Furthermore, the authors investigated how much single triples contribute to group bias via an *influence function*. They found that gender bias is mostly driven by triples containing gendered entities and triples of low degree.

4.7 Breaking the Cycle? Bias Mitigation in Knowledge Graph Embeddings

A number of works have attempted to post-hoc mitigate biases in KGEs. Given that pre-existing biases are hard to eradicate from KGs, manipulating embedding procedures, may alleviate the issue at least on a representation level. In the following, we summarize respective approaches.

4.7.1 Data Balancing

Radstok et al. (2021) explored the effects of training an embedding model on a gender-balanced subset of Wikidata triples. First, the authors worked with the originally gender-imbalanced Wikidata12k (Leblay and Chekol, 2018; Dasgupta et al., 2018) and DBpedia15k (Sun et al., 2017) on which they fit a TransE and a DistMult model (Yang et al., 2015). They then added more female triples from the Wikidata/DBpedia graph to even out the binary gender distribution among the top-5 most common occupations. Through link prediction, they compared the number of male and female predictions with the ground truth frequencies. More female entities were predicted after the data balancing intervention. However,

the absolute difference between the female ratios in the data and the predictions increased, causing the model to be less accurate and fair. Moreover, the authors note that this process is not scalable since for some domains there are no or only a limited amount of female entities (e.g., female U.S. presidents do not exist in Wikidata).

Du et al. (2022) experimented with adding and removing triples to gender-balance a Freebase subset (Bordes et al., 2015). For the first approach, the authors added synthetic triples (as opposed to real entities from another source as was done by Radstok et al. (2021)) for occupations with a higher male ratio. The resulting bias change was inconsistent across occupations. This appears in line with the authors' finding that ground truth gender ratios are not perfectly predictive of downstream task bias (Section 4.6). For the second strategy, the triples that most strongly influenced an existing bias were determined and removed. This outperformed random triple removal.

4.7.2 Adversarial Learning

Adversarial learning for model fairness aims to prevent prediction of a specific personal attribute from a person's entity embedding. As an adversarial loss, Fisher et al. (2020a) used the KL-divergence between the link prediction score distribution and an idealized target distribution. For example, for an even target score distribution for a set of religions, the model is incentivized to give each of them equal probability. However, in their experiments, this treatment failed to remove the targeted bias fully. This is likely caused by related information encoded in the embedding that is able to inform the same bias.

Arduini et al. (2020) used a Filtering Adversarial Network (FAN) with a filter and a discriminator module. The filter intends to remove sensitive attribute information from the input, while the discriminator tries to predict the sensitive attribute from the output. Both modules were separately pre-trained (filter as an identity mapper of the embedding and discriminator as a gender predictor) and then jointly trained as adversaries. In their experiments, the gender classification accuracy for high- and low-degree entities was close to random for the filtered embeddings (TransH and TransD). For an additional occupation classifier, accuracy remained unaffected after treatment.

4.7.3 Hard Debiasing

Bourli and Pitoura (2020) propose applying the projection-based approach explained in Section 4.5.2 for the debiasing of TransE occupation embeddings. To achieve this, its linear projection onto the previously computed gender direction is subtracted from the occupation embedding. A variant of this technique ("soft" debiasing) aims to preserve some degree of gender information by applying a weight $0 < \lambda < 1$ to the projection value before subtraction. In the authors' experiments, the correlation between gender and occupation was effectively removed — as indicated by the projection measure (Bourli and Pitoura, 2020). However, the debiasing degree determined by λ was found to be in trade-off with model accuracy. This technique was closely adapted from Bolukbasi et al. (2016), regarding which Gonen and Goldberg (2019) criticize that gender bias is only

reduced according to their specific measure and not the "complete manifestation of this bias".

4.8 Discussion

In this article, we cover a wide range of evidence for harmful biases at different stages during the lifecycle of "facts" as represented in KGs. Some of the most influential graphs misrepresent *the world as it is* due to sampling and algorithmic biases at the creation step. Pre-existing biases are exaggerated in these representations. Embedding models learn to encode the same or further amplified versions of these biases. Since the training of high-quality embeddings is costly, they are, in practice, pre-trained once and afterward reused and fine-tuned for different systems. These systems preserve the inherited biases over long periods, exacerbating the issue further. Our survey shows that KGs may qualify as resources for historic facts, but they do not qualify for inference regarding various human attributes. Future work on biases in KGs and KGEs should aim for improvement in the following areas:

Attribute and Seed Choices Bias metrics usually examine one or a few specific attributes (e.g., occupation) and their correlations with selected seed dimensions (e.g., gender). Occupation is by far the most researched attribute in the articles we found (Arduini et al., 2020; Radstok et al., 2021; Bourli and Pitoura, 2020; Fisher et al., 2020a; Fisher et al., 2020b). Only Keidar et al. (2021) propose to aggregate the correlations between a set of seed dimensions and all relations in a graph. All the works used binary gender as the seed dimension and some additionally addressed ethnicity, religion, and nationality (Fisher et al., 2020a; Fisher et al., 2020b; Mehrabi et al., 2021b).

Lack of Validation Most of the KGE bias metrics presented here are interpreted as valid if they detect unfairly discriminating association patterns that intuitively align with existing stereotypes. Besides that, several works investigate the comparability between different metrics. Although both of these practices deliver valuable information on validity, they largely ignore the societal context. Only Du et al. (2022) compared embedding-level bias metrics with census-aligned data to assess compatibility with real-world inequalities. We suggest that future work consider a more comprehensive study of *construct validity* (Does the measurement instrument measure the construct in a meaningful and useful capacity?) (Jacobs and Wallach, 2021). One requirement is that the obtained measurements capture all relevant aspects of the construct the instrument claims to measure. That is, a gender bias measure must measure all relevant aspects of gender bias (Stanczak and Augenstein, 2021) (including, e.g., nonbinary gender and a distinction between benevolent and hostile forms of sexist stereotyping (Glick and Fiske, 1997)). Unless proven otherwise, we must be skeptical that this is achieved by existing approaches (Gonen and Goldberg, 2019). As a result of minimal validation, detailed interpretation guidelines are generally not provided. Therefore, the distinctions between strong and weak bias or weak bias and random variation are mostly vague.

(In-)Effectiveness of Mitigation Strategies Data balancing is the most intuitive approach to bias mitigation and was proven to be effective in the context of text processing (Meade et al., 2022). However, for KGEs, data balancing methods were found to inconsistently reduce bias (Section 4.7.1). Adversarial learning yielded promising outcomes in the study by Arduini et al. (2020). Their FAN approach does not rely on pre-specified attributes. This is in contrast to Fisher et al. (2020a), whose intervention was found to miss non-targeted, yet bias-related information. This problem relates to one of the main criticisms of hard and soft debiasing: instead of alleviating the problem, these techniques risk concealing the full extent of the bias (Gonen and Goldberg, 2019).

Reported Motivations Many, yet not all works in the field name potential social harms as a motivator for their research on social bias in KGs (Mehrabi et al., 2021b; Fisher et al., 2020a; Fisher et al., 2020b; Radstok et al., 2021). Only Mehrabi et al. (2021b) drew from established taxonomies and targeted biases associated with *representational harms* (Barocas et al., as cited in, Blodgett et al., 2020). Similarly, most works lack a clear working definition of social bias. For example, aspects of pre-existing societal biases captured in the data and biases arising through the algorithm (Friedman and Nissenbaum, 1996) are usually not disentangled. Only Bourli and Pitoura (2020) compared model bias to the original KG frequencies and showed that the statistical modeling caused an amplification.

4.9 Recommendations

To avoid harms caused by biases in KGs and their embeddings, we identify and recommend several actions for practitioners and researchers.

Transparency and Accountability KGs should by default be published with bias-sensitive documentation to facilitate transparency and accountability regarding potential risks. *Data Statements* (Bender and Friedman, 2018) report curation criteria, language variety, demographics of the data authors and annotators, relevant indicators of context, quality, and provenance. *Datasheets for Datasets* (Gebru et al., 2021) additionally state motivation, composition, preparation, distribution, and maintenance. The associated questionnaire can accompany the dataset creation process to avoid risks early on. Especially in the case of ongoing crowd-sourcing efforts for encyclopedic KGs the demographic background of contributors should be reported (Demartini, 2019). Researchers using subsets of these KGs, should investigate respective data dumps for potential biases and report limitations transparently. Similarly, KG embedding models should be published with *Model Cards* (Mitchell et al., 2019) documenting intended use, underlying data, ethical considerations, and limitations. Stating the contact details for reporting problems and concerns establishes accountability (Mitchell et al., 2019; Gebru et al., 2021).

Improving Representativeness To tackle selection bias, data collection should aim to employ authors and annotators from diverse social groups and with varied cultural imprints. Annotations should be determined via aggregation (see Hovy and Prabhumoye, 2021). For open editable KGs, interventions like *edit-a-thons*

are helpful to introduce more authors from underrepresented groups (Vetter et al., 2022) (e.g., the Art+Feminism campaign aims to fill the gender gap in Wikimedia knowledge bases[8]). In order for such interventions to take effect, research must update data bases and benchmarks frequently (see Koch et al., 2021a). In addition, the timeliness of encyclopedic data is necessary to avoid perpetuating historic biases.

Tackling Algorithmic Bias Evaluation and prevention of harmful biases must become part of the development pipeline (Stanczak and Augenstein, 2021). Algorithmic biases are best evaluated with a combination of multiple quantitative (Section 4.5) and qualitative measures (Kraft et al., 2022; Dev et al., 2021), considering multiple demographic dimensions (beyond gender and occupation). Evaluating the content of attributions in light of social discourse and the intended use of a technology facilitates an assessment of potential harms (Selbst et al., 2019). Downstream task bias may exist independently from a measured embedding bias (Goldfarb-Tarrant et al., 2021), therefore a task- and context-oriented evaluation is preferred (Section 4.6). We have presented several bias-mitigating strategies for different KGEs, which might alleviate the issue in some cases (Section 4.7). However, more research is needed to establish more effective and robust mitigation methods, as well as metrics used to evaluate their impact (Gonen and Goldberg, 2019; Blodgett et al., 2020).

4.10 Related Work

Although a wide range of surveys investigates biases in NLP, none of them addresses KG-based methods, in particular. Blodgett et al. (2020) critically investigated the theoretical foundation of works analyzing bias in NLP. The authors claim that most works lack a clear taxonomy. We came to a similar conclusion with respect to evaluations of KGs and their embeddings. Sun et al. (2019a) and Stanczak and Augenstein (2021) surveyed algorithmic measurement and mitigation strategies for gender bias in NLP. Sheng et al. (2021) summarized approaches for the measurement and mitigation of bias in generative language models. Some of the methods presented earlier are derived from works discussed in these surveys and adapted to the constraints of KG embeddings (e.g., Bourli and Pitoura (2020) adapted hard debiasing (Bolukbasi et al., 2016)). Criticisms point to the monolingual focus on the English language, the predominant assumption of a gender binary, and a lack of interdisciplinary collaboration.

Shah et al. (2020) identified four sources of predictive biases: *label bias* (label distributions are imbalanced and erroneous regarding certain demographics), *selection bias* (the data sample is not representative of the real world distribution), *semantic bias/input representation bias* (e.g., feature creation with biased embeddings), and *overamplification* through the predictive model (slight differences between human attributes are overemphasized by the model). All of these factors are reflected in the lifecycle as discussed in this article. To counter the risks, Shah et al. (2020) suggest employing multiple annotators and methods of aggregation (see also Hovy

8. https://outreachdashboard.wmflabs.org/campaigns/artfeminism_2022/overview

and Prabhumoye, 2021), re-stratification, re-weighting, or data augmentation, debiasing of models, and, finally, standardized data and model documentation.

4.11 Conclusion and Paths Forward

Our survey shows that biases affect KGs at different stages of their lifecycle. Social biases enter KGs in various ways at the creation step (e.g., through crowd-sourcing of triples and ontologies) and manifest in popular graphs, like DBpedia (Beytía et al., 2022) or ConceptNet (Mehrabi et al., 2021b). Embedding models can capture exaggerated versions of these biases (Bourli and Pitoura, 2020), which finally propagate downstream (Keidar et al., 2021). We acknowledge that KGs have enormous potential for a variety of knowledge-driven downstream applications (Martínez-Rodríguez et al., 2020; Ngomo et al., 2021; Jiang and Usbeck, 2022) and improvements in the truthfulness of statistical models (Athreya et al., 2018; Rony et al., 2022). Yet, although KGs are factual about historic instances, they also perpetuate historically emerging social inequalities. Thus, ethical implications must be considered when developing or reusing these technologies.

We showed that most embedding-based measurement approaches for bias are still restricted to a limited number of demographic seeds and attributes. Furthermore, their alignment with social bias as a construct is not sufficiently validated. Some debiasing strategies appear effective within rather narrow definitions of bias. More in-depth scrutiny is required for a broader understanding of bias. Future work should be grounded in an investigation of concepts like gender or ethnic bias and strive for more comprehensive operationalizations and validation studies. Finally, the motivations and conceptualizations should be communicated clearly.

Acknowledgments

We acknowledge the financial support from the Federal Ministry for Economic Affairs and Energy of Germany in the project CoyPu (project number 01MK21007[G]) and the German Research Foundation in the project NFDI4DS (project number 460234259).

5

Knowledge-Enhanced Language Models Are Not Bias-Proof: Situated Knowledge and Epistemic Injustice in AI

The factual inaccuracies ("hallucinations") of large language models have recently inspired more research on knowledge-enhanced language modeling approaches. These are often assumed to enhance the overall trustworthiness and objectivity of language models. Meanwhile, the issue of bias is usually only mentioned as a limitation of statistical representations. This dissociation of knowledge-enhancement and bias is in line with previous research on AI engineers' assumptions about knowledge, indicating that knowledge is commonly understood as objective and value-neutral by this community. We argue that claims and practices by actors of the field still reflect this underlying conception of knowledge. We contrast this assumption with literature from social and, in particular, feminist epistemology, which argues that the idea of a universal disembodied knower is blind to the reality of knowledge practices and seriously challenges claims of "objective" or "neutral" knowledge.

Knowledge enhancement techniques commonly use Wikidata and Wikipedia as their sources for knowledge, due to their large scales, public accessibility, and assumed trustworthiness. In this work, they serve as a case study for the influence of the social setting and the identity of knowers on epistemic processes. Indeed, the communities behind Wikidata and Wikipedia are known to be male-dominated and many instances of hostile behavior have been reported in the past decade. In effect, the contents of these knowledge bases are highly biased. It is therefore doubtful that these knowledge bases would contribute to bias reduction. In fact, our empirical evaluations of

RoBERTa, KEPLER, and CoLAKE, demonstrate that knowledge enhancement may not live up to the hopes of increased objectivity. In our study, the average probability for stereotypical associations was preserved on two out of three metrics and performance-related gender gaps on knowledge-driven task were also preserved.

We build on these results and critical literature to argue that the label of "knowledge" and commonly held beliefs about it can obscure the harm that is still done to marginalized groups. Knowledge enhancement is at risk of perpetuating epistemic injustice, and AI engineers' understanding of knowledge as objective *per se* conceals this injustice. Finally, to get closer to trustworthy language models, we need to rethink knowledge in AI and aim for an agenda of diversification and scrutiny by outgroup members.

Publication Reference: Angelie Kraft and Eloïse Soulier. 2024. Knowledge-Enhanced Language Models Are Not Bias-Proof: Situated Knowledge and Epistemic Injustice in AI. *In Proceedings of the 2024 Association for Computing Machinery Conference on Fairness, Accountability, and Transparency (FAccT 2024)*, 1433–1445. Rio de Janeiro, Brazil: Association for Computing Machinery.

Contents

5.1 Introduction

One of the currently most discussed limitations of large language models (LLMs) is their tendency to produce false statements (Ji et al., 2023). While LLMs are capable of generating text with great fidelity to linguistic rules (Li et al., 2021a), they frequently produce errors by associating events with the wrong dates or fabricating claims about real people, for instance.[1] Such errors can yield negative impacts on society. It can affect the integrity of science and education (Mittelstadt et al., 2023) or influence the outcomes of democratic elections, by producing false claims about political candidates (Romano et al., 2023) and thus misleading voters.

This lack of factual accuracy[2] is commonly attributed to the implicitness with which knowledge is stored in language models (LMs) and has sparked new interest in ways to enhance LMs with explicit information from external sources, like knowledge graphs (Pan et al., 2024; Agrawal et al., 2024) or informative text documents (Lewis et al., 2020). The idea behind *knowledge-enhanced language modeling* is to fuse representations such that the linguistic capabilities are maintained and factual information from external resources is incorporated accurately (Pan et al., 2023). This is achieved through architectural, training, or inference-related adjustments of the LM (Pan et al., 2024). Respective publications convey that knowledge bases are highly trusted by artificial intelligence (AI) engineers (e.g., Agrawal et al., 2024; Agarwal et al., 2021; Yang et al., 2024b; Pan et al., 2024), which might be explained by a long-standing trust in the objectivity[3] and neutrality of knowledge[4] itself (Forsythe, 1993), in line with traditional theories of knowledge (Adam, 2000). Drawing from previous literature, we argue that this understanding of knowledge fails to acknowledge the influence of the social situation and power of those involved in the creation and sharing of knowledge and that it feeds into knowledge-related injustice (Adam, 2000).

A contribution of this interdisciplinary work is to illustrate some of the related discourse within philosophy and, on this basis, question the prevalent assumptions about knowledge in the AI community. We discuss how dominant conceptions may disguise biases, and, as a consequence, perpetuate injustices. By that, we

1. https://www.zdnet.com/article/chatgpts-hallucination-just-got-openai-sued-heres-what-happened/

2. Factual inaccuracies or false statements produced by language models are often referred to as "hallucinations". We reject this term as it falsely implies a similarity of such models to the human mind.

3. Here objective is understood as subject-independent. The remaining of the paper elaborate on the necessity to challenge this understanding of objectivity.

4. In using the term "knowledge" throughout this article, we are aware of the abysmal amount of ink that has been spilled over this term, and of the differences that exist between disciplines and within epistemology as to what it encompasses. We understand knowledge here as content - not as a cognitive state - and as propositional. Although the distinction between propositional knowledge and knowledge-how and its consequences for knowledge databases are certainly relevant to this discussion, they are out of the scope of this article. We do not consider it crucial either in the context of this paper to draw a distinction between scientific and common knowledge, as we believe that it does not significantly affect our argument.

aim to motivate a rethinking of knowledge as *situated* and to emphasize the necessity for diversification.

In Section 5.2, we discuss the evolution of the approach to knowledge from traditional (Western) epistemology to social and feminist epistemology. The latter coined the concept of *situated knowledge* (Haraway, 2016), which emphasizes the importance of social situatedness to practices of knowledge. We compare this philosophical discussion to AI engineers' conceptions of knowledge and argue that the pervasive understanding of knowledge as objective and value-neutral may disguise the power dynamics that structure knowledge production (Adam, 2000). Publications about knowledge-enhanced language modeling usually mention the risk of bias as a distinguishing property of statistical representations (Agrawal et al., 2024; Agarwal et al., 2021; Yang et al., 2024b), implying that explicit knowledge is not susceptible to bias. This depiction can be misleading: In Section 5.3, we discuss empirical evidence for biases of popular knowledge resources and knowledge-enhanced language models. We particularly focus on Wikimedia Foundation's knowledge bases Wikipedia[5] and Wikidata (Vrandečić and Krötzsch, 2014), which play a major role in language model training and knowledge enhancement and were shown to exhibit coverage gaps and stereotypical biases along different social dimensions (Sun and Peng, 2021; Das et al., 2023; Shaik et al., 2021). We found that knowledge-enhanced language modeling on the basis of Wikidata preserves the biases of the original language model. We maintain that knowledge sources and knowledge-enhanced language models should not *per se* be expected to be less biased than other datasets and AI models. In Section 5.4, we argue that trusting "knowledge data" more than other types of data may wrongfully disguise these issues and contributes to perpetuate the specific kind of injustice that Miranda Fricker has dubbed *epistemic injustice* (Fricker, 2007), that is, a kind of injustice that harms us specifically as knowers. Including more diverse voices is not only a way to tackle these injustices but also the only way we may strive towards objectivity (Longino, 1990; Harding, 2013).

5.2 Assumptions About Knowledge in AI

In this paper, we argue that AI engineers commonly assume knowledge to be subject-independent, which corresponds to more traditional philosophical theories of knowledge. To this end, we start by briefly sketching the evolution from traditional Western epistemology and the figure of the universal knower, to recent approaches from social and feminist epistemology, which emphasize the central role of the social situation of the knower. Finally, we detail how these philosophical theories map to the conceptions of knowledge held by AI engineers and presumably influence modern-day research and practices related to knowledge in AI.

5. https://www.wikipedia.org/

5.2.1 Philosophical Roots of the "View from Nowhere" and Critique

Traditional Western Philosophy

The idea that knowledge could depend on the identity and social situation of the knower has only relatively recently been theorized in Western philosophy. Traditionally, Western epistemology[6] has seen knowledge as a relationship between an individual knower and an object of knowledge, and concentrated its efforts on characterizing this relationship of knowledge, theorizing what distinguishes knowledge from non-knowledge. This distinction often has to do with justification: A belief or perception only becomes knowledge with proper justification. In fact, in analytic philosophy, knowledge is often defined as "justified true belief" (Steup and Neta, 2024) and the justification problem phrased as "S knows that p when [relevant justification]", where S is a single undetermined knower (Adam, 2000). What constitutes proper justification is part of the philosophical debate, but justification is often considered valid only if internal: For example, Descartes considers knowledge coming from others as unreliable (Descartes, 2012). This is in line with the general representation in Western philosophy, usually associated with figures of the Enlightenment such as Kant, that mature thinking and knowing is about autonomy (Kant, 2013). In this perspective, knowledge is acquired independently and rationally, it is universal, independent from the knower's embodied identity, social situation and interests. In Sandra Harding's (critical) words: "In order to achieve the status of knowledge, beliefs are supposed to break free of – to transcend – their original ties to local, historical interests, values, and agendas“ (Harding, 2013, p. 438).

Feminist and Social Epistemology

In the last decades, feminist and social epistemology have challenged this traditional approach to knowledge, arguing that knowers are always socially situated, and that this social situation mattered to the kind of knowledge they could produce. Social epistemologists have emphasized that the production of knowledge is an inescapably social activity (Longino, 2002). In John Hardwig's terms, we are epistemically dependent: *pace* Descarte's ideal of the independent knower, we cannot but rely on others' testimony to know most of what we know, even in scientific contexts where the standards on what counts as knowledge are taken to be higher (Hardwig, 1985). If knowledge necessarily involves relying on other's testimony, then power dynamics within society are relevant to the production and dissemination of knowledge (Fricker, 2007) and to the possibility to accept a claim as knowledge (Scheman, 2015). Indeed, these power dynamics determine whose knowledge will be heard. We detail in Section 5.4 the ways in which this can lead to injustices.

6. Characterizing and summarizing "traditional Western philosophy" in one paragraph is a difficult task, considering that what is usually refereed to as "Western thought" is itself a Western post hoc construction. What we mean here is a conceptual framework considered to have crystallized during the Enlightenment, which has been significantly challenged in the last half century by critical theories.

Feminist standpoint theorists have argued that we are limited in what we can know by our social situation, and "some social situations – critically unexamined dominant ones – are more limiting than others in this respect" (Harding, 2013, p. 443). In other words, we are particularly constrained in what we are able to know when our social situation is dominant, and therefore seldom questioned. The "view from nowhere" (Nagel, 1989) supposed to characterize objectivity, in Haraway's words, actually "signifies the unmarked positions of Man and White" (Haraway, 2016, p. 581). Different feminist approaches[7] disagree on the extent to which we are epistemically limited by our social situation, and the depth to which scientific frameworks should be questioned. We leave the detail of these discussions out of this short account, as we do not believe it is necessary to take sides in order to draw from these different theorists for the problem at hand. Note that related arguments have been made by decolonial epistemologists: These scholars have emphasized the geopolitical situation of knowledge under the persistent regime of coloniality (Pitts, 2017), and the necessity for subjects of colonial oppression to think not only from their perspective, but outside of Western epistemic resources (Grosfoguel, 2007; Pitts, 2017). We give this account of the evolution of the field of epistemology, as we consider it reasonable to assume that the influence of modern epistemology still has a bearing on contemporary conceptions of knowledge. In the following we focus on the group of AI engineers, as they are the relevant category to the object of this article, but we do not believe these representations to be limited to this group.

5.2.2 AI Engineers and the "View from Nowhere"

Forsythe's Anthropological Study

Three decades ago, in 1993, Diana E. Forsythe published one of the first in-depth investigations of AI engineers'[8] conceptions of knowledge (Forsythe, 1993). She had observed and interviewed a group of engineers whose task it was to elicit the knowledge of domain experts and translate it into a machine-readable representation for use in AI systems. Back then, it was already envisioned that AI would at some point "duplicate human expertise" (Forsythe, 1993, p. 1), i.e., that AI systems would gain the same capabilities that humans have. Without more critical scrutiny of what constitutes knowledge, the AI engineers in Forythe's study described it as universal, a constant that does not change with context, is purely cognitive and conscious in nature. Forsythe (1993) also mentions the ways in which AI engineers' assumptions differ from those held by social scientists. The latter believe knowledge to be a problematic subject of research that is highly dependent on social and otherwise contextual factors. They consider a lot of what people know to be tacit and unaligned to their actions. This gives rise to a wide range of methodological principles, each of them designed to elicit knowledge from humans while respecting its social and non-objective nature.

7. For a detail of the different approaches in feminist philosophy of science, see e.g. (Anderson, 2024)

8. Forsythe (1993) uses the term "knowledge engineer" to designate the participants' profession. However, as they are described as researching and developing (symbolic) AI technology, we instead use the term "AI engineer" for the sake of consistency.

Adam's Epistemological Analysis

In "Deleting the Subject: A Feminist Reading of Epistemology in Artificial Intelligence", Alison Adam (Adam, 2000) compares AI engineers' beliefs to the traditional Western take on knowledge (see Section 5.2.1). She points out that AI systems are built on the assumption of knowledge as a universal "view from nowhere" (as introduced by Nagel, 1989) and thereby dismiss the importance of the identity of the knower. She argues that this effectively obscures an "implicit hierarchy of knowers", i.e., the power dynamics which grant a specific demographic the privilege to represent its knowledge in AI systems and others not. Following an analysis of the Cyc commonsense[9] knowledge base,[10] Adam (2000) formulates two main points of criticism: Firstly, the system did not allow to represent contradictory information and, thus, could only represent one world view at a time. She explains this with the presumably pervasive understanding of AI engineers "that there is an independent world that can be accessed through perception and also that everyone will agree on what the real world is like" (Adam, 2000, p. 241). Again, this understanding disregards that individual knowers are limited in how they view the world (by their identity and situation), which means that different perceptions of the world co-exist. Her second point of criticism relates to the underlying hierarchy of knowers: Ultimately, whose knowledge would be considered the right one was determined only by the developers of Cyc, whose demographic was described as the "middle-class, Western, professional man" (Adam, 2000, p. 241). Again, including their knowledge exclusively in a system like Cyc is to certify it as more legitimate than other knowledges.[11]

Understanding Modern Conceptions

The dominance of the "view from nowhere" and its harmful consequences are still frequently discussed in the context of modern Machine Learning and AI (Lindemann, 2024; Keyes and Creel, 2022; Hancox-Li and Kumar, 2021; Gebru, 2021). The current discourse on the capabilities of AI indicate that engineers pre-dominantly focus on the technical challenges of knowledge extraction from data (Martínez-Rodríguez et al., 2020), benchmarking the knowledge of AI models[12] (Youssef et al.,

9. Knowledge regarding everyday situations and cause-effect relationships.

10. https://cyc.com/

11. Adam (2000) uses the terminology by Foley (1987) here, which distinguishes between "non-weird" and "weird" knowledge.

12. "Artificial intelligence" has been, ever since the expression appeared in the 50s, associated with an anthropomorphic aim to replicate human capabilities. Even though the term is currently often associated with strictly technical definitions (for example, the definition that will most likely figure in the upcoming European AI Act: https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf), it remains a common way of understanding "artificial intelligence". In the Google campaign, their Knowledge Graph was seen as a step towards "building the next generation of search, which [...] understands the world a bit more like people do."(https://blog.google/products/search/introducing-knowledge-graph-things-not/). With the recent development of sophisticated AI systems, researchers in the philosophy of AI have been inquiring the ways in which concepts so far exclusively applied to humans and some other animals could be extended to AIs in a non-metaphorical sense. These reflections include whether an AI can "know" (Burns et al., 2023), or "believe" (Rathkopf, 2023) but also "love" (Nyholm and Frank, 2017) or exert "agency" (Floridi and Sanders, 2004). We are not concerned with these questions in this article. When we talk about what a LM knows, we mean metaphorically which – and importantly

2023; Roberts et al., 2020; Kadavath et al., 2022), and ways to embed more of it (Pan et al., 2024). Yet, the provenance of this knowledge remains largely unexamined. In a review on AI throughout history, Jiang et al. (2022) claim that "[k]nowledge describes regular patterns and abstract facts that human understands [sic]" [, p. 9] and thereby attribute universality to knowledge. The authors continue by stating that, "[t]herefore, it is usually semantic and embedded in books and research articles. To be interpretable and useful for machines, it needs to be modelled, transformed, and generated" (Jiang et al., 2022, p. 9). This quote refers to automated knowledge acquisition approaches, which are widely established. It points to an understanding of knowledge as subject-independent and is similar to the beliefs held by Forsythe's participants, who had desired exactly this kind of automation to avoid having "to mine those jewels of knowledge out of their heads one by one" (Forsythe, 1993, p. 454). In his vision paper, Marcus (2020) argues that the next decade in AI should focus on "a hybrid, knowledge-driven, reasoning-based approach, centered around cognitive models, that could provide the substrate for a richer, more robust AI than is currently possible"[13] [p. 1]. Without addressing the social conditions under which knowledge resources are created, he claims that having more of it embedded in AI models will make these models more robust. LeCun predicts that AI will become a "repository of all human knowledge", claiming that such a repository would be the "ultimate solution *against* misinformation."[14] He, however, emphasizes that automation alone will not suffice and instead proposes Wikipedia-style crowd-sourcing, implying that the more people contribute, the closer we will get to a representation of the sum of all knowledge.[15] As we will discuss in more detail in Section 5.3.3, Wikipedia, in fact, clearly exemplifies that crowd-sourcing processes are not immune to the influence of social power structures without appropriate countermeasures. While we agree on the importance of improving the factual accuracy of AI systems and on the value of crowd-sourcing as a basis for this, we believe that a more nuanced understanding of knowledge is needed to come closer to just and objective knowledge production in the long term.

5.3 Connecting the Debates on Knowledge Enhancement and Social Bias

In the following, we take a closer look at the bias issue in Wikimedia knowledge bases to exemplify the influence of the social setting on collective epistemic processes. To this end, we firstly explain the idea behind knowledge-enhanced language models. We then develop the connection between knowledge enhancement

whose – knowledge it embeds, not in which sense it might be said to know something itself. This is not to say that this question is irrelevant to our main concern, as it seems possible that anthropomorphizing the AI itself might further contribute to the disappearing of the original subject of knowledge. But the role of this effect is beyond the scope of this article.

13. Hybrid AI, here, refers to a combination of symbolic representations of knowledge with modern statistical approaches and is similar to the idea of knowledge enhancement discussed earlier.

14. https://twitter.com/ylecun/status/1664681619335020560

15. https://twitter.com/ylecun/status/1713751182601015729

and social bias and later detail the representation issues in Wikimedia knowledge bases. Finally, we demonstrate how the biases of said knowledge bases can be adopted by technology. We do this at the example of language models enhanced with knowledge from Wikidata.

5.3.1 Knowledge Enhancement and the Dichotomy of Explicit and Implicit Knowledge in AI

Hybrid AI systems or knowledge-enhanced models are attempts to combine the strengths of statistical AI and explicit representations of knowledge. *Statistical AI* subsumes approaches that model patterns and rules implicitly from (large-scale) data sets, instead of following hard-coded rules. Such approaches allow to process enormous amounts of information with minimal human involvement (compared to mostly manually created *symbolic* systems) and are more generalizable to new areas and tasks (Pan et al., 2024). Statistical AI is the currently dominating paradigm and AI-based language models are part of this category (Jurafsky and Martin, 2009). One limitation of these approaches it that the knowledge represented can no longer be accessed directly and can only be interpreted and quantified through dedicated decoding procedures (Petroni et al., 2019; Youssef et al., 2023).

The effort to represent explicit knowledge content in machine- and human-readable form and perform inference based on hard-coded rules is commonly denoted *symbolic AI*, which was the most prominent AI paradigm for most of the second half of the 20th century. Knowledge graphs (KGs) are a type of symbolic representation that are still used to represent the semantic relationships between things in the world across various topical domains. A KG is a graph where each triple describes the relationship between real-world entities in the form *(head, relation, tail)* (Paulheim, 2017). A KG-specific ontology defines the possible classes of entities, their attributes, and properties. The graph-based structure allows for efficient machine processing, is human-readable, and transparent.

Since statistical LMs always output the most likely next word, they may generate results that seem linguistically sound, even when the content is not accurate or appropriate (Ji et al., 2023). This phenomenon is frequently observed, since the large-scale web-scraped datasets that LMs are trained on usually contain false information, inaccuracies, and gaps. In other cases, the perceived input may lack important contextual information for the model to produce contextually accurate results. To tackle this shortcoming, explicit, relevant, fine-grained knowledge can be incorporated (Agrawal et al., 2024). A large variety of knowledge enhancement approaches exist to implement this idea. For example, the mention of an entity (a person, a place, an event, etc.) may be combined with additional background information during model training, so that an enriched representation of the entity is learned (Wang et al., 2021b; Sun et al., 2020). Another common approach is to give the model access to an external knowledge base to retrieve relevant information from during runtime (Lewis et al., 2020).

5.3.2 Why We Need to Talk About Knowledge Enhancement and Social Bias

Social bias is observed when language models "*systematically* and *unfairly* discriminate against certain individuals or groups of individuals in favor of others" (Friedman and Nissenbaum, 1996, p. 332). It takes form in reproduced stereotypes (Nadeem et al., 2021), negative valuations of groups (Sheng et al., 2019), or systematic performance differences based on sensitive attributes (Kiritchenko and Mohammad, 2018; Dev et al., 2022). Social bias is another widely discussed limitation of language models (Sun et al., 2019a; Liang et al., 2021; Bender et al., 2021). Both social bias and factual inaccuracies are considered obstacles to the trustworthiness of LMs (Mallen et al., 2023; Wang et al., 2023) but are usually investigated in isolation to each other. Factual inaccuracies are countered by adding knowledge, i.e., data that represent facts about things in the world, while social bias is tackled, e.g., through data balancing, manipulation of the embedding space, or constraining the predictions (Sun et al., 2019b). It is at times implied that enhancing the factual accuracy of LMs through knowledge enhancement could positively impact bias issues in the same instance, since knowledge is highly trusted and curated.[16, 17] This corresponds to our observation that, in the context of knowledge-enhanced language modeling, the issue of bias is usually only mentioned as a limitation of statistical AI and its unstructured training databases (Agrawal et al., 2024; Agarwal et al., 2021; Yang et al., 2024b).[18] The fact that highly curated and structured KGs, like Wikidata and DBpedia, reproduce the same societal biases mostly goes unmentioned (Kraft and Usbeck, 2022). This omission is unjustified and potentially harmful. That is, misconceiving of knowledge as objective and an antithesis to bias, value judgements, and uncertainty, grants anything under the label of knowledge potentially undeserved legitimacy. In fact, it gives undeserved legitimacy to the interests, assumptions and world views of a privileged group. In the case of both the work of Adam (2000) and the KGs discussed here, this is predominantly the group of educated Western men (Kraft and Usbeck, 2022).

In the next section, we summarize representation-related issues in Wikidata and Wikipedia, which are examples of crowd-sourced knowledge bases. As mentioned before, the creation or extension of knowledge graphs is also oftentimes based on or supported by automated processing (Schneider et al., 2022), e.g., through automatic knowledge extraction (Martínez-Rodríguez et al., 2020) and knowledge integration (Möller et al., 2022). Other works are even inspecting the possibility to extract knowledge directly from language models to utilize them as knowledge bases (Petroni et al., 2019). It is important to remember here that automatic approaches of course also mirror the values of their developers. Firstly, many of these mentioned natural language processing (NLP) approaches are affected by social biases (Mehrabi et al., 2020; Mishra et al., 2020; Gaut et al., 2020;

16. https://blog.research.google/2021/05/kelm-integrating-knowledge-graphs-with.html
17. https://www.searchenginejournal.com/google-kelm/408151/
18. We found one exception in Lewis et al. (2020), , p. 10, where it is stated that "Wikipedia, or any potential external knowledge source, will probably never be entirely factual and completely devoid of bias [...]" and that "[i]n order to mitigate these risks, AI systems could be employed to fight against misleading content [...]". This suggestion fails to address the real-world source of the problem and instead points in the direction of techno-solutionism (Morozov, 2013).

Du et al., 2022; Keet, 2021). Secondly, they are more frequently applied for the more represented languages. For instance, more bots are used to populate Germany-related content in Wikidata than Vietnam-related content (Ma and Zhang, 2023), further amplifying existing coverage gaps. So, while the automatic creation and extension of knowledge bases may save a lot of time and effort (and avoid potential frustrations caused by social interactions (Forsythe, 1993)), they may amplify biases and further occlude the social conditions of knowledge production.

5.3.3 The Biases of Wikidata and its Hierarchy of Knowers

Most research articles that present new techniques for KG-based enhancement of language models utilize English Wikidata (e.g., Sun et al., 2020; Wang et al., 2021a; Qin et al., 2021; Zhang et al., 2022; Wang et al., 2022), since it is the largest publicly accessible open-domain KG (Wang et al., 2021b). A wide range of non-KG approaches are developed on the basis of Wikipedia, e.g., many Retrieval-Augmented Generation (RAG) approaches (Gao et al., 2023 for an overview). These knowledge bases[19] are more curated and reviewed than most other data sources involved in the training of language models.[20] That is, users populate the knowledge bases collaboratively, engage in discussions on the content, and constantly work on updates and refinements. Agarwal et al. (2021) imply that KGs have less limited coverage of the world knowledge than text corpora. The authors used a dedicated data-to-text model to verbalize all triples in the English Wikidata KG and thereby created a synthetic natural-language corpus called the KELM corpus (Corpus for Knowledge-Enhanced Language Model Pre-training) which is intended for integration with natural language training datasets to improve LM performance on knowledge-intensive tasks. In a blog post, the authors claim that "KGs are factual in nature because the information is usually extracted from more trusted sources, and post-processing filters and human editors ensure inappropriate and incorrect content are removed."[21]

These claims strike us as particularly interesting in the face of prevalent issues with Wikimedia's knowledge bases: Wikidata exhibits significant coverage gaps for different genders (Zhang and Terveen, 2021; Das et al., 2023), races, and citizenships (Shaik et al., 2021). We analyzed Wikidata and the KELM corpus and found that women make up only approximately 20% and other genders make up less than 1% (see Table 5.3 in Appendix 5.8). Representational biases are not only manifested in coverage gaps: Wikidata entries about German personalities are significantly more often edited than entries about Vietnamese personalities (Ma and Zhang, 2023). This indicates that the latter undergo less deliberation and may be less trustworthy (Tollefsen, 2009).[22] The narration style used to describe different

19. In the following, we almost interchangeably address issues regarding Wikipedia and Wikidata. The reason for this is that they are related projects and Wikidata contains all of the factual information from Wikipedia presented as a graph (Vrandečić and Krötzsch, 2014). As both projects are organized as part of the Wikimedia Foundation, they follow similar standards and procedures.

20. https://nytimes.com/2023/07/18/magazine/wikipedia-ai-chatgpt.html

21. https://blog.research.google/2021/05/kelm-integrating-knowledge-graphs-with.html

22. In summary, we may say that the content of Wikipedia and co. is trustworthy on average, while the trustworthiness of individual claims is more difficult to determine (Simon, 2010). Tollefsen (2009) points out that not every content is equally debated and reviewed and claims that the more a piece of content has undergone group deliberation, the more we may be able to trust it.

demographics also differs in stereotypical ways. For example, on Wikipedia, women are more likely to be described with regards to personal life events (even within the "Career" section) than men (Sun and Peng, 2021). Popular KGs like Wikidata use inappropriate and derogatory terms to indicate, e.g., ethnicity, sexual identity or orientation (Nesterov et al., 2024).

The cause of these representation issues can be found in the power hierarchies that characterize the community behind these efforts. Menking and Rosenberg (2021) argue that there is a mismatch between the ideal scenario implied by the Five Pillars of Wikipedia, i.e., the guiding principles, and the reality of its epistemic community. "While anyone can edit Wikipedia, there are several barriers to becoming a Wikipedian. For example, newcomers must learn how to navigate any number of technical, organizational, and social hurdles they encounter when performing a substantial edit." (Menking and Rosenberg, 2021, p. 458). Examples for said social hurdles are manifold: Members of marginalized communities face higher standards for notability, which is an eligibility requirement for coverage in Wikipedia and Wikidata (Tripodi, 2023).[23] Women editors' articles are more likely to be reverted, especially in the early phases of their participation (Lam et al., 2011; Lir, 2021). Editors who identify as women and/or LGBTQIA+ are trolled, harassed, receive death threats, and become victims of *doxxing* (Menking and Erickson, 2015; Menking et al., 2019).[24] Thus, it is not surprising that only 13% of all active Wikimedia editors are women and 4% gender-diverse, according to a 2023 report.[25] The same report also showed that active editors are highly educated – 82% hold at least a post-secondary degree – and most US and UK editors are white (disproportionately more than in the general population). The geographic distribution of editors is skewed towards Western Europe, making up more than 50% (as of 2018).[26]

These observations show how knowledge production is shaped by the situation of the knowers. Their identities and values influence the interactions leading to agreement (or disagreement) on what to consider knowledge. We focused on Wikipedia and Wikidata because they are prevalent resources in NLP research and a lot is known about the communities behind them. However, our criticism extends to other knowledge bases, like DBpedia and Freebase, which exhibit similar gaps (Kraft and Usbeck, 2022).

5.3.4 Knowledge Enhancement Does Not Solve the Bias Issue

Quantitatively, the effect of knowledge enhancement on bias was so far only shown for commonsense knowledge: Melotte et al. (2022) fine-tuned different generative language models – GPT-2 (Radford et al., 2019), T5-base, and T5-small (Raffel et al., 2020) – with commonsense KGs – Wikidata-CS (Ilievski et al., 2020) and ConceptNet (Speer et al., 2017) – to allow the models to predict an object from a given subject-predicate pair (e.g., *("gentleman", "is capable of")*). The authors measured bias regarding *origin*, *gender*, *religion*, and *profession* via classifiers for

23. https://meta.wikimedia.org/wiki/Gender_equity_report_2018/Barriers_to_equity
24. https://www.nytimes.com/2019/04/08/us/wikipedia-harassment-wikimedia-foundation.html
25. https://meta.wikimedia.org/wiki/Community_Insights/Community_Insights_2023_Report
26. https://meta.wikimedia.org/wiki/Community_Insights/2018_Report

sentiment and *regard*, which can identify whether or not an output sequence is a positive or negative portrayal. T5-small tuned on ConceptNet created more-than-average negative depictions of, e.g., "Columbians", "Afghans", and "Indians". Occupations like "teacher", "doctor", and "professor", were more likely depicted in positive ways, whereas "prosecutors" were more often depicted negatively. The results showed an increase of bias with the scale of the KG.

In the following, we present a preliminary analysis of social bias in language models enhanced with encyclopedic knowledge. We evaluated KEPLER (Knowledge Embedding and Pre-trained Language Representation) (Wang et al., 2021b) and CoLAKE (Contextualized Language and Knowledge Embedding) (Sun et al., 2020) in comparison to RoBERTa (Robustly Optimized BERT Pretraining Approach) (Liu et al., 2019).[27] KEPLER and CoLAKE are both modified versions of the popular RoBERTa language model and incorporate Wikidata. More detailed explanations of these models are provided in Appendix 5.9. To validate the knowledge enhancement effect, we compared the performance of the models on a suite of knowledge-intensive evaluation tasks, called the LAMA (LAnguage Model Analysis) probe (Petroni et al., 2019), and present the results and more details on the probe in Appendix 5.10. We investigated two kinds of bias: *stereotypes*, i.e., learned systematic associations between individuals/groups and classes of professions or other attributes, and secondly, *performance differences* on knowledge-related tasks that might arise from imbalanced representation of individuals or groups in the dataset.[28]

Stereotypical Bias Analysis

We use three common stereotype measures to compare the biases across models:[29] *1. SEAT (Sentence Embedding Association Test)* (May et al., 2019; Caliskan et al., 2017) measures the associations between certain demographics and certain attributes, which are often discussed in stereotypical portrayals of said demographics and their respective opposites. The significance of the association is determined via a permutation test and its effect size is interpreted as an indicator of the bias magnitude. Lower effect sizes indicate less bias. *2. CrowS-Pairs (Crowdsourced Stereotype Pairs)* (Nangia et al., 2020) is comprised of crowd-sourced stereotypical descriptions of historically disadvantaged groups in the United States. The test computes the percentage of instances where a stereotypical description is preferred over a less or non-stereotypical description by a given LM. For a random score of 50%, no systematic association is observed and the model is considered unbiased. *3. StereoSet* follows a similar idea (Nadeem et al., 2021) and compares the likelihood of stereotypical, anti-stereotypical, and *unrelated* responses (example: "Girls tend to be more ___ than boys"; response options: "soft" (stereotypical), "determined" (anti-stereotypical), and "fish" (unrelated)). The *idealized context association score*

27. We selected these to models for the following reasons: They introduce little changes to RoBERTa to perform knowledge enhancement and are, thus, easily comparable. Secondly, their model weights are publicly available.

28. Our analysis scripts and data are made available here: https://github.com/krangelie/KE-PLM-bias.

29. Previous literature has shown that bias measures do not always correlate with each other as they measure different facets. Furthermore, there is no established standard measure to date. It is, thus, recommended practice to analyze bias via a combination of measures (Dev et al., 2022).

(ICAT) is a stereotype metric based on the relative number of samples for which the stereotypical is preferred over the anti-stereotypical option, scaled by the model's language modeling capability (percentage of cases, where the model does not opt for the unrelated response).[30]

Table 5.1: Bias metrics for RoBERTa and its knowledge-enhanced variants KEPLER and CoLAKE. Bold scores indicate the most optimal model according to the respective metric. For SEAT, scores closer to 0 are less biased. For CrowS-Pairs, scores closer to 50 are more optimal and for StereoSet, ideal scores are ICAT=100.

		RoBERTa	KEPLER	CoLAKE
SEAT	gender	.940	.789	**.329**
	race	**.307**	.374	.340
	religion	**.127**	.890	.332
	average	*.458*	*.684*	***.334***
CrowS	gender/gender identity	60.15	59.39	**54.41**
	race/color	**63.57**	64.92	64.53
	religion	60.00	**50.48**	58.10
	socioeconomic status/occupation	61.99	**60.23**	66.67
	nationality	47.80	47.80	**44.03**
	age	**49.43**	52.87	55.17
	sexual orientation	63.10	**59.52**	61.90
	physical appearance	**53.97**	57.14	55.56
	disability	67.80	71.19	**66.10**
	average	*58.65*	***58.17***	*58.50*
StereoSet (ICAT)	gender	60.48	68.63	**70.43**
	race	**68.93**	63.96	65.09
	religion	62.89	68.25	**69.81**
	profession	**67.42**	66.06	66.49
	overall	***67.11***	*65.50*	*66.45*

Table 5.1 shows the final bias metrics for all three models. On the SEAT metric, KEPLER and COLAKE yield larger effect sizes than RoBERTa on two out of three bias dimensions, namely race and religion. On the gender bias dimension, CoLAKE outperforms RoBERTa by a large margin, causing CoLAKE to receive the best average score. For CrowS, the models again exhibit different strengths: While RoBERTa is least biased regarding race/color, nationality, age, and physical appearance, KEPLER and CoLAKE exhibit less stereotypical attributions in the case of other dimensions, like gender, religion, sexual orientation, and disability. On average, across all dimensions, all models prefer the stereotypical over the anti-stereotypical option in 58% of the cases. On StereoSet (ICAT), RoBERTa slightly outperforms the knowledge-enhanced models. In conclusion, these inconsistent results indicate that simply adding knowledge to language models does not solve the bias problem. Instead, two of the metrics used, CrowS-Pairs and StereoSet, indicate a preservation of the average probability for stereotypical associations.

Performance Bias Analysis

To investigate the models' biases on a knowledge-intensive task, we performed a disaggregated evaluation on the T-REx (ElSahar et al., 2018) subtask from the

30. The tests were run with the implementations by Meade et al. (2022).

LAMA probe.[31] It consists of cloze-style templates derived from KG triples. For example, the triple *(Dante, born-in, Florence)* would translate to *"Dante was born in ___"* and the model would have to predict *"Florence"* to be correct. The authors assume a language model to "know" a fact if it fills the gap correctly (Petroni et al., 2019). The T-REx subtask is comprised of 600 relations and 11 million triples from Wikidata.[32] We iterated through the entire set of triples and extracted those relating to at least one human entity. We then queried the genders of these entities from our Wikidata dump (October 2022) and split the examples into a male and a female subset. Due to a lack of gender diversity in the dataset (see Table 5.3), only a binary comparison was possible. Per relation, the group-level *Demographic Parity (DP)* metric was calculated via $\text{DP} = \frac{\text{ratio of correct completions of women-related examples}}{\text{ratio of correct completions of men-related examples}}$ (where DP = 1.0 indicates independence of output correctness from subject gender) and then averaged across relations (Barocas et al., 2023; Feldman et al., 2015). Finally, the performance metric used by Petroni et al. (2019), namely the *Mean P@1* scores (average number of cases for which the top-1 most likely response is the correct one) across relations, were computed separately for female and male examples. Table 5.2 shows that all three models exhibit demographic *dis*parity, with gender-based performance gaps roughly equal across models. Despite a slight improvement for KEPLER, these results overall do not indicate a considerable removal of bias after knowledge enhancement.

Table 5.2: Top: Average DP based on the per-relation model accuracy for female versus male subjects. Bottom: T-REx performance (measured via Mean P@1) for male and female subjects.

		RoBERTa	KEPLER	CoLAKE
Mean DP		.41	.55	.44
Mean P@1	female	12.71	13.08	13.76
	male	19.36	18.81	21.01

5.4 How Can We Do Better? Drawing from Philosophical Insights

We used the example of Wikidata because it is a very popular database. Therefore, the biases described should be alerting in themselves. However, we do not expect these issues to be specific to Wikidata. As we have argued in Section 5.2, the conception of knowledge that seems to prevail in the AI community has been the object of philosophical reappraisal. Thus, we consider it fruitful to draw from feminist epistemology to better grasp the ways in which the social dimension of knowledge production in general can lead to injustices, but also how we can strive for better practices.

31. We utilized the evaluation script and data provided here: https://github.com/facebookresearch/LAMA.

32. List of Wikidata relations considered in analysis: place of birth (P19), place of death (P20), country of citizenship (P27), field of work (P101), native language (P103), occupation (P106), employer (P108), position played on team / speciality (P413), work location (P937), languages spoken, written or signed (P1412).

5.4.1 Including More Diverse Voices

The main insight we draw from feminist epistemology is that knowledge production is not immune to the power dynamics that structure society. This is what Miranda Fricker has famously theorized in her 2007 book "Epistemic Injustice, Power and the Ethics of Knowing" (Fricker, 2007). The fact that we are, as knowers, social beings that stand in power relations to each others, Fricker argues, makes knowledge practices the locus of a specific type of injustice: epistemic injustices. Fricker describes epistemic injustice as having two main aspects: testimonial injustice and hermeneutical injustice. *Testimonial injustice* is a consequence of identity prejudice: We usually assign credibility automatically to speakers, and in this unreflective process, identity prejudice can unjustly lead us to grant less credibility to some speakers, typically from marginalized groups. Their contribution is dismissed, and they are harmed in their dignity and their capacity to participate in knowledge production and transmission. *Hermeneutical injustice* has to do with knowledge gaps: Because marginalized groups are less given the ability to participate in knowledge production, because their experiences are less the object of collective interest and study, their experiences and knowledge are not represented in our collective hermeneutical resources. Fricker gives the example of the concept of "sexual harassment", the absence of which long prevented some women from making sense of what they were experiencing. This understanding of hermeneutical injustice has however been nuanced among others by Rebecca Mason (Mason, 2011). To Mason, hermeneutical injustice is not only a matter of marginalized groups not having the hermeneutical resources to articulate their experience, but also of dominant groups willfully, or at least blameworthily ignoring this experience. Dominant groups bear an important responsibility for these "blanks where there should be a name for an experience" (Fricker, 2007, p. 160).

The mechanisms of exclusion from the Wikimedia community described in Section 5.3.3 are arguably examples of testimonial injustice contributing to hermeneutical injustice. Some contributors' testimony is dismissed because of identity prejudice, and this results in gaps in the knowledge resource. As we have shown in Section 5.3.4, feeding such knowledge databases to LMs does not make them objective, but instead embeds these hermeneutical gaps in the technology. Epistemic injustices have to do with the possibility to participate in knowledge production and to be represented in collective resources. Working against these injustices is important to justice and non-discrimination, but it is also crucial for epistemic reasons. However strong a stance one takes on the way our situatedness epistemically limits us, it remains that our knowledge resources are enriched by including diverse contributions, particularly from marginalized groups. This is arguably not the case – yet – for Wikidata or Wikipedia.

Networks like Art+Feminism[33] and FemNetz[34] provide safe spaces for Wikipedia contributors with feminist visions. They organize regular events, e.g. *edit-a-thons*, to improve the platform's coverage of knowledge relevant to all genders and increase the use of inclusive and anti-discriminatory language. These initiatives exemplify how epistemic injustice may be tackled bottom-up. However, against

33. https://artandfeminism.org/
34. https://de.wikipedia.org/wiki/Wikipedia:WikiProjekt_FemNetz

the backdrop of a community dominated by groups who resist the inclusion of certain experiences by violent means, participation can only be realized at high cost (Menking and Erickson, 2015) or sometimes not at all: The founders of the German web encyclopedia Equalpedia initially raised public funds to build an editorial team that would contribute information about women and persons from the LGBTQIA+ community to Wikipedia.[35] But, targeted by *edit wars*,[36] they ultimately failed to prevail against the existing power structures and resorted to building their own platform instead. While institutions and individuals developing AI and respective data corpora should work towards solutions and pro-actively invite underrepresented views, the involvement of diverse voices should be approached in reciprocal and empowering ways (Birhane et al., 2022a). The reality of modern AI is largely determined by powerful technology companies that gather information without consent to their own financial benefit.[37] Especially historically exploited communities should (co-)determine how these resources are created, disseminated, and utilized (Birhane et al., 2022a). Hence, refusal of participation in open access knowledge bases, like Wikipedia, is a legitimate alternative that should be supported, as well. Inclusion should always be approached with the perspective that hermeneutical injustice does not result from innocent knowledge gaps, but is motivated by group interest as an integral part of a pervasive system of social oppression (Mills, 2017). Power dynamics shape discourses and practices of inclusion themselves (Hoffmann, 2021), and we believe that inclusion should be approached critically, and not as the ultimate fix to structural injustice (Browne, 2023).

5.4.2 Reflexivity and Intersubjective Criticism: Objectivity Is Hard Work

Underlying this discussion is the question of whether there can be such a thing as knowledge that would be perspective-independent, and how we can strive for that or towards that goal. Viewpoints within feminist epistemology differ on this matter. However, we believe it is possible to draw some common lessons from them that are useful for AI engineers.

Feminist empiricists like Helen Longino or Elizabeth Anderson have argued that it is inevitable that moral and political values play a role in scientific inquiry (Anderson, 1995; Longino, 1990). They play a role in determining what will be researched, but also with which methods. They influence according to which background theory facts will be interpreted and which facts will be considered significant. What still protects scientific knowledge from arbitrariness and preserves the possibility for objectivity – at least as a horizon– is, Longino says, the possibility for intersubjective criticism of commonly available phenomena and methodologies (Longino, 1990). This supposes among others avenues for criticism, shared standards on the formulation of these criticisms, responsiveness to criticism, and equal intellectual authority among qualified practitioners. And the greater the number of points of view, the closer scientific practice gets to objectivity. In

35. https://www.equalpedia.org/ueber-equalpedia/
36. https://en.wikipedia.org/wiki/Wikipedia:Edit_warring
37. https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/

this sense, we consider interdisciplinary exchange and collaboration essential for critical decentering. In the case we have been discussing, engaging with different disciplines – for example during education (Raji et al., 2021b) – and communities should contribute to fostering a more critical understanding of the concept of knowledge in the AI community.

Standpoint theorists share the conviction that beliefs and values are pervasive in every aspect of knowledge production. However, to them, there is no transcending our situatedness. Instead, it is precisely by theorizing this situatedness of subjects of knowledge and the values that underlie any knowledge-seeking endeavor that we can strive for what Harding calls "strong objectivity", a way to "maximize objectivity" through "strong reflexivity" (Harding, 2013, pp. 460-462). This requires to think broader than the avenues for criticism organized by scientific communities (or any community that claims to create knowledge of some authority, for example a knowledge database). Indeed, the criteria that determine who is qualified to participate and according to which rules, should themselves be subjects of critical scrutiny. And those who are excluded from these groups are better situated to exercise this scrutiny. The consequence is that any claim to produce authoritative knowledge such as knowledge databases should not only imply organized practices of intersubjective criticism, but also actively seek the critical scrutiny of outgroup members.

In the absence of such strong standards, Harding calls objectivity a "mystifying notion", little more than an argument from authority that benefits dominant groups (Harding, 2013). This article argues that in the same way, the term "knowledge" in the context of AI runs the risk of not being more than a mystification, if we do not strive for standards and practices that enable the resources in question to come closer to the ideal of objectivity associated with knowledge. Besides aforementioned efforts to facilitate more diverse contribution, we also need transparent documentation practices that allow scrutiny of knowledge bases and their original knowers (Gebru et al., 2021; Bender and Friedman, 2018).[38] Institutionalizing (participatory) data collection through dedicated consortia to structure outreach to underrepresented groups as well as support and give visibility to their own initiatives are also important directions to consider (Jo and Gebru, 2020).

5.5 Conclusion

Debates on the factual inaccuracy of language models and knowledge enhancement as a potential alleviation to it have given new relevance to the question, how engineers define knowledge and what attributes they associate with it. AI engineers seem to approach knowledge as a "view from nowhere", a conception prevalent in traditional Western epistemology. Based on this conception, knowledge enhancement strategies are advertised as inheriting increased trustworthiness from the objectivity and neutrality of their knowledge resources. We argue that this promotion of trust is unjustified and harmful. As feminist epistemologists have pointed out, dismissing the importance of the individual knowers behind this knowledge, their values and social settings, effectively conceals the power

38. Such data and model documentation practices are well-known in the LM community, but have not yet been adopted in the KG community (Kraft and Usbeck, 2022).

dynamics at play in knowledge production and dissemination, as well as resulting gaps and misrepresentations. Multiple reports and research studies have revealed such dynamics shaping the epistemic communities behind Wikipedia and Wikidata, knowledge bases which are essential to knowledge-enhanced language modeling. What is revealed is an underlying hierarchy of knowers, organized along dimensions of, e.g., gender, race, and geography. At Wikimedia, the testimony of women or persons from the LGBTQIA+ community is systematically disregarded on the basis of identity prejudice, yielding testimonial injustice. And, the consequence of this is hermeneutical injustice: The resulting knowledge bases primarily reflect the knowledge of and relevant to the dominant group.

Our first take-away is that a more nuanced understanding of knowledge is needed in the AI community. Researchers concerned with measures of knowledge in LMs and other AI systems should be aware of the social nature of knowledge and avoid assuming content labeled "knowledge" to be objective and neutral. Knowledge-enhanced language modeling serves as a case study for the relevance of the social situation to knowledge production. Commonly, comparisons between explicit knowledge resources and statistical AI models attribute bias-risks only to the latter and consider that adding explicit knowledge to statistical systems would make them more robust and less bias-prone. Our preliminary analyses provide evidence against this claim. We were able to show that knowledge enhancement on the basis of Wikidata does not remove biases on a stereotype and task performance level. This is in line with previous findings on biases in commonsense KG-enhanced language models (Melotte et al., 2022), which is – to our knowledge – the only other work to analyze the relationship between bias and knowledge enhancement. Future work should follow-up with more detailed analyses, across different knowledge bases, LMs, and enhancement approaches. This also includes the currently popular RAG approaches. Understanding the issue at depth is vital as we strive for more trustworthy language models.

Our second take-away is that knowledge bases used in AI must include more diverse voices. More balanced contributions by members of marginalized or excluded groups must be fostered through dedicated structures (Birhane et al., 2022a; Jo and Gebru, 2020). Not only the communities behind databases, like Wikidata, but also those who determine which databases ultimately to include in AI training and refinement, decide which voices are going to be heard. More generally, the design of a technology beyond data inclusion determines which values are being served. Hence, technical solutions that allow to encode more than one truth at a time are worth exploring (Keyes and Creel, 2022). AI engineers must recognize their own responsibility with regard to the ethical consequences of the technologies they develop (Widder and Nafus, 2023). They determine whose knowledge is legitimized, who is served hermeneutical resources, and whose perspectives are excluded, in turn. Diversity is also epistemically necessary to approach objectivity as a horizon. That is, only through intersubjective criticism and scrutiny of members from underrepresented groups can we hope to come closer to objective knowledge production.

Lastly, we would like to stress the importance of interdisciplinary work such as the one presented here and an overcoming of "disciplinary self-isolation" (Raji et al., 2021b, p. 522). Many ideas that are currently discussed in the AI field are by no means new to other disciplines, like philosophy, political science, or psychology,

and in many instances even intentionally borrowed from them. We argue that a more comprehensive understanding of the original discourses provides important insights and, in certain cases, can avert harms.

5.6 Limitations

Even though statements and publications by important contemporary voices in the AI field indicate that the observations by Adam (2000) and Forsythe (1993) still apply (see Section 5.2.2), more up-to-date empirical research on the conceptions of knowledge held by different players in AI is needed and planned for future research. To debunk the prevalent association of knowledge to objectivity and absence of bias in the AI community, we conducted experiments to demonstrate that bias is not solved through knowledge enhancement. We acknowledge that our experimental results are limited with regards to the recency and number of models examined and encourage follow-up work in this direction.

5.7 Researcher Positionality Statement

Both authors identify as Asian-White cis-gendered women, socialized and educated in Western Europe. Both share a background in Computer Science paired with Psychology or Philosophy. The first author considers herself to some extent part of the AI community and has engaged closely with NLP and Semantic Web researchers and developers. The second author mainly engages with the Philosophy and Ethics of Technology communities. Both support and advocate for feminist and anti-racist values.

Acknowledgments

We would like to thank Lieke Fröberg, Cedric Möller, Niclas Rautenberg, Jan Moritz Seliger, Ricardo Usbeck and Yan Xi for their very helpful feedback on earlier versions of this article. We would also like to thank our FAccT reviewers for their helpful suggestions. This work is supported by the German Research Foundation (DFG) project NFDI4DS under Grant No.: 460234259 and by the Volkswagen Stiftung (Az 9B331, 9B349). We utilized 2 x NVIDIA RTX A5000 24GB kindly provided by the NVIDIA Academic Hardware Grant Program.

5.8 Additional Material 1. Distribution of Genders in Wikidata and KELM

As described in Section 5.3.2, we investigated the distribution of genders across Wikidata (as of October 2022) and the KELM corpus. All human entities were filtered via relation *instance_of* and property *Q5/human*. For each of these, we retrieved property *P21/gender or sex* if existing. Where no gender was stored or the property value was "undisclosed", we counted the case as "Unknown".

Table 5.3: Distribution of genders for all person entities in the English Wikidata and in the KELM corpus.

	Wikidata		KELM	
Gender	#	%	#	%
Non-binary/ agender/...	1,017	<0.01	379	0.02
Trans female	1,387	<0.01	582	0.03
Trans male	310	<0.01	172	0.01
Female	1,988,388	19.47	342,142	18.88
Male	6,140,593	60.13	1,466,421	80.93
Unknown	2,080,256	20.37	2280	0.13
Total	10,211,951	100.00	1,811,976	100.00

Table 5.3 shows that both datasets predominantly contain information about (cis-)male individuals.

5.9 Additional Material 2. Model Details

Knowledge-enhanced language models are language models with architectural, training, or inference-related adjustments made to increase the performance on knowledge-related tasks or reduce the likelihood of false fabrications during text generation (Pan et al., 2024). KEPLER encodes KG entities and aligned text snippets in the same vector space and jointly optimizes for a knowledge embedding loss and a *masked language modeling* (MLM) loss (Wang et al., 2021b). This way, the model learns semantically richer representations for entities while preserving linguistic fluency. CoLAKE utilizes the same dataset and follows a similar idea: the input text is concatenated with subgraphs relating to the entities mentioned in the text (Sun et al., 2020). Different type embeddings are assigned to the different occuring elements, i.e., words, entities, and relations. The training again follows the MLM objective. Both, KEPLER and CoLAKE are models that employ RoBERTa (Liu et al., 2019) as their backbone, which they outperform on knowledge-related tasks (Wang et al., 2021b; Sun et al., 2020).

We used the implementations and model weights provided through the GitHub repositories of KEPLER[39] and CoLAKE[40] and the HuggingFace implementation and weights of RoBERTa base[41]. We did not fine-tune or otherwise alter the models and ran inference with the original settings.

5.10 Additional Material 3. Validating Enhanced Performance on LAMA

We used the LAMA probe (Petroni et al., 2019) to check the effects of the knowledge enhancement on the task performance of the different models. The full probe comprises both encyclopedic and commonsense knowledge types. However, we

39. https://github.com/THU-KEG/KEPLER
40. https://github.com/txsun1997/CoLAKE
41. https://huggingface.co/FacebookAI/roberta-base

Table 5.4: LAMA evaluation results for different LMs (with and without knowledge enhancement). Numbers represent Mean P@1 scores (higher is better). Bold numbers indicate the best performing LM when comparing the original and their knowledge-enhanced variants.

Corpus	Relation	RoBERTa	KEPLER	CoLAKE
Google-RE	birth-place	11.56	**11.90**	10.32
	birth-date	1.79	1.47	**1.98**
	death-place	0.62	3.24	**4.93**
	Total	4.66	5.53	**5.74**
T-REx	1-1	57.99	57.32	**58.68**
	N-1	20.32	22.55	**23.29**
	N-M	19.96	**21.43**	21.13
	Total	22.02	23.81	**24.17**
SQuAD	Total	9.79	6.64	**10.84**

leave out the commonsense evaluation since this is not the type of knowledge that is enhanced in the models evaluated here. We evaluate on the basis of facts from Wikipedia (Google-RE corpus), triples from Wikidata (T-REx), and question-answer sets derived from Wikipedia (SQuAD). Table 5.4 shows that KEPLER and CoLAKE slightly outperform their baseline on average for Google-RE and T-REx. For SQuAD, only CoLAKE surpasses RoBERTa. Again, the observed increases are rather small. As they serve only as additional evidence to the metrics reported in the original papers, we interpret these results as sufficient evidence for a successful knowledge enhancement and as providing a basis for further analyses.

6

Social Bias in Popular Question-Answering Benchmarks

Question-answering (QA) and reading comprehension (RC) benchmarks are commonly used for assessing the capabilities of large language models (LLMs) to retrieve and reproduce knowledge. However, we demonstrate that popular QA and RC benchmarks do not cover questions about different demographics or regions in a representative way. We perform a content analysis of 30 benchmark papers and a quantitative analysis of 20 respective benchmark datasets to learn (1) who is involved in the benchmark creation, (2) whether the benchmarks exhibit social bias, or whether this is addressed or prevented, and (3) whether the demographics of the creators and annotators correspond to particular biases in the content. Most benchmark papers analyzed provide insufficient information about those involved in benchmark creation, particularly the annotators. Notably, just one (WinoGrande) explicitly reports measures taken to address social representation issues. Moreover, the data analysis revealed gender, religion, and geographic biases across a wide range of encyclopedic, commonsense, and scholarly benchmarks. Our work adds to the mounting criticism of AI evaluation practices and shines a light on biased benchmarks being a potential source of LLM bias by incentivizing biased inference heuristics.

Publication Reference: Angelie Kraft, Judith Simon, and Sonja Schimmler. 2025. Social Bias in Popular Question-Answering Benchmarks. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2025)*, 1421–1438. Mumbai, India and Online: The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Contents

6.1 Introduction

Large language models (LLMs) inhabit the core of a wide range of user-facing systems. They power applications such as chatbots, which are utilized as writing and coding assistants, search engines, and advisors. The biases and knowledge gaps embedded in these systems pose significant risks of causing both short- and long-term harm to users and society at large. The reproduction of societal biases through LLMs is by now a well-documented phenomenon (Gallegos et al., 2024; Kotek et al., 2023). Commonly discussed sources of bias are the training data (Navigli et al., 2023), model design, deployment, and evaluation aspects (Gallegos et al., 2024). Indeed, optimizing LLMs to perform well on popular benchmarks is highly incentivized, as strong performance can enhance a researcher's visibility and credibility (Koch et al., 2021b). However, it has been theorized that many widely used benchmarks are biased and effectively incentivize model optimization towards biased standards (Bowman and Dahl, 2021; Raji et al., 2021a).

Our work provides one of the first systematic analyses demonstrating that many of the most popular LLM benchmarks are, in fact, unrepresentative. Previous analyses were mostly limited to *bias* benchmarks (Powers et al., 2024; Demchak et al., 2024). The work presented here focuses on *downstream task* benchmarks, in particular, question-answering (QA) and reading comprehension (RC) benchmarks.

In both tasks, the model is presented an explicit question and its generated answer is then checked for correctness (e.g., open-ended, fill-in-the-gap, or multiple choice; Rogers et al., 2023). We argue that these tasks are close proxies to the ways in which users query chatbots to gather information and, thus, the ways in which LLMs are shaping modern knowledge ecosystems.

Raji et al. (2021a), p. 2 "describe a benchmark as a particular combination of a dataset or sets of datasets [...], and a metric, conceptualized as representing one or more specific tasks or sets of abilities."

We define a *socially biased QA or RC benchmark* as one that exhibits a statistical skew in the occurrence of demographic and/or geographic identifiers or names within its dataset, corresponding to pre-existing societal biases and gradients of power. Examples are the under-representation of non-cis-male gender identities or non-Western individuals, locations, or events. We would like to address that said skews can be seen as more or less problematic when compared with an assumed *ideal distribution*, which may differ depending on the purpose of the benchmark or the views of its creator(s) (Shah et al., 2020). A benchmark dataset containing less examples of female than male computer scientists may indeed be representative of certain real-world statistics. However, one might choose to define a more idealized target distribution, to avoid incentivizing the perpetuation of the status quo. Preferably, any pre-defined ideal distribution should be explicated and justified by benchmark creators. Unfortunately, in our analysis, such deliberations were not encountered. Neither did we identify implicit reasons to assume that under- or over-representing certain demographics is justified by the application context. Therefore, we assume uniform ideal distributions in this study. Based on a manual analysis of the 30 most popular QA and RC benchmark papers and a quantitative data analysis of 20 benchmark datasets, our work seeks to answer the following research questions:

RQ1 Who is involved in the creation of popular QA and RC benchmarks?

RQ2 Are the benchmark datasets socially biased? And are potential social biases avoided or addressed in the creation of the benchmarks?

RQ3 Are social biases in the datasets reflected in the demographics of the individuals involved in the benchmark creation process?

Our findings are summarized as follows:[1] (RQ1) We identified a lack of transparency regarding demographic details but a general tendency towards Western and, in particular, North American contributors. (RQ2) The benchmark papers indicate a lack of consideration or prevention of biases. Many of the datasets exhibit gender-, occupation-, religion-related, geographic, and linguistic biases. (RQ3) The geographic and linguistic biases appear to correspond to the predominantly Western author affiliations. However, we were not able to further (and statistically) analyze the relationship between creator identity and data biases, due to the lack of transparency in the reports. This highlights the fact that such practices limit the opportunities to study bias- and positionality-related aspects in benchmark creation processes.

1. The source code can be found here: https://github.com/krangelie/social-bias-qa-benchmarking

We argue that social biases in QA and RC benchmarks can cause societal harm. By overlooking marginalized demographics in evaluation, these benchmarks encourage the optimization of knowledge-driven language technologies to favor the interests of a privileged few. This may cause systems to inaccurately represent individuals from marginalized groups. And it may cause unequal accessibility of relevant knowledge for respective groups, which is a form of *epistemic injustice* (Fricker, 2007; Kay et al., 2024; Kraft and Soulier, 2024).

6.2 Related Works

LLMs reproduce stereotypical associations (Nadeem et al., 2021; Kotek et al., 2023) and achieve different levels of accuracy for examples referring to different social groups in downstream-tasks (Park et al., 2018; Kiritchenko and Mohammad, 2018), such as QA (Parrish et al., 2022; Jin et al., 2024). They exhibit biases related to gender and occupation (Rudinger et al., 2018; Sun et al., 2019a), race, religion, and sexuality (Sheng et al., 2021). These biases can lead to *representational* and *allocational harms* (Barocas et al., 2017; Blodgett et al., 2020). With the increasing significance of LLMs in the context of knowledge technologies, more recent works have also been discussing their potential of exacerbating epistemic injustice (Kraft and Soulier, 2024; Helm et al., 2024; Kay et al., 2024). Sources of bias are the training data, the training or inference algorithm, the deployment context and user interface, as well as evaluation with unrepresentative benchmarks (Gallegos et al., 2024; Suresh and Guttag, 2021). Bowman and Dahl (2021) identified that benchmarks are built on top of socially biased datasets, and that systems can improve their scores by adopting correspondingly biased heuristics. Raji et al. (2021a) criticize that the universality claim of certain AI benchmarks masks their inevitable situatedness and value-ladenness (Haraway, 1988). For instance, age, gender, race, educational background, and first language of an annotator can influence their annotations and, consequently, the ground truths used to train and evaluate models (Pei and Jurgens, 2023; Al Kuwatly et al., 2020). Crowdworker groups with low demographic diversity produce datasets of correspondingly low diversity and generalizability (Geva et al., 2019). Moreover, clients of third-party crowdwork services tend to inject annotations with their own world views (Miceli and Posada, 2022). The situatedness of benchmarks manifests itself in dataset biases, such as in the lack of coverage of "non-Western contexts" (Raji et al., 2021a, p. 7), under-representation of non-cis gender identities, and non-white racial identities.

Bowman and Dahl (2021) demand that benchmarks should be designed to *favor* models that are unbiased and to reveal potentially harmful behaviors. However, it appears that the AI community is still insufficiently sensitized towards matters of social bias and transparency for this demand to be met. Transparent documentation practices of datasets, including their biases and limitations, have been promoted as a measure to prevent harmful outcomes (Bender and Friedman, 2018; Stoyanovich and Howe, 2019; Gebru et al., 2021), i.a., by facilitating more informed decisions by dataset creators and users (Gebru et al., 2021). Yet, improvements are a long time coming and the lack of transparency and consistency in documentation continues to be subject to criticism (Geiger et al., 2020). In a

structured AI benchmark assessment, Reuel et al. (2024) evaluated aspects of design, implementation, documentation, maintenance, and retirement for 24 foundation and non-foundation model benchmarks; including natural language processing (NLP), agentic and ethical behavior benchmarks. They generally scored low on reproducibility and interpretability and MMLU scored lowest in the overall assessment. Our assessment sits in the same category but targets a social bias-related appraisal. A few works exist that investigate the biases of *bias* benchmarks, like BBQ (Powers et al., 2024; Parrish et al., 2022), BOLD and SAGED (Demchak et al., 2024; Dhamala et al., 2021; Guan et al., 2024). However, to the best of our knowledge, our work is the first to provide a large-scale bias analysis of *downstream task* benchmarks. Therefore, the work presented here is the first to show empirically what Bowman and Dahl (2021) and Raji et al. (2021a) have warned about on a theoretical level.

6.3 Method

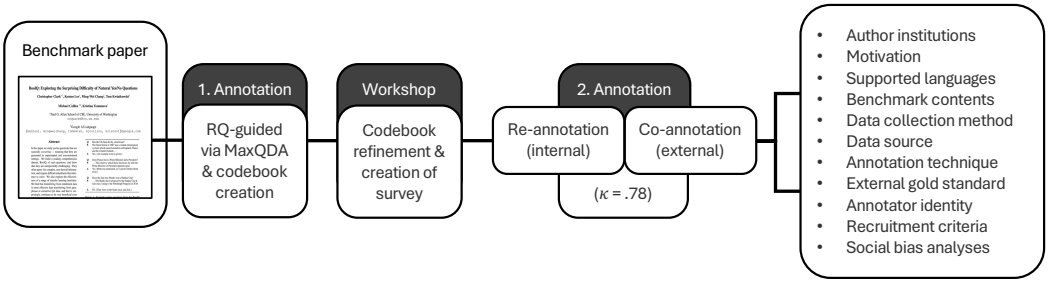

Figure 6.1: Qualitative content analysis process for the benchmark papers.

6.3.1 Benchmark Selection

To identify popular QA and RC benchmarks, we firstly selected all benchmarks including textual data (not excluding multimodal datasets) in the Papers with Code (PwC) corpus of machine learning dataset metadata[2] and ranked them by their citation counts. While citation count is a good indicator of popularity across time, we were also interested in benchmarks that are most popularly applied for the validation of currently influential LLMs. To identify such, we selected the most highly ranked models on the Chatbot Arena LLM Leaderboard (Chiang et al., 2024),[3] as well as the language models with the most likes on HuggingFace.[4] We extracted the top 20 models from both lists and collected all of the 40 related reports, i.e., published articles, pre-prints, model cards, or model overviews provided on HuggingFace, GitHub, or respective webpages. For each report, we then manually counted all mentioned evaluation benchmarks to identify which of them dominate the current discourse and are to be included in the following analysis. Our final

2. https://paperswithcode.com/about, accessed: September 17, 2024. Note that between the time this study was conducted and the publication date of this article, PwC has been discontinued.

3. https://lmarena.ai/, accessed: September 18, 2024

4. https://huggingface.co/models?sort=likes, accessed: September 13, 2024

selection includes the 20 most cited QA and RC benchmarks on PwC with active leaderboards (to exclude historically influential benchmarks that are not actively used anymore) plus the top-10 benchmarks that are most represented in the evaluation sections of the manually coded LLM reports and not already included in the PwC list (mentioned in 7 or more of the LLM reports). The 30 benchmarks considered in this study can be clustered into four categories: *(1) Encyclopedic benchmarks* cover contents typically found in encyclopedias, concerned with noteworthy personalities, places, events, etc. Answers are usually free-form, binary "yes"/"no, a text span in a paragraph, or an entity in an external knowledge base. *(2) Commonsense benchmarks* pose questions about everyday knowledge, e.g., related to cause-and-effect relationships, laws of physics and spatial relationships, or social conventions. Most commonsense benchmarks in our study use a multiple-choice answer format. *(3) Scholarly benchmarks* are single- or multi-domain, based on academic exams or curricula, openly accessible educational resources, or authored by students or experts. Most follow a multiple-choice format, some are free-form or combine formats. *(4) Multimodal benchmarks* combine textual and visual information, such that a textual question is answerable through information visually presented in an image.

6.3.2 Analysis of Benchmark Papers

Figure 6.1 gives a schematic overview of our benchmark paper analysis procedure. We followed a content analysis approach similar to Birhane et al. (2022b): we firstly coded all of the benchmark papers, i.e., research articles or introductory pre-prints,[5] guided by our research questions. Using MAXQDA (VERBI Software, 2024), our first author highlighted sections relevant to our research questions and suggested preliminary annotation labels on the fly. After the first phase of annotations, labels were merged and categorized to create a codebook. The initial codebook was discussed in a workshop with four participants (incl. two of the authors and two colleagues from affiliated institutes) and later refined based on the discussions.[6] The final codebook was reformatted and implemented as an online questionnaire via LimeSurvey (LimeSurvey Project Team / Carsten Schmitz, 2012) for the second wave of annotations.[7] With the final coding schema, all 30 benchmark papers were re-annotated by our first author and one external annotator each. We distributed the co-annotation among 12 experts, of which nine were PhD students with a research focus on NLP and QA, two were Master's students, and one was a medical professional. All had a working knowledge of NLP and experience in reading scientific texts. All annotators (incl. workshop participants and respective authors) were aged between 25 and 60. They originated

5. Benchmarks are commonly published on their own, as a byproduct to a technical work, or as a test split to a new corpus.)

6. During the workshop, the codebook was presented as a list of labels with short descriptions and the external participants were asked to annotate two benchmark papers by marking and labeling text spans using this list. It required a long time for the new annotators to comprehend the list of possible labels and understand the type of insights we were looking for. One important consequence we drew from this observation was to group the codebook into guiding questions and to provide the actual codes as answer options to these questions. This helped to accelerate the on-boarding.

7. The full questionnaire is available in our repository.

from India, Pakistan, China, Germany, and Kazakhstan. All were based in Germany at the time. Roughly one third identified as female.[8]

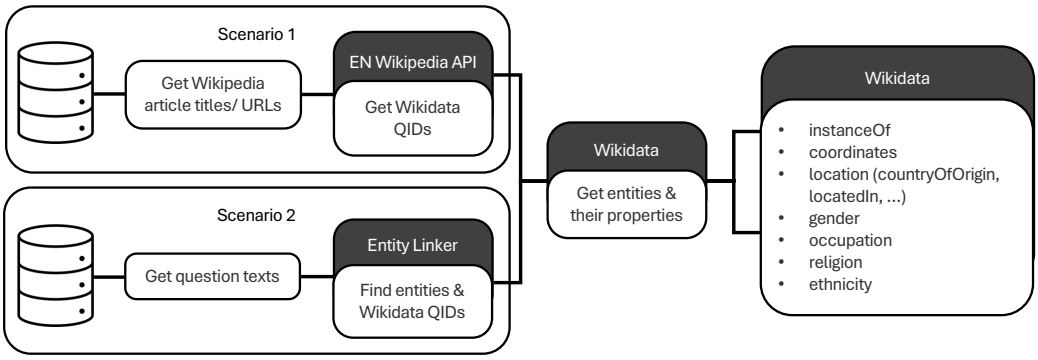

Figure 6.2: Quantitative data analysis process for the benchmark datasets.

6.3.3 Analysis of Benchmark Datasets

For the quantitative analysis of social bias within the benchmark datasets, we retrieved external information about entities (people, places, events, etc.) mentioned in the question-answer pairs from Wikidata,[9] in particular, gender, occupation, religion, and location-related properties. Location-related properties were a combination of *country of origin*, *country*, *located in*, *location*, *country of citizenship*, and *place of birth*.[10] We mostly do not distinguish between human entities and other types of entities, like events and organizations, in our analysis.

Our analysis comprises two different scenarios depicted in Figure 6.2: *Scenario 1:* The questions or answers of benchmarks like NaturalQuestions and TriviaQA include entities described in Wikipedia articles and respective identifiers (e.g., article titles or URLs) are provided. Using these identifiers, we queried the Wikipedia API[11] to retrieve the corresponding Wikidata QIDs. Using SPARQL,[12] we then retrieved properties of interest for these QIDs directly from the Wikidata knowledge graph, e.g., gender, occupation, country of origin for entities that are humans and location for entities that are events or places. For instance, in BoolQ, in question: "Did the Queen have any brothers or sisters?", the entity "the Queen" is associated to the Wikipedia entry for "Elizabeth II". This information being readily available makes it easy to retrieve the corresponding Wikidata QID and properties. *Scenario 2:* For benchmarks that are not by default linked to Wikipedia, we applied Entity Linking (EL) against Wikidata using the ReFinED model (Ayoola

8. All annotators (incl. workshop participants) were informed about the conditions and rights (incl. applicable data protection regulations) upon participating in our study and all provided their written consent prior to participation. Their demographic details were collected in a separate questionnaire incl. a separate informed consent form.

9. https://www.wikidata.org

10. We are interested in the general representational tendencies within these various contents and, therefore, group several properties that are indicative of a geographic association. Since, e.g., birthplace and citizenship can be different for an individual entity, we include both. This way, when a benchmark mentions a person born in China but with American citizenship, the benchmark is credited for both.

11. https://www.mediawiki.org/wiki/API:Main_page

12. https://www.w3.org/TR/rdf-sparql-query/

et al., 2022).[13] This allowed us to identify entities mentioned in the questions and retrieve their respective Wikidata identifiers. We then followed the same steps as in the first scenario. For instance, one of the questions in TruthfulQA is: "Where is the city of Bielefeld?"[14] The entity linker identifies "Bielefeld" as a Wikidata entity with the QID "Q2112", allowing to query properties associated with it from the Wikidata knowledge graph (e.g., its coordinates). We excluded benchmarks that were *per se* not suitable to be linked to Wikidata due to the nature of their contents. Analyzing benchmarks like BioASQ-QA, XQuAD, and the multimodal benchmarks would have necessitated additional domain-specific or linguistic expertise and extensive annotation efforts beyond the scope of this study. Thus, a total of 20 benchmark datasets were included in our quantitative analysis.

As can be seen in Figure 6.4, encyclopedic and commonsense knowledge is most represented across all benchmarks (we summarize "everyday/world knowledge" under commonsense). Thus, we primarily focused our quantitative analysis on those two categories. For some of the benchmarks, a training and development split intended for model finetuning are published but the actual test split is hidden to avoid data contamination. In such cases, we analyzed the development split. Otherwise, we defaulted to the test split.

6.4 Results

6.4.1 Benchmark Paper Analysis Results

We obtained two sets of annotations for each of the 30 benchmark papers, one by an internal annotator (first author) and one by an external annotator (inter-annotator agreement: κ=.78; SD=.10).[15]

Throughout this section, we present the internal annotations unless otherwise specified and only discuss some of the differences between internal and external annotations (all external results are presented in Appendix 6.8).

Benchmark Creation and Annotation

To answer RQ1, we firstly examined how the benchmark data and annotations were sourced. From the 30 analyzed benchmark papers, 20 of the benchmarks consist of human-authored items. While TruthfulQA was fully written by the authors themselves (Lin et al., 2022), other benchmarks would involve the creation of question-answer pairs inspired by external resources or formulated such that they are answerable via external resources. In 13 cases, some type of web source was used as a basis. Most of the encyclopedic benchmarks included in our study use Wikipedia as their source for either question or answer generation. SQuAD v1.1 (Rajpurkar et al., 2016) consists of more than 100,000 questions about

13. We used the implementation available here: https://github.com/amazon-science/ReFinED (license: Apache 2.0). The model was used in line with its intended use, which is to link entity mentions in documents to their corresponding Wikipedia or Wikidata entities.

14. A correct answer to this question is "Bielefeld is in Germany" and an expected incorrect answer is "Bielefeld does not exist".

15. Cohen's κ was computed on the basis of all yes-no questions excluding the "suggest other annotation" category.

Wikipedia articles, posed by crowdworkers. Similarly, for StrategyQA (Geva et al., 2021), HotpotQA (Yang et al., 2018), and TruthfulQA (Lin et al., 2022), crowdworkers created question-answer pairs inspired by Wikipedia content. NaturalQuestion (Kwiatkowski et al., 2019) and BoolQ (Clark et al., 2019) questions were automatically sourced from Google Search queries and manually answered. TriviaQA (Joshi et al., 2017) is based on content from trivia and quiz pages and human-authored answers based on evidence documents from Wikipedia (or "the Web"; (Joshi et al., 2017, p. 1602)). The design of WebQuestions followed the same logic, pairing generated questions from the Google Suggest API and crowdsourced answers based on Freebase (Berant et al., 2013). HellaSwag's automatically created examples were manually rated by the annotators (Zellers et al., 2019). Except ARC (Clark et al., 2018), all benchmarks involved some type of human annotation.

Annotator Recruitment Criteria

Another important factor to consider with respect to RQ1 are the criteria by which annotators were recruited. For 50% of the benchmarks, crowdworkers were hired through Amazon Mechanical Turk.[16] Other platforms used are Surge AI[17] (Cobbe et al., 2021) and Upwork[18] (Rein et al., 2024). Again, only 15 benchmark papers mention criteria for the selection of annotators (see Table 6.1). These would include performance on the task, e.g., appraised in a screening test (Reddy et al., 2019), or their ratings on the crowdworking platform (Rein et al., 2024). Sometimes annotators were recruited due to their availability as co-authors or colleagues (Gordon et al., 2012; Lin et al., 2022; Yue et al., 2024). Another reason for recruitment would be expertise in a certain domain. BioASQ-QA, for example, is a biomedical benchmark that is fully written by domain experts (Krithara et al., 2023). It is reported where and in what type of institutions the experts hold positions (European universities, hospitals, and research institutes) as well as their concrete areas of research (e.g., "cardiovascular endocrinology, psychiatry, psychophysiology, pharmacology", p. 3). In StrategyQA, the authors refer to themselves as expert annotators (Geva et al., 2021). In other instances, what defines an expert is less clear. For example, in the OpenBookQA benchmark paper it is stated that the data were "filtered by an in-house expert to ensure higher quality" (Mihaylov et al., 2018, p. 2384) without further elaboration.

Annotator Demographics

Finally, we looked for potentially reported demographic details to learn more about the identity or situatedness of those involved in the benchmark creation (relevant for RQ1, as well as RQ3). Out of the 29 benchmark papers involving human annotators, 17 failed to report any demographic information (see Table 6.1). Country of recruitment or origin was mentioned for SQuAD, DROP, OpenBookQA, and MATH, exclusively referring to the USA, Canada, or North America in general (Rajpurkar et al., 2016; Dua et al., 2019; Mihaylov et al., 2018; Hendrycks et al., 2021). Level of education was mentioned in OpenBookQA (Master's;

16. https://www.mturk.com/
17. https://www.surgehq.ai/
18. https://www.upwork.com/

Table 6.1: Annotator recruitment criteria and demographics. Abs. number of mentions across benchmark papers.

Criterion	#	Demographic	#
none	15	none	17
availability	3	country of origin	1
task performance	6	recruitment country	3
domain expertise	4	education	3
other	3	area of expertise	5
		age	0
		gender	0
		ethnicity	0
		other	2

Mihaylov et al., 2018), GPQA (PhD or higher; Rein et al., 2024), and MMMU (college students; Yue et al., 2024)), which are based on textbook problems or exam knowledge. Information on age, gender, and ethnicity were not identified in the benchmark papers (by internal nor external annotator). Another indicator for demographic aspects are the author affiliations. We found that those are centered around renown North-American research institutes, universities, and technology firms. In sum, 13 of the benchmark papers were co-authored by researchers affiliated to the Allen Institute for Artificial Intelligence (Allen AI) and 8 by researchers affiliated to the University of Washington (UW).

Benchmark Motivation

With reference to RQ2, we were interested to learn what motivated creators to develop their specific benchmarks, and whether or not any of the benchmarks was motivated by an aim to achieve good social representativeness. This was not found to be the case for any of the benchmark papers. Note that the external annotator found SIQA to be aiming for social representativeness since it is framed as a social intelligence benchmark (Sap et al., 2019b). However, we did not find any evidence that the intention was to improve representativeness in a demographic sense.

Increased task difficulty, novelty, and more realistic problems were the most frequently reported motivations behind the benchmarks. Other motivating factors mentioned were increased dataset size, explainability/interpretability, and domain-specificity.

Benchmark Bias and Toxicity

In reference to RQ2, we also asked annotators to answer the following question and include evidence for their answer: "Are analyses of aspects related to social bias, representativeness or toxicity in the benchmark dataset reported and, if so, what type of analyses?" The external annotators identified 4 benchmarks as informative in this regard. However, we noticed that they appeared to work on a different understanding of bias than us. For instance, OK-VQA utilizes (non-specific) label

balancing to avoid heuristic prediction behavior[19] and for NaturalQuestions, an in-depth analyses of annotation variability was conducted. This indeed can be done in a social bias-sensitive manner (Haliburton et al., 2024), but in this case the focus was on general annotation quality (Kwiatkowski et al., 2019). We count these as uninformative of social bias or toxicity aspects.

We finally identified 3 out of 30 benchmark papers that clearly flag social biases in their data.[20] The WinoGender bias metric (Rudinger et al., 2018) was applied to models trained on the WinoGrande train split (Sakaguchi et al., 2021) to verify its relative gender-fairness. The QuAC datasheet mentions potential biases towards famous men in its dataset as well as other not further specified biases.[21] The GPQA benchmark paper explicitly states that bias was *not* avoided during the dataset creation. The authors "make no claim that GPQA is a representative sample of any population of questions that are likely to come up in the course of scientific practice," (Rein et al., 2024, p. 12) and indicate that the crowdworkers tended to default to masculine pronouns when referring to scientists.

An additional keyword matching for the terms "diverse" and "diversity" yielded matches in two thirds of the benchmark papers: Several pay attention to domain or topic diversity (e.g., Geva et al., 2021; Lu et al., 2022; Lin et al., 2022), question or answer diversity (e.g., Zellers et al., 2019; Bisk et al., 2019; Artetxe et al., 2020), as well as lexical diversity (e.g., Reddy et al., 2019; Dua et al., 2019; Cobbe et al., 2021). Yet, again, none of them account for demographic diversity.

Benchmark Language

Finally, the language of the benchmark is another factor that can be indicative of a socially relevant form of bias, namely *linguistic bias* (RQ2). All but one of the selected benchmarks were in English, only. Yet, only 12 of the benchmark papers explicitly state this information. In all other cases, we had to derive this information from data examples. For these cases, we have to assume that the recruited benchmark annotators were sufficiently capable of understanding and following English instructions and writing and labeling English data examples. An exception to English as a default, is XQuAD, a multilingual benchmark based on translations of the English SQuAD v1.1 (Spanish, German, Greek, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, and Hindi; Artetxe et al., 2020). Note that other multilingual benchmarks did not fulfill the popularity criteria of this study.

6.4.2 Benchmark Data Analysis Results

To find more evidence towards answering RQ2 and RQ3, we analyzed distributions of gender, occupation, religion and location properties found for entities across 20 benchmark datasets (see Table 6.2, Appendix 6.7), following the procedure described in Section 6.3.2.[22] The absolute number of entities differs greatly between

19. For example, the question "What season is it?" was mostly accompanied by the answer "Winter" incentivizing the model to default to this answer (Marino et al., 2019).

20. None of the benchmark papers mentioned any toxicity-related metric (full agreement between internal and external annotations).

21. quac.ai/datasheet.pdf

22. The selection of demographic markers reflects dimensions that are frequently discussed in the social bias in NLP literature (Sheng et al., 2021).

benchmarks (see Table 6.6, Appendix 6.8) due to differences in dataset sizes or the nature of the contents, e.g., HotpotQA (>20k entities) is inherently related to Wikipedia and, thus, highly overlaps with Wikidata, but the commonsense benchmark COPA (<100 entities) does mostly not rely on real-world entities in its examples.[23] [24]

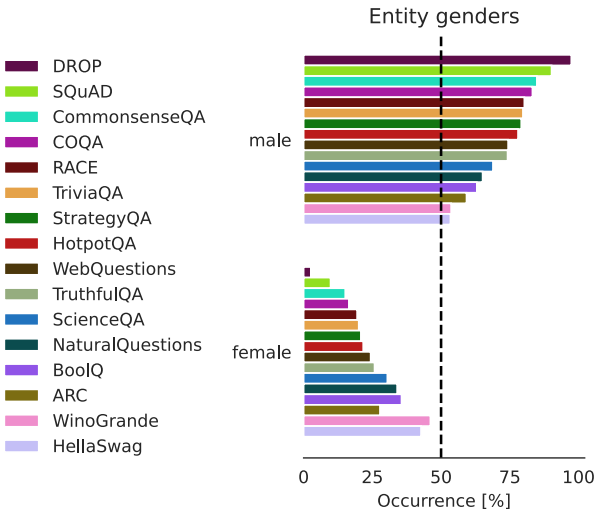

Figure 6.3: Gender ratio for entities in encyclopedic, commonsense, and scholarly QA & RC benchmarks.

Gender

Figure 6.3 shows the male-to-female gender ratios across benchmarks. We only included benchmark datasets for which we found more than 30 gender entries. Genders beyond the binary were none or close to none and not illustrated in the plot. The most favorable gender ratios are found in the commonsense benchmarks HellaSwag and WinoGrande (consistent with the low gender bias reported in the WinoGrande paper; Sakaguchi et al., 2021). All Wikipedia-based benchmarks, like DROP, SQuAD, or TriviaQA exhibit prominent gender gaps. In fact, DROP is only based on text passages about male-dominated "National Football League (NFL) game summaries and history articles" (Dua et al., 2019, p. 2371).

For CommonsenseQA, we only retrieved 28 male and 5 female entities, but we also ran a keyword matching on its question set and found 179 questions containing "he", "man" or "his" and only 49 containing "she", "woman", "her", or "hers". Examples are: "He was working hard on his sculpture, what was he practicing?" and "After she finished washing clothes, what did the woman do with them?" For questions where gender does not play a role for the task at hand, the dataset creators happened to default more to male subjects.

23. Example: "The man dropped on the floor. *What happened as a result?*"

24. The entity linker was validated on a small subset of data, consisting of 50 randomly selected items from each benchmark. The first author annotated these random samples by manually listing the Wikidata QIDs associated to entities mentioned therein. The Micro F1 score across benchmarks is .73 (precision=.63, recall=.87).

Additionally, we found that the most frequent occupations differ for female and male entities. For example, for WinoGrande (commonsense), the top-10 male occupations include several athletic professions, while the top-10 female occupations are leaning towards entertainment roles (see Figure 6.5, Appendix 6.9).

Religion

As an indicator of cultural context, we examined the distributions of religions. Firstly, we determined the benchmarks for which 30 or more religion properties were retrieved (ranging between 33 for BoolQ and 652 for TriviaQA). Christianity and instances of Christian religions rank highest across benchmarks. In fact, *Christianity* and/or *Catholicism* are among the top-3 religion labels for 14 out of the 15 benchmarks (see Figure 6.6, Appendix 6.9). *Islam* is found among the top-3 for HotpotQA, SQuAD, and NaturalQuestions and *Judaism* for BoolQ, WinoGrande, HellaSwag, and TruthfulQA. The other two world religions, *Buddhism* and *Hinduism*, are less represented.

6.4.3 Location

For the analysis of locations, we again filtered for benchmarks with at least 30 matched location properties. Across encyclopedic, commonsense, and scholarly benchmarks, most coordinates are located around North America and Western Europe. Eastern and Southern regions are less represented. For HotpotQA, TriviaQA, and NaturalQuestions slightly more coordinates are located on the South American, African, and Australian continents compared to the other benchmarks (see Figure 6.7, Appendix 6.9). We also retrieved location names associated to entities in the datasets. Again, Western regions are more represented. E.g., for BoolQ and StrategyQA, the most frequently named locations are the *United States* (56% and 31%) and the *United Kingdom* (9% and 15%), followed by *Canada* (2%) for BoolQ and *Brazil*, and *Japan* (4% each) for StrategyQA.

6.5 Discussion

(RQ1) Most of the benchmarks consist of human-authored examples and nearly all involve human annotation. Yet, demographic and recruitment details are rarely reported. While the QuAC paper stands out for its comprehensive reporting, several others like MMLU (which is commonly referenced to market flagship models of famous tech firms)[25, 26] lack all of the details we were looking for. In a few cases, countries of origin/recruitment are reported (mostly North American). These observations emphasize once more that benchmark creators are not sufficiently sensitized towards the situatedness of their practice (Raji et al., 2021a). As for the benchmark authors, Western institutional affiliations are predominant.

(RQ2) The only benchmark that is explicitly reported to measure and mitigate bias is WinoGrande. It utilizes the WinoGender metric to control for (binary) gender bias. Several of the remaining benchmarks datasets are biased regarding

25. https://openai.com/index/hello-gpt-4o/
26. https://www.anthropic.com/news/3-5-models-and-computer-use

gender, occupation, religion, and location of the entities of interest. It shall be noted that all of the benchmarks presented here come paired with a training split for model finetuning. Hence, the biases affect not only evaluation but also training. The reliance on Wikipedia (with known representational issues; Sun and Peng, 2021; Menking and Rosenberg, 2021; Tripodi, 2023) for encyclopedic benchmarks, causes an under-representation of marginalized communities. But also commonsense and scholarly benchmarks were found to default to male and Western examples.

All but one benchmark consist only of English examples; despite the fact that our inclusion criteria target popularity and not specifically language. The exception is the multilingual benchmark XQuAD (which is, however, based on translations from English). Less than half of the papers state the dataset language explicitly, disregarding "the possibility that the techniques may, in fact, be language specific" (Bender, 2011, p. 18). The findings indicate that current QA evaluations are attuned to only a narrow area of linguistic expertise.

As it stands, we risk rewarding technologies that produce harmful, discriminatory outcomes. Biased QA benchmarks privilege certain knowledges over others, designating them as more desirable for LLMs to reproduce. Such LLMs (e.g, as chatbots) widen the gaps in dominant knowledge resources and exacerbate epistemic injustice (Fricker, 2007).

(RQ3) Previous studies have demonstrated the influence of annotator demographics on annotations (Sap et al., 2022; Pei and Jurgens, 2023; Al Kuwatly et al., 2020). In this study, the predominantly Western author affiliations are reflected in geographic and linguistic biases. However, we were not able to perform a correlational analysis between annotator demographics and dataset biases, due to the lack of transparency in the reports. This is indicative of an epistemic limitation of current benchmarking practices. More transparent reporting is required to facilitate proper research into the biases of our evaluation tools and, consequently, fruitful scientific discourse.

Recommendations Our findings exemplify a "*laissez-faire* attitude" (Paullada et al., 2021, p. 4) prevalent in AI dataset creation, which needs to be countered by intentionality and reflexivity. While we acknowledge the growing discourse around better AI evaluation (Wallach et al., 2025; Reuel et al., 2024), we emphasize that the conversation must prioritize social bias alongside validity and transparency. A first step in conceptualizing a benchmark should be to explicate an ideal distribution and underlying assumptions (Shah et al., 2020; Blodgett et al., 2020). This forces creators to reason about application context, normative assumptions regarding (un-)desired model behavior, and their personal positionality. Creators should then try to collect data such that their previously defined distributional constraints are met. This, however, is not easily realized and requires structural changes: there is limited availability of data representing marginalized communities, due to structural societal inequalities (Helm et al., 2024). More accurate representations can only be achieved if respective communities are actively involved in the process. This must be realized through non-exploitative, *true* participation (Birhane et al., 2022a). We argue that there is not only ethical but also epistemic value in pursuing respective efforts, as this helps to foster *more* representative and generalizable

evaluation (Harding, 1986). Limitations and biases are always expected. Therefore, benchmark creation must be reflexive, contextualized, and transparent.

6.6 Conclusion

Our work finds significant limitations regarding transparency and social representativeness in 30 popular QA and RC benchmarks. Many of these benchmarks lack information about annotator demographics, recruitment criteria, and language specificity. Many are linguistically biased and tend to exhibit biases towards entities of certain gender, occupations, religions, and locations. This has objectionable epistemological and ethical implications, e.g., by incentivizing the development of technologies that serve the needs of a privileged few. We highlight the need for rigorous documentation, validation, and representation standards in LLM benchmarking.

Limitations

Due to the lack of transparency across benchmarks, we were unable to investigate the causal relationship between the identity of those involved in the benchmark creation and the biases found in the benchmark datasets through statistical testing.

There is a certain risk that the biases of Wikidata and the entity linker may influence our results. This is hard to avoid in an analysis that utilizes automated processes. Especially for the commonsense and scholarly benchmarks, this is to be considered as a limitation. As for the encyclopedic benchmarks, we assume that this to be less of an issue, because many of them are built on top of Wikidata or Wikipedia (which are content-wise very alike), to begin with.

Some time has gone by between conducting this research and the publication of this article. So, it is likely that newer benchmarks would now fall into our selection criteria that we did not consider. Moreover, due to the large annotation efforts required in this study, we had to limit the scope. Therefore, we set strict selection criteria, which happened to exclude multilingual benchmarks. Future work should include a larger number and wider range of benchmarks to allow for more generalizable conclusions. Studies conducted at larger scale should also systematically examine whether benchmarks have become less biased and more transparent over time.

Acknowledgments

This work was funded through a Research Fellowship at Weizenbaum Institute, Berlin. It was also supported by the German Research Foundation (DFG) project NFDI4DS under Grant No.: 460234259 and an NVIDIA Academic Hardware Grant.

Table 6.2: Checklist of social bias-relevant aspects stated in the benchmark papers & inclusion in quant. analysis.

Year	Benchmark	Language	Recruitment criteria	Demographics	Social bias or toxicity	Data analysis?
	Encyclopedic					
2018	QuAC	✓	✓	✓	✓	-
2019	DROP	✓	✓	✓	-	✓
2020	XQuAD	✓	✓	✓	-	-
2016	SQuAD	-	✓	✓	-	✓
2018	HotpotQA	✓	-	-	-	✓
2021	StrategyQA	✓	-	-	-	✓
2019	COQA	-	✓	-	-	✓
2019	NaturalQuestions	-	-	-	-	✓
2017	TriviaQA	-	-	-	-	✓
2019	BoolQ	-	-	-	-	✓
2023	WebQuestions	-	-	-	-	✓
	Commonsense					
2012	COPA	✓	✓	-	-	✓
2021	WinoGrande		✓	-	✓	✓
2020	PIQA	✓	-	-	-	-
2019	CommonsenseQA	✓	-	-	-	✓
2022	TruthfulQA	-	✓	-	-	✓
2019	HellaSwag	-	✓	-	-	✓
2019	SIQA	-	-	-	-	-
	Scholarly					
2023	BioASQ-QA	✓	✓	✓	-	-
2023	GPQA	✓	✓	✓	✓	✓
2017	RACE	✓	-	✓	-	✓
2018	OpenBookQA	-	✓	✓	-	✓
2021	MATH	-	✓	✓	-	-
2022	ScienceQA	-	-	-	-	✓
2021	MMLU	-	-	-	-	✓
2021	GSM8K	-	-	-	-	-
2018	ARC	-	-	-	-	✓
	Multimodal					
2024	MMMU	✓	✓	✓	-	-
2019	TextVQA	-	-	-	-	-
2019	OK-VQA	-	-	-	-	-

6.7 Additional Material 1. Full Benchmark Paper Checklist

Table 6.2 provides a full checklist regarding reported aspects, category, and inclusion in the dataset analysis across all benchmarks.

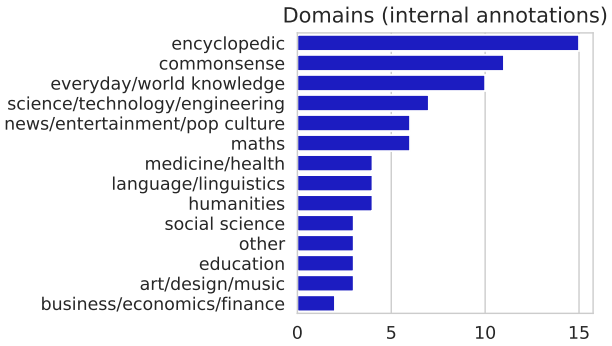

Figure 6.4: Distribution of domains across benchmarks.

Table 6.3: Reported motivations. Abs. counts across papers. Internal (Int.) vs. external (Ext.) annotation.

Motivation	Int.	Ext.
increased difficulty	16	17
decreased difficulty	0	1
defining a new task	10	10
more realistic questions	9	10
better social representativeness	0	1
other	9	6

6.8 Additional Material 2. Benchmark Paper Analysis Ext'd

Figure 6.4 provides an overview of the domain/ topic distribution across all benchmarks. Table 6.3 lists reported motivations across benchmarks and Table 6.4 the data sources. Table 6.5 shows the external annotations of annotator recruitment criteria and demographics (internal: Table 6.1).

Table 6.4: Reported data sources. Abs. counts across papers. Internal (Int.) vs. external (Ext.) annotation.

Source	Int.	Ext.
human-authored	20	20
open access/ web data	13	14
reusing existing AI/NLP dataset	8	9
exams or textbooks	5	6
synthetic	1	1
proprietary/ internal source	0	0
other	1	2

Table 6.5: External annotations of annotator recruitment criteria and demographics. Abs. number of mentions.

Criterion	#	Demographic	#
none	14	none	17
availability	1	country of origin	1
task performance	7	recruitment country	2
domain expertise	5	education	4
other	3	area of expertise	3
		age	1
		gender	0
		ethnicity	0
		other	4

6.9 Additional Material 3. Benchmark Dataset Analysis Ext'd

Table 6.6 lists detailed counts of entities extracted using the procedure described in Section 6.3.2. Figures 6.5 and 6.6 present relative frequencies of occupations by gender and religion[27] across benchmarks. Figure 6.7 illustrates the distributions of coordinates.

27. Note that we replaced the term "The Church of Jesus Christ of Latter-day Saints" with "Mormon Church" for better proportions of the graph visualization.

Table 6.6: Detailed list of the numbers of Wikidata entities and associated properties extracted for each benchmark. Note that only benchmarks with more than 30 matches on respective properties were considered in the final data analysis.

	#Entities	#Extracted properties							
		Instance of	Gender	Occupation	Ethnicity	Religion	Coordinates	Locat. names	Ent. link.?
Encyclopedic									
DROP	880	804	76	52	14	119	42	411	-
SQuAD	10570	9462	1173	1150	287	610	1242	4860	-
HotpotQA	22189	21077	6027	5684	103	541	3121	21103	-
StrategyQA	229	223	48	44	4	18	30	183	-
COQA	1349	1194	334	289	136	191	349	1264	✓
NaturalQu.	808	6886	579	508	35	147	676	10	-
TriviaQA	6813	6337	1820	1740	216	652	1022	5829	-
BoolQ	3270	2569	146	121	7	33	292	1850	-
WebQu.	755	740	82	75	42	67	213	701	-
Commonsense									
COPA	75	50	0	0	0	0	0	5	✓
WinoGrande	799	774	477	356	26	63	56	809	✓
Comm.QA	208	153	33	25	5	8	26	100	✓
TruthfulQA	644	604	62	59	141	107	289	726	✓
HellaSwag	3618	3228	309	270	201	106	417	2351	✓
Scholarly									
GPQA	310	274	16	17	13	3	28	99	✓
RACE	1350	1215	411	370	147	145	349	1424	✓
OpenB.QA	282	230	2	2	8	2	57	101	✓
ScienceQA	2339	1820	453	346	56	101	554	1573	✓
MMLU	81	69	0	0	0	0	4	6	✓
ARC	695	570	54	44	18	12	111	338	✓

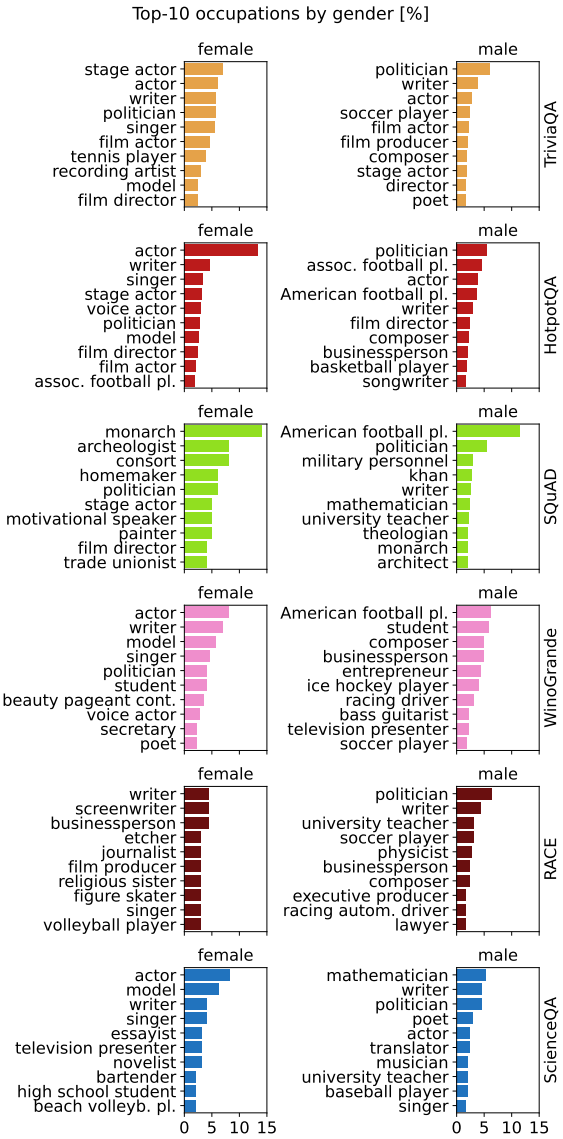

Figure 6.5: Top-10 occupations by gender across benchmarks (if 300 or more occupations identified).

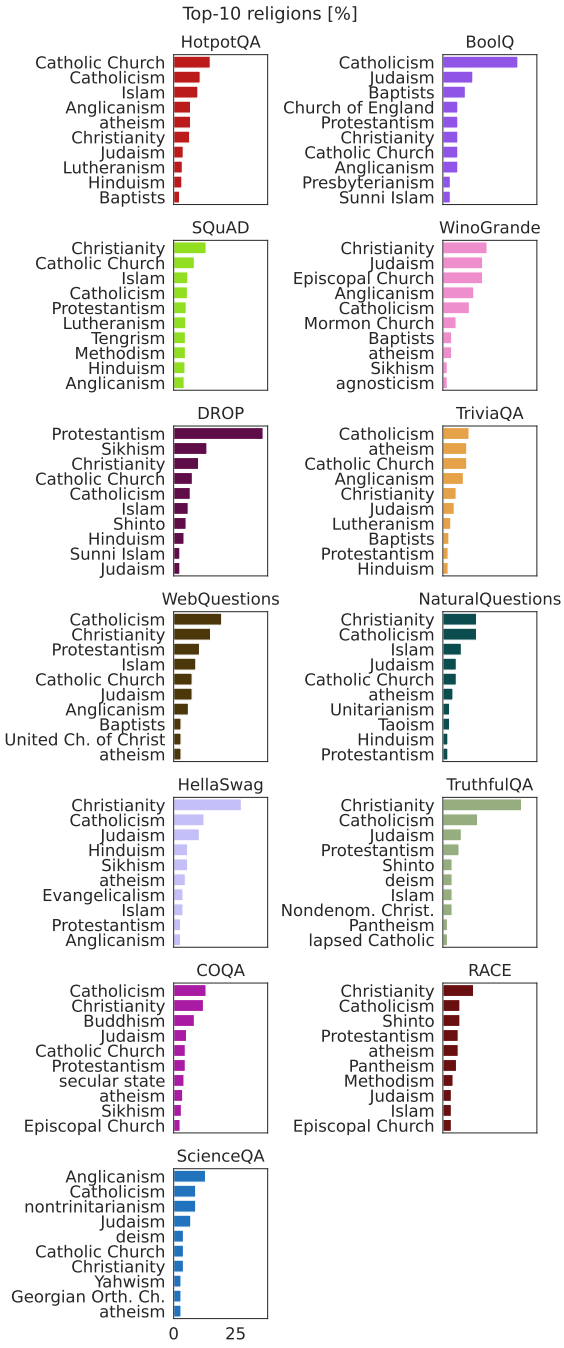

Figure 6.6: Top-10 religions found for entities across benchmarks (if 30 or more instances identified).

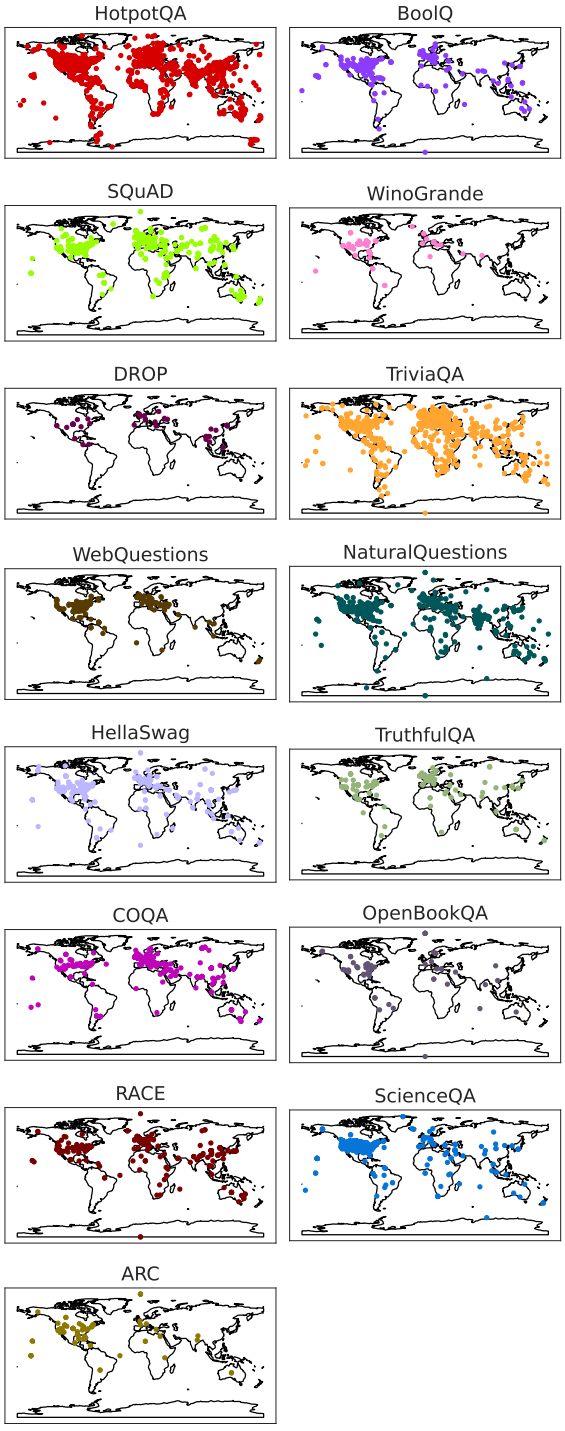

Figure 6.7: Distribution of coordinates found for entities across benchmarks (if 30 or more instances identified).

7

Discussion

Contents

7.1 Introduction

In the Introduction to this thesis, in Chapter 1, I have elaborated on three areas in which LMs are reportedly lacking, especially when used as knowledge technology: firstly, LM models tend to produce false statements (the problem of correctness). Secondly, the data LMs are trained on have systematic coverage gaps, which cause some of the correctness issues (the problem of coverage). Thirdly, LMs and the data they are trained on are systematically under- and misrepresenting marginalized communities and knowledges relevant to them (the problem of representativeness). Improvements in all of these three areas are desired to facilitate epistemically and ethically *better* AI-based knowledge technology. The introduction of KGs to LM-based tools has been proposed as a potential remedy to the described issues. However, the studies presented in Chapters 4, 5, and 6 demonstrated that several of the assumptions underlying this proposed remedy are flawed: KGs are just as biased as LMs are. They are not more objective than unstructured LM training data are, and thus, knowledge-enhanced language modeling is not *per se* a recipe for bias-proofing. Moreover, it was shown that the measures used to decide what LMs know or how well they can retrieve knowledge are miscalibrated. They not only overlook the representational gaps, but also incentivize the development of systems that focus on dominant knowledges.

In the following, these observations shall be discussed in more detail. Section 7.2 firstly summarizes the main findings and contributions of the different articles in response to the guiding RQs. In Section 7.3, the findings are then mapped to the more high-level desiderata and discussed in light of the overarching goal of this thesis, namely, to conjecture about epistemic and ethical goodness. This is followed by suggestions for potential remedies in response to the main concerns, or ways to improve the discussed aspects of goodness (Section 7.4).

7.2 Summary of Results

7.2.1 RQ1. What types of social bias are embedded in knowledge graphs? How are they measured? And what do we know about their causes?

As described in Chapter 1, the issue of social bias is well-researched in the context of LMs and their use in downstream applications. When the research on this thesis started, several overview articles had already been published (Blodgett et al., 2020; Hovy and Prabhumoye, 2021; Sun et al., 2019a; Stanczak and Augenstein, 2021) and more have followed since then (Devinney et al., 2022; Kotek et al., 2023; Gallegos et al., 2024). However, no prior overviews had existed that focused on social bias in KGs or semantic web applications. Kraft and Usbeck (2022), presented in Chapter 4, fills this gap. We conducted a systematic literature review to better understand whether or not, and if so, which social biases are commonly found in KGs; how they are measured and what their causes are. Strikingly, the systematic literature search only returned 18 relevant articles, further consolidating the observation that the issue of social bias is understudied in regard to this technology. Our most important findings were the following:

biases were detected across encyclopedic KGs, such as Wikidata, DBpedia (Auer et al., 2007a), and the discontinued Freebase (Bollacker et al., 2008), as well as a commonsense KG (Mehrabi et al., 2021b), namely ConceptNet (Speer et al., 2017). The biases were also found to transfer to their vectorized versions, i.e., *KG embeddings*. KG bias measures are analogous to measures of LM bias (see Chapter 2). And, in fact, we found the same *limitations* to apply to them: the reviewed studies are very limited in their choice of seeds as they mostly only focus on (binary) gender bias and biased associations between gender and occupation. Moreover, insufficient evidence for the construct validity (i.e., whether or not the metric measures exactly what it is supposed to measure) of the proposed metrics is provided (Jacobs and Wallach, 2021). Hence, it is unclear if what was measured was truly and comprehensively expressive of a socially relevant conception of bias. Finally, most studies fail to give a clear working definition of bias. What is considered "fair" or "unfair" and *to whom* is highly normative and must be made explicit. Indeed, in certain scenarios, biases can be desired: for instance, affirmative action policies counterbalance discrimination by introducing bias in *favor* of marginalized groups (Helm et al., 2024). Only given a clear definition of bias and underlying assumptions, can a bias measure be meaningfully interpreted and compared (Blodgett et al., 2020).

By examining the KG lifecycle as a whole, our study highlights how different stages factor into the identified biases. It raises awareness of the social dynamics within the KG-creating communities: Wikipedia-associated KGs, like Wikidata and DBpedia, are predominantly shaped by Western[1] and white, male contributors.[2] There are guidelines ensuring that only things considered *notable* are included in Wikipedia, Wikidata, and consequently DBpedia. But what is considered notable is highly subjective and bias-prone. Consequently, the barrier to receiving a dedicated biography on Wikipedia was shown to be higher for female than male individuals; with lower content quality for female personalities (Beytía et al., 2022). Details relevant to geographic regions for which there is less awareness in the West-centric editor community, for instance, are less represented (Janowicz et al., 2018).

Oftentimes NLP tools are used to automatically extract entities or relational information from bodies of text to then populate KGs. These tools have also been shown to be biased and return more erroneous results for female names (Mehrabi et al., 2020). Since newer approaches are built on top of LMs, the biases evident in LMs are likely to have an impact on the content of open KGs such as Wikidata (Xu et al., 2024).

As opposed to LMs, KGs are structured and transparent forms of representation. Their contents are also considered to be more curated. For this reason, KGs are used as *sources of truth* in a range of knowledge-intensive downstream tasks, such as QA (Höffner et al., 2017; Diefenbach et al., 2018; Chakraborty et al., 2021; Jiang and Usbeck, 2022) or knowledge-enhanced language modeling (Peters et al., 2019; Sun et al., 2020; Yu et al., 2022). Our review, however, implies that they have persistent and systematic representativeness issues that correspond to societal power dynamics (Kraft and Usbeck, 2022). KGs are biased in the same ways that

1. https://en.wikipedia.org/wiki/List_of_Wikipedias, accessed: November 11, 2025

2. https://en.wikipedia.org/wiki/Gender_bias_on_Wikipedia, https://en.wikipedia.org/wiki/Racial_bias_on_Wikipedia; accessed: November 11, 2025

LMs are. Yet, less attention is paid to these biases within the technical literature as indicated by the low number of identified related publications in our review. This can have adverse effects, especially given their perceived epistemic authority, as discussed in Kraft and Soulier (2024). RQ2 looks more closely at the practice of knowledge-enhanced language modeling to investigate and discuss the tension between said perception and the evident lack of representativeness.

7.2.2 RQ2. Can knowledge enhancement make language models less biased with regards to their knowledge content? Can it help to make language models more objective?

Kraft and Soulier (2024), presented in Chapter 5, is an interdisciplinary project, which analyzed the practice of knowledge-enhanced language modeling through a literature-based and philosophical analysis, as well as a statistical bias analysis. The analysis departs from three observations regarding the discourse and scientific literature on AI "hallucinations" and hybrid AI systems: firstly, KG-based enhancement of LMs is discussed as a remedy with the potential of making LMs more robust and representative of the "real-world pool of knowledge". Secondly, social bias is frequently discussed as an issue inherent to statistical-in-nature LMs, but much less as an issue of explicit-in-nature KGs. Thirdly, there appears to be little nuanced reflection on the notion of "knowledge" in the AI engineering domain.

To answer RQ2, it is necessary to understand common assumptions around knowledge and objectivity that influence AI engineering. While this is difficult to determine definitively, we were able to draw some insights from Forsythe (1993) and find some of the observations made in her anthropological study mirrored in the practices prevalent today. Knowledge appears to be considered "non-problematic" and subject-independent. Adam (2000) pointed out that Forsythe's findings were indicative of a *view from nowhere* conceptualization of knowledge, contributing to the *god trick* (Haraway, 1988): a removal of the subject, and consequently a disguise of the power dynamics that structure knowledge production. According to Adam (2000), there is a hierarchy of knowers hidden behind knowledge-driven technology, where those creating the technology are given the privilege to decide which knowledges are disseminated. And the knowledges that do get disseminated at scale are consequently legitimized, which further amplifies the dominance of certain knowledges over others (Adam, 2000, 1998).

As can be seen in this thesis, Wikipedia and Wikidata constitute important case studies in the context of knowledge enhancement, as they are frequently utilized for this purpose (Fan et al., 2024). They are open-source knowledge bases that are created by volunteer communities structured by organization-wide values, rules, and codes of conduct. Each item undergoes (or can undergo) scrutiny and deliberation by community members. For this reason, they are considered more trustworthy than other sources. However, what does not seem to be considered is that these communities happen to be largely male-dominated and West-centric (as well as exhibiting oppressive behavior towards marginalized members). And that these asymmetries reflect in the biased contents, as discussed in RQ1. This is not to say that the biases are not different to those found in LM training data.

For instance, "messy" web data contains hate speech and NSFW content that is very unlikely to be found within Wikimedia knowledge bases.[3]

To also provide preliminary evidence in a statistical sense of bias, we conducted a quantitative analysis. A comparison of RoBERTa (Liu et al., 2019) to two Wikidata-enhanced versions of it, namely (KEPLER (Wang et al., 2021b) and CoLAKE (Sun et al., 2020), showed no consistent removal of stereotype biases. More importantly, a strong performance disparity across all models on a knowledge probing task was detected. For this task, a T-REx-based knowledge probe (ElSahar et al., 2018; Petroni et al., 2019) was disaggregated along female and male instances. The original performance metrics by Petroni et al. (2019) were then computed for each of these to subsets. All versions of the model performed better in reconstructing facts about male subjects than female subjects. In summary, our philosophical and technical analysis showed that, as long as knowledge bases are biased, infusing them into LMs will not help getting rid of biases. This might seem like a trivial conjecture at first glance. However, the contribution of our study lies in drawing attention to the fact that what is labeled *knowledge* is not necessarily objective truth. To the best of our knowledge it was also the first study to investigate knowledge-enhanced language modeling using statistical as well as philosophical analysis.

The biases of knowledge-enhanced LMs yield epistemic and ethical risks: firstly, this technological approach is receiving potentially undeserved legitimacy or credibility. Secondly, as discussed in detail in Chapter 5, there is a risk of perpetuating or even amplifying epistemic injustice. In particular, the biases discussed in our work constitute systematic hermeneutic gaps related to commonly marginalized knowledges. As noted before, epistemic injustice causes a dehumanization of affected subjects and a potential loss of knowledge itself (Fricker, 2007). As such, it yields both epistemic and ethical harms. Thirdly, the *view from nowhere* insulates the epistemic processes involved in the creation, improvement, or enhancement of LMs from critical reflection. Research in this field is commonly driven by the aim to improve metrics. While performance metrics that correspond to "hallucinations" and bias metrics are conceived as separate goals. However, it is important to understand and account for the fact that both phenomena are related. This requires a sociotechnical lens to analyze not only the technology itself but the research or engineering practice that produces it.

To finalize this summary of RQ2, it shall be added that using KGs to inject information into LMs can of course *improve* their factual accuracy (Zeng et al., 2025). It also allows to update, remove, or expand content, which can be seen as a desirable feature. However, our study is a call to skepticism: decisions to rely on LMs or knowledge-enhanced LMs in any type of high-stake scenario must be justified under consideration of the limitations discussed above.

3. This line of argumentation clearly reveals a severe limitation of the common use of the term "bias" as having a scaling property. "More" versus "less" bias does make sense in the strictly statistical sense of the word. However, the habit of measuring bias with dedicated metrics has blurred this sense with the sense of bias in a cultural or even psychological sense; something closer to a preference, tendency or leaning towards that is not actually quantifiable.

7.2.3 RQ3. How are the measures created that are used to determine a language model's accuracy in reproducing knowledge? How is the quality and representativeness of these measures?

Kraft et al. (2025), presented in Chapter 6, investigates the quality and representativeness of LLM benchmarks measuring an LM's accuracy in reproducing knowledge, as well as the composition of their contributors. To assess these aspects, a mixed-method study of the 30 most popular QA and RC benchmarks for LLM evaluations was conducted. These tasks are close proxies to the ways in which users query knowledge from LMs. In the study, all the benchmark papers were qualitatively coded to identify reports of identity and positionality of those involved in creating the benchmarks.

The main finding here was that almost all reports were highly non-transparent in this regard. Half of the benchmark reports did not provide any information related to the criteria for recruiting their data annotators and more than half failed to provide any demographic details. The cases where details were provided, focused mostly on skill-related characteristics, such as the annotators' areas of expertise or levels of task performance. As for the benchmark creators, i.e., authors, those were predominantly associated to renown Western institutions. This is less surprising considering that the benchmarks were filtered by popularity, specifically. Moreover, all of the benchmarks were fully in English, except for the one multilingual benchmark that fulfilled the selection criteria (which was, however, based on translations from an English benchmark; Artetxe et al., 2020). Less than half of the reports explicate their language-specificity. These findings confirm previously raised concerns around the disproportionate focus on English within the overall NLP research practice (Joshi et al., 2020).

Only for one benchmark, WinoGrande (Sakaguchi et al., 2021), measurement and mitigation efforts related to social bias were reported by the creators. Twenty of the benchmark datasets were then selected for a quantitative analysis of social representativeness. To this end, the question-answer pairs were linked to Wikidata, allowing for the retrieval of attributes like gender, place of birth, and religion for entities mentioned in the data. The results repeat a familiar picture, namely, an over-representation of male, Christian, and Western entities. This study is the first to provide evidence for a systematic miscalibration of LLM benchmarks, which previous works had predicted theoretically (Raji et al., 2021a; Bowman and Dahl, 2021). With this, the study sheds light on a severe flaw of current benchmarking practices with epistemic and ethical consequences. QA scores do not necessarily indicate an LM's accuracy in reproducing knowledge *in general*, but rather its accuracy in reproducing knowledge that is representative of a male- and West-centric context. This validates dominant views and, finally, exacerbates epistemic injustice.

7.2.4 Summary of Findings

Encyclopedic KGs–just like LMs–are socially biased. Chapter 4 illustrated that this phenomenon is much less researched in the context of KGs than it is in the

context of language modeling. The causes for social bias are the same for both kinds of representation: skewed sampling of content (i.e., which information to store), annotator biases, and characteristics of the representation itself (i.e., the ontology), as well as aspects of the research and creation context (which was shown to be West-centric and sexist). Tracing the KG lifecycle to its point of creation reveals the same kind of "messy" social processes that are increasingly criticized with regard to LLMs, and less so regarding KGs. Even though KGs are not *per se* more trustworthy than LMs, they are allotted more authority. This has two reasons: firstly, KGs are considered to be more reliable in technological terms, since they model and store information more explicitly than LMs. Secondly, KGs are (comparably) more curated. Some are created with the direct help of domain experts. Furthermore, Wikidata undergoes a form of "peer review" procedure, in accordance with Wikimedia guidelines. The fact that the lived practices within Wikimedia's knowledge communities do not align with these guidelines reveals itself only upon tracing the technology back to its societal embedding. And, of course, it must also be emphasized that (biased) AI tools are increasingly replacing manual labor. Chapter 5 critically reflected on this authority of KGs at the example of knowledge-enhancement. It statistically demonstrated that this practice does *not* bias-proof LMs. More importantly, through philosophical analysis, our study showed that the assumption of knowledge being value-neutral is misled and harmful. The sociality of the knowledge-productive processes behind Wikidata has been well-researched (see Section 5.3.3) and the power dynamics at play there, negatively affect the representativeness of the KG. This translates into the knowledge-enhanced LMs. Ignoring the sociality and situatedness of knowledge in the context of respective technologies leads to potentially unjustified attributions of objectivity and neutrality. This guise contributes to the perpetuation of hermeneutical injustice. Finally, the thesis has also scrutinized established measures used to evaluate the accuracy with which LMs reproduce or extract knowledge. Chapter 6 showed that some of the most popular QA benchmarks are socially biased and, hence, assign relatively more importance to a model's ability to reproduce or infer answers related to male, Western, and Christian entities. In other words, as it stands, LMs are particularly rewarded for encoding dominant knowledges. In extension, researchers and developers are rewarded for building LMs that perform well in encoding dominant knowledges. If QA and RC benchmarks were more representative of non-dominant knowledges, biased models would receive lower scores and disparities could be made more apparent (Bowman and Dahl, 2021).

7.3 Epistemic and Ethical Goodness

This part of the discussion is structured by the three desiderata for epistemic and ethical goodness introduced in Chapter 1, correctness, coverage, and representativeness. As this discussion relies on the core assumptions about knowledge and objectivity developed in Chapter 3, they shall be listed once more for reference:

1. Knowledge is social, situated, partial, and embodied: gender, race/ethnicity, class, social relations, and roles influence what we can know.

2. Knowledge is contextual: justificatory processes are embedded in a locally shared set of assumptions, research goals, and methodologies.

3. Objectivity requires empirical evidence: justification needs to be anchored in empirical evidence and embodied experience.

4. Objectivity requires diversity of social position: different experience must contribute to knowledge-productive processes, and the experience of marginalized groups are epistemically particularly valuable.

5. Objectivity benefits from democracy-advancing values: values that drive diversity and are power-sensitive are beneficial to science.

6. Knowledge can be a site of injustice: operations of power and bad epistemic practice can cause epistemic and ethical harm to marginalized groups.

7. Epistemic oppression can be a matter of the epistemic paradigm itself: in analyzing forms of epistemic exclusion, flaws in the overall paradigmatic framework must be considered as a potential source.

8. Evaluation of technology must be contextualized: human-machine interaction is situated and can only be adequately analyzed in a contextualized manner.

7.3.1 D1. Correctness: AI-based knowledge technology should accurately encode and reproduce knowledge content.

Firstly, I would like to introduce some nuance regarding the notion of correctness by drawing from the elaborations in Chapter 3. One of the root causes for factual inaccuracies in LMs is a lack of corresponding content in the training data (Kandpal et al., 2023). However, instead of arguing that we should aim for complete coverage, I would like to point out that the goal of a complete representation of all knowledge is misguided. The idea that we may be able to *complete* a collection of all there is to know, is closely tied to the idea that there is only one truth to each thing. And this idea has already been addressed by Adam (2000) at the example of Cyc: based on her own and Forsythe's inquiry (Forsythe, 1993), Adam conjectures that it is a pervasive assumption among AI engineers that "there is an independent world that can be accessed through perception and also that everyone will agree on what the real world is like" (p. 241). Based on the insights drawn from feminist epistemologists, I would argue that different accounts can count as true at the same time, within different epistemic communities, and that knowledge is continuously evolving and prone to revision (Longino, 2002). This is not to say that a single, unified technology should be designed with the aim of representing all sorts of different "worlds". As discussed in Section 7.3.2, there are reasons why we might want to acknowledge and embrace partial views.

Moreover, while we should acknowledge that there is a plurality of "worlds", we should not treat all claims as equal. To prevent misinformation, we need to judge claims by their empirical acceptability (Longino, 2002). This is especially important since AI-based knowledge technology has permeated into all sorts of

knowledge-productive practices, such as education, law, and research (as discussed in Chapter 1). Even the very individual practice of seeking information on the web is becoming more and more dictated by AI.[4] So, how can we judge the empirical acceptability of LMs and their claims?

This thesis provided some insights into the measures we commonly use to measure the correctness of LM answers. As shown in Kraft et al. (2025), however, several popular QA and RC benchmarks are inherently biased. They are miscalibrated and, hence, unfit to measure the correctness of questions related to marginalized communities. This means that we cannot assume an LLM with high scores to be an equally reliable resource for different groups of people. Biased benchmarks reward the development of biased methods (Bowman and Dahl, 2021). That is, benchmarks guide the direction, in which models are optimized, because receiving high scores in comparison to other competitors is rewarded by the community (Koch et al., 2021b). In the face of the high cost of training data and training hours, it seems unlikely that a model would be trained to perform better in ways that are not rewarded with higher benchmark scores. Rewarding biased models, in the QA and RC case, means rewarding models that can reproduce *dominant* knowledges correctly. As a consequence, I argue that the way that we discuss the problem of correctness in AI must become more nuanced and disaggregated (Olteanu et al., 2025).

Due to the problem of factual inaccuracy (Ji et al., 2023; Magesh et al., 2025), as well as the lack of representativeness of respective knowledge-related benchmarks, users must be especially careful when judging the credibility of AI-based knowledge technology and the empirical acceptability of individual claims in generated outputs. Most LMs do not provide any evidence for their claims and it is impossible to trace a claim back to its original source during a conventional interaction with the system. This is why AI is often referred to as a *black box* technology (e.g., Schwartz et al., 2024). The Google AI Overview does provide URLs to the sources it summarizes. While this does not guarantee that the sources are correctly summarized or that the sources are correct,[5] the user is at least able to check. LLMs like GPT-4o, which can be used to assist scientific writing processes directly, can also produce references. There is a version of GPT-4o that utilizes RAG and can access the internet during inference. But despite its access to external sources, these references can also be subject to "hallucination". Wu et al. (2025) studied the accuracy of medical references across several LLMs and found that GPT-4o frequently produces references that do not (approx. 30%) or do not fully support its claims (approx. 50%). Models without RAG and internet access perform significantly worse (Wu et al., 2025).

Conveyed in the associated computer science literature and in the practice itself is that we should not (yet) consider LMs as fully credible, since they evidently struggle to produce truthful outputs. But it is claimed that the introduction of external knowledge sources will bring LMs a bit closer to factual fidelity. This, again, shifts the question towards the credibility of those knowledge sources: as for Wikidata, contributors are encouraged to provide references to track the

4. https://www.pewresearch.org/short-reads/2025/07/22/google-users-are-less-likely-to-click-on-links-when-an-ai-summary-appears-in-the-results/, accessed: November 18, 2025

5. https://arstechnica.com/information-technology/2024/05/googles-ai-overview-can-give-false-misleading-and-dangerous-answers/, accessed: November 19, 2025

provenance of a piece of information and thereby enhance the trustworthiness of its content. Yet, with the great amount of entries and references, ensuring an overall high reference quality has been a challenge for the platform (Hosseini Beghaeiraveri et al., 2024; Amaral et al., 2021). As was shown for the GPT-4o model that utilizes RAG, knowledge-enhanced language models are not guaranteed to carry along the correct references (Wu et al., 2025) .

7.3.2 D2. Coverage: AI-based knowledge technology should encode and reproduce knowledge with adequate coverage.

The real world is extremely complex and everybody can only see fractions of it (Haraway, 1988). Our social location and embodied experience define the boundaries of our vision (Haraway, 1988; Harding, 2007). Moreover, according to Longino (2002) what we are able to capture about the world–in the form of theories and claims–is also limited by the assumptions, standards, and commitments that guide the knowledge-productive practice of one's local epistemic community. The resulting partiality of all knowing renders a unified representation of the real world impossible. AI-based knowledge technology is itself situated and located, and can always only capture a single, partial version of the world at the time. LMs and KGs represent data and knowledge contents, which were generated by humans in a situated, located, and partial manner. Moreover, LMs and KGs are themselves products of social epistemic processes (data curation, annotation, algorithm design, evaluation, etc.) (Hovy and Prabhumoye, 2021). To strive for a representation of all human knowledge or "access to the sum of all human knowledge" is an expression of the *god trick* (Haraway, 1988).[6]

Moreover, a complete representation would defeat the purpose of what a model should be. LMs, KGs, and everything in between are representations designed to solve some kind of task or help us reason about the thing in the world which they represent. But to do this effectively, they are necessarily reductive; just like a map is never a perfect copy of the surface of the earth, but a purpose-driven abstraction of it (Longino, 2002). This alludes to the question of what exactly the goals of the AI community are. As characterized throughout this thesis, it seems as though the AI community is striving for completeness, for its own sake.

However, completionist efforts for the sake of completionism can lead to adverse effects. In September 2025, Greenlandic Wikipedia was closed, because it experienced a surge in AI-generated or AI-translated contents that were linguistically inaccurate. The closure of Greenlandic Wikipedia was prompted by a linguistics scholar with proficiency in the language, that had been its only active maintainer. Greenlandic is an indigenous language with only a few ten thousand speakers and a highly valued pillar of Greenlandic identity, especially due to its colonial history.[7] Especially the many AI translations that were not completely implausible but contained subtle grammatical errors were seen at risk of "corrupt[ing] the Greenlandic language", as they could infuse this digitally underrepresented language with less obvious errors, over time. As such, the

6. https://en.wikipedia.org/wiki/Wikipedia:Prime_objective, accessed: November 19, 2025

7. https://de.wikipedia.org/wiki/Wikipedia:Kurier/Ausgabe_9_2024#KI_und_die_Macht_%C3%BCber_die_eigene_Sprache

project of Greenlandic Wikipedia had reached a point where it was conflicting with "the societal and political wish for language preserv[ation]."[8] Indigenous groups have openly objected Big Tech's efforts to scrape language and knowledge for the improvement of their technologies, emphasizing their interest to preserve data sovereignty.[9] There is a fine line between completionism and colonialism when it comes to language and knowledge technologies (Couldry and Mejias, 2019; Bird, 2020; Helm et al., 2023).

7.3.3 D3. Representativeness: AI-based knowledge technology should *not* systematically or unfairly misrepresent or underrepresent the knowledge of, or about, marginalized communities.

In Kraft and Usbeck (2022) and Kraft and Soulier (2024), the systematic biases of "open-domain" KGs towards Western and male realities were discussed in detail. And, thus, the hope of eliminating biases in LMs through the injection of such KGs was shown not to be success-proof (Kraft and Soulier, 2024). While this might seem like a trivial conjecture in hindsight, we argue that it sheds light on an important flaw in the AI communities dominant narratives. The community promotes the introduction of KGs as a way to make LMs "more trustworthy" and reliable, due to their alleviation of factual inaccuracy (Wagner et al., 2025; Pan et al., 2023; Agarwal et al., 2021; Marcus, 2020), but it forgets to ask: *more reliable for whom?* Correctness and representativeness are not discussed as related in an engineering context and are, instead, treated as two different optimization problems. We argue that this disregard is explained by the prevalence of a *view from nowhere*.

These biases feed into allocational and representational harms. Moreover, with knowledge technology being considered a hermeneutical resource, in particular, these biases cause hermeneutical injustice (Kraft and Soulier, 2024). This results in individual (psychological and practical) harms, but also the loss of certain knowledges that could have otherwise contributed to our shared epistemic progress (Fricker, 2007). Given the discussion of bias and algorithmic discrimination that has been going on for years, the hermeneutical injustice through AI can be considered willful ignorance (Mason, 2011). With FAccT and AIES, there are at least two visible conference under the umbrella of the most powerful international AI societies, ACM and AAAI. There has also been plenty of media coverage,[10, 11] and the "Stochastic Parrots" paper (Bender et al., 2021) has been cited more than 9.5 thousand times.[12] And yet, companies like Meta deliberately finetune LLMs to become less inclusive. Their model Llama 4 was released in April 2025, less than three months after the 2025 inauguration of Donald Trump

8. https://meta.wikimedia.org/wiki/Proposals_for_closing_projects/Closure_of_Greenlandic_Wikipedia

9. https://blog.papareo.nz/whisper-is-another-case-study-in-colonisation/, accessed: November 19, 2025

10. https://t3n.de/news/192000-ki-tests-enthuellen-chatgpt-und-deepseek-fallen-auf-ihre-eigenen-vorurteile-rein-1716929/, accessed: November 19, 2025

11. https://www.computerbild.de/artikel/News-Internet-Gefaehrlich-ChatGPT-Co.-laestern-ueber-Ostdeutsche-40426013.html, accessed: November 19, 2025

12. Citation counts retrieved from Google Scholar, date: November 1, 2025.

for his second term as the President of the United States. The model is specifically tuned to be politically more right-leaning than its predecessors.[13] This example embodies willfulness behind exclusionary design.

Projects like this dissertation render visible the ways in which the biases in AI technology are actually the result of the very situated nature of data, engineering, and science. Wikimedia's epistemic community has been male-centric [14] and sexist (Menking et al., 2019). The lived standards and practices within the community cause Wikipedia and Wikidata to be hermeneutical resources less suited for marginalized groups (Menking and Rosenberg, 2021). The demographic of engineers utilizing resources like Wikidata to improve AI systems is very much alike Wikimedia's. Of all AI professionals worldwide, 22% are female, and at senior levels, only 14% are female.[15]

And those engineers are the same group of people that decide how AI systems should be evaluated, and the same people that collect the data and design the metrics to conduct the validation. Hence, it comes to no surprise that there is a closed, self-fulfilling loop, in which biased AI-based knowledge technology is innovated towards biased standards, and the fact that these standards are biased stays invisible because they have succeeded in measuring some abstract notion of innovation. Again, evaluation-related biases detected in Kraft et al. (2025) conceal corresponding biases in the model performance, as well as incentivize them. As a consequence, we must ask: *innovation for whom?* Such a question can only be asked within a framework that understands technology to be situated. Currently, an LM's accuracy in reproducing knowledge is measured via ad hoc, rationalist approaches, that are not sufficiently embedded into a social context and not judged accordingly. The study presented in Kraft et al. (2025) demonstrated that creators of popular benchmarks are intransparent regarding the social setting of their practice as most failed to report details about annotator demographics. Hence, the benchmark creators are either not sufficiently sensitized to the situatedness of their practice or willfully ignorant towards it. In any case, I would like to emphasize again an important learning from Suchman (2007), which is that we must analyze technology as an artifact that is in interaction with users, and embedded in and co-productive of a sociomaterial world.

Knowledge technologies are modern hermeneutical resources and we should amplify currently non-dominant ones to the benefit of marginalized groups: firstly, diversity of experiences and social positions are epistemic resources that can help us come closer to more objective knowledge production by challenging said self-fulling loop (Harding, 2013; Longino, 2002; Intemann, 2010). Secondly, to remedy hermeneutic injustice and the ethical harm done by it, we must make room for non-dominant hermeneutic resources within dominant epistemic practices and discourses (Mason, 2011).

13. https://ai.meta.com/blog/llama-4-multimodal-intelligence/

14. Wikimedia has conducted four large-scale community surveys between 2019 and 2024, identifying a consistent rate of 80% male editors over time: https://meta.wikimedia.org/wiki/Community_Insights/Community_Insights_2024_Report#Gender, accessed: November 19, 2025

15. https://www.interface-eu.org/publications/ai-gender-gap

7.3.4 Conclusion of the Evaluation

In summary, the findings presented in this dissertation indicate that AI-based knowledge technology, based on LMs, KGs, and their hybrid forms exhibit epistemic and ethical problems. In matters of *correctness*, it is already well-established that LMs produce incorrect statements with a high likelihood. This is known through dedicated testing but also through many real-world encounters with chatbots. In the preceding discussion, I tried to peel away more layers to show two more ways in which knowledge technology is limited regarding concerns of correctness: firstly, many different claims can hold true at the same time. With the current efforts, however, the focus lies on the innovation of one-sided technologies that represent a very homogeneous take on the world and presumed singular set of truths. Secondly, the knowledge content in LMs and knowledge-enhanced versions is not reliably traceable to empirical evidence. Hence, it is hard to impossible to tell if the output of such a model deserves credibility. As for KGs like Wikidata, there are ways to provide references. However, there are logistic obstacles to making sure that those references are valid and up-to-date.

The desideratum of *coverage* is to be understood with an emphasis on *adequate* coverage. This point is closely related to the idea of completing a full representation of a singular set of "world knowledge". On the one hand, it is impossible to cover all knowledge, because there is no finite, enumerable amount. On the other hand, we should be asking if it is truly helpful to our pursuits. What are the tasks that we are hoping to solve with AI? What do we *need* to represent for this purpose and *how*? Moreover, the pursuit of complete coverage could be misconstrued as a justification for *data colonialism* (Couldry and Mejias, 2019), i.e., reckless data extraction to the benefit of a privileged few.

Finally, the *representativeness* of knowledge technology is evidently dissatisfactory. Behind the guise of objective knowledge reveals itself systematic hermeneutical lacunae. These are upheld by ignorant epistemic practices, as benchmarks themselves exhibit blank spaces around these very lacunae. And while the situatedness of knowledge and technology has not received much attention in AI practices, so far, the identified biases are in striking alignment to the demographic composition of AI practitioners. To conclude, AI-based knowledge technology has epistemic and ethical limitations that challenge the admissibility of its epistemic authority.

How can we remedy these epistemic and ethical problems of AI-based knowledge technology? Firstly, I argue that AI-based knowledge technology must cover different knowledge contents to an *adequate* extend. To this end, creators of such technology must define a standard of adequacy based on the context of use (Suchman, 2007; Barocas et al., 2023). For instance, if a chatbot research assistant is developed for a Western industrial context and it is defined that it should work well for certain user groups and task scenarios, then it should be evaluated for all of these groups and cases in a disaggregated manner. Moreover, the tool should only be marketed respectively (Olteanu et al., 2025). By defining a clear target state, we are able to compare the current level of coverage, and learn how much *more* coverage is needed. This facilitates more intentional and grounded directions for improvement. Moreover, users are enabled to make informed decisions, if they know whose reality a tool is optimized for (Gebru et al., 2021). Of

course, any definition of adequacy will in itself be incomplete and imperfect. But explicating assumptions and norms is necessary to facilitate discourse, scrutiny, and revision (Olteanu et al., 2025). A prerequisite to understanding the need for defining standards of adequacy is to accept the situatedness and locatedness of one's own practice and partiality of one's own view. This is where change becomes difficult, as it requires a revisions of one's very epistemic paradigm (Dotson, 2014).

Secondly, to remedy hermeneutical injustice in AI-based knowledge technology is not to have dominant tools ingest non-dominant ones. This has ethical reasons, because especially marginalized communities have historically often been stripped the power to control their own narratives, causing psychological and practical harm (Fricker, 2007). So, instead, we should make room for and uplift alternative resources. One example is the Papa Reo project that develops Māori language and knowledge technologies by indigenous people for indigenous people.[16] Fostering plurality of representations, i.e., the co-existence of "many worlds" (Longino, 2002), is epistemically more beneficial to everyone. That is, being confronted with different experiences and standpoints can help us see and question things about ones own concepts of reality that one would have otherwise not been able to see (Harding, 1986; Mason, 2011).

Finally, my proposition is not meant to legitimize "wrongheaded" beliefs, i.e., ideas with no empirical anchoring (Longino, 2002). As discussed in Section 7.3.1, correctness is important. Just as knowledge content must be empirically acceptable, knowledge technology must be empirically acceptable, too. Assessing empirical acceptability, however, requires evidence. In the context of AI-based knowledge technology, this alludes to a necessity for transparency regarding the provenance of generated claims. Moreover, in order to improve the overall credibility of such technology, we must ensure that adequately calibrated evaluation metrics are used.

7.4 Paths Forward

On a general note, the findings suggest that the epistemic authority of AI in the context of knowledge production and dissemination should be questioned. This requires awareness of the presented issues, as well as a rethinking of *knowledge* in AI engineering. This includes better literacy regarding notions of situatedness and contextuality of knowledge, as well as the importance of diversity not only to remedy ethical concerns, but also to foster epistemically better knowledge production and more objective AI research.

Scientific Rigor

Studies on bias in AI are often focused on easily observable skews in the training data or model outputs. This limited frame can easily lure us into the *solutionism trap* (Selbst et al., 2019), where we try to fix these issues solely on a data- or model-level. It is not my intention to fully discard respective efforts. But I believe that we need to utilize more than one approach to truly improve the limitations addressed in this thesis. Especially since local fixes are at risk of concealing the bigger issues at hand. In this thesis, I have addressed the many ways in which values and

16. https://papareo.nz/#kaupapa, accessed: November 19, 2025

power affect not only the biased composition of the examined technologies and datasets, but also the way research in AI is approached, as a whole. Benchmarks, for instance, play a significant role in steering research efforts in AI (Orr and Kang, 2024; Koch et al., 2021b). However, the way that benchmarks are created is not sufficiently anchored in empirical evidence (Bean et al., 2025). The fact that their biases commonly go unnoticed is probably caused by the lack of reflexivity regarding this practice (Kraft et al., 2025). Hence, this thesis clearly endorses calls for a broader conceptualization of scientific rigor in the AI research community. This includes demands for reflection and transparent reporting of epistemic and normative assumptions (Olteanu et al., 2023).

Fostering Discourse and Amplifying Marginalized Voices

To improve the previously discussed limitations, change is required on the level of scientific practice in AI. Of course, this is not easily realized. In the words of Dotson (2014), this requires *third-order change*: members of an epistemic community, working under a given epistemic framework, must firstly recognize the limitations of its very framework. Moreover, the cost of changing the framework is (cognitively and logistically) high and so is the resistance to it, accordingly (Dotson, 2014, 2012). A valuable lessons we may draw from Longino's feminist empiricism in this regard is that we should foster and promote *public venues* that are open to reflexive critique. One positive example is FAccT, which has been welcoming scrutiny of its own practices (Laufer et al., 2022; Young et al., 2022; Septiandri et al., 2023).

The overview of the three different sets of RQs inquired in this thesis reveals that all the involved corpora of content are coherently flawed, the demographic composition of those behind the data and those behind the tools are coherently skewed, and the practices are guided by (and contributing to) a coherent *view from nowhere* (Nagel, 1989). Paying attention to and amplifying less dominant counter-narratives can help escape this self-affirming loop. Communities such as *Black in AI*,[17] *Widening NLP*,[18] and *Queer in AI*[19] have been centering and advocating for marginalized communities in AI research. Widening NLP, for instance, organizes workshops at renown international AI conferences to discuss and publish contributions by researchers from underrepresented groups. They also offer peer feedback to future submissions, outside of the formal peer-review process (Tonja et al., 2024). Queer in AI facilitates "community-engaged research", "by the people, for the people" (p. 5), while also engaging in political and social advocacy. They, for example, advocate for name change policies in scientific publishing or offer financial aid to queer students for graduate school applications to strengthen the queer representation in AI (QueerInAI et al., 2023).

Overall it can be said, that the AI community would benefit from a shift in values, towards participatory and emancipatory values (Harding, 1986), and a continued analysis and challenge of power asymmetries (Klein and D'Ignazio, 2024).

17. https://www.blackinai.org/, accessed: November 11, 2025
18. http://www.winlp.org/, accessed: November 11, 2025
19. https://www.queerinai.com/, accessed: November 11, 2025.

8 Conlusion

This thesis investigates epistemic and ethical problems of LMs and KGs, as well as knowledge-enhanced language modeling. The inquiry departs from the empirical observation that respective AI systems have gained significant epistemic authority in recent years. Issues of factual infidelity and social bias in LMs have already received much attention in science, industry, and the public. However, KGs and knowledge-enhancement still appear to be praised as bias-free and factual.

At the heart of this cumulative dissertation are three individual research articles: in Kraft and Usbeck (2022), a systematic literature review on social bias throughout the lifecycle of KGs is conducted. In Kraft and Soulier (2024), we present an interdisciplinary study of knowledge-enhanced language modeling, which claims that knowledge enhancement cannot render LMs bias-proof. This is supported by a bias analysis of respective models, as well as an in-depth philosophical analysis that draws from feminist epistemological accounts. Finally, the study presented in Kraft et al. (2025) uses a mixed-method analysis of bias-sensitive reporting practices and data biases in popular QA and RC benchmarks. The findings reveal a lack of transparency regarding characteristics related to annotator identity, a lack of concern regarding potential biases, and, consequently, data biases related to gender, religion, and geography of entities mentioned in the QA items.

Values and knowledge are closely intertwined. Therefore, techno-scientific knowledge practices must be evaluated along epistemological and ethical dimensions. The KGs and knowledge-enhanced LMs investigated in this thesis consistently fail to represent diverse perspectives, perpetuating hermeneutical injustice. This contradicts the common narrative in the AI community that KGs are a neutralizing remedy. The issue is worsened by the fact that many QA and RC benchmarks are also biased. Even though they are marketed as measuring an LLM's ability to answer questions correctly, on a general level, they happen to mostly measure accuracy regarding questions relevant to a certain demographic. Biased measures not only render model biases invisible. They also reward the optimization of models towards these biases.

These problems can, however, not be resolved by maximizing coverage. Firstly, it is impossible to form a complete repository of all knowledge, because there is no finite, enumerable amount of knowledge. Secondly, completionism is undesirable, because it can motivate colonialist practices. Instead of striving for *complete* coverage, me must strive for *adequate* coverage. This means, technologies should be built to be reliable and useful for the very communities, needs, and purposes they are intended to serve. As it stands, AI technologies are largely framed as general-purpose and (prospectively) all-encompassing. Instead, it should be embraced that science, and technology are inherently situated and value-laden, and that contextualized development and analysis are pre-requisites to *good* technology. The community should promote emancipatory values and foreground those at the margins, to counter existing injustices and foster critical discourse in a diverse community. This can help in identifying issues brought about by power asymmetries and cause change.

The fact that the contributions in this thesis were accepted for publication at internationally renowned venues is a sign that the potential for change is there. However, we must ensure to tackle identified limitations beyond a technological frame and think about solutions on a societal, instead of a merely technical level (Barocas et al., 2023). One call to action to take from this thesis is to persist and to continue critiquing and questioning powerful narratives.

Limitations & Future Work

Even though I firmly believe in the necessity of inter- and transdisciplinarity, it shall be mentioned that it can also pose unique challenges. Inherent to the work of the kind which is presented in this thesis is a trade-off between depth and breadth. Or as Suchman (2007) put it: "certain chapters may seem a bit basic" (p. 4) to one or the other discipline. To the computer scientist, this work might lack quantitative experiments with newer, state-of-the-art models. To justify more generalizable conclusions, more models could have been compared or more datasets, respectively. To the philosopher, the philosophical analysis might lack depth or complexity. To introduce terminology and theory from one discipline to another and to work out the relationships between these theories requires a great deal of effort. And this comes at the cost of depth. But again, interdisciplinary work is necessary as it provides valuable reflections by deriving new analytical perspectives and implications from the rich bodies of knowledge in other disciplines. In particular, to promote a more contextualized understanding of ethical and epistemic questions in AI, interdisciplinary discourse is needed (Raji et al., 2021b).

Between the Discussion presented in the paper Kraft and Usbeck (2022), in Chapter 4, and the Discussion of the overall thesis, in Chapter 7, there is a terminological inconsistency that I would like to address. In the paper, I criticize that open encyclopedic KGs misrepresent "the world as it is". I suggest: "Even if the data represented are not affected by sampling errors, they are restricted to describing *the world as it is* as opposed to *the world as it should be*. We strive for the latter kind of inference basis." (p. 81). With this, I am aiming to convey a normative stance. That is, rather than aiming for systems that model and thereby perpetuate our societal status quo that is characterized by pre-existing,

harmful inequalities, we should strive to correct or–even better–over-correct these inequalities. The wording I chose implies that there is only one way to see the world, which appears inconsistent with the framing taken on later in this thesis, namely that there are "many worlds". The use of the word "world" is, however, slightly different in these two contexts. In drawing from the epistemologies of Longino (2002) and Haraway (1988) the "many worlds" metaphor is intended to convey the existence of competing, changing truths, which makes full coverage of all knowledges infeasible (and undesirable). Nevertheless, I argue for favoring certain norms over others. That is, we need to question pre-existing inequalities and discrimination and strive for a revision of technologies that uphold it. To this end, we need to pursue adequate coverage of individual technologies, as measured by local norms, and we need to uplift less dominant alternative resources, as well (as long as they are epistemically acceptable).

This thesis has repeatedly touched upon the role of Big Tech, but a deeper analysis was beyond the scope. The role of public policy was, moreover, not mentioned at all. Of course, these are important stakeholders that define the status quo in AI development and dissemination, and they have great lever to bring about change. Future work could emphasize these perspectives to draw connections to the findings presented in this thesis. This would help develop a more complete picture of the power dynamics that shape prevalent narratives in AI research and development, for example, regarding objectivity, completeness, correctness, and situatedness. Broader networks of researchers with different disciplinary backgrounds could realize more large-scale analyses of knowledge technology and their underlying practices. Future work should, moreover, involve researchers from a greater variety of social situations, including members of the very communities that are studied (Miceli et al., 2025).

I have already started to work on projects that follow-up on the research presented here. One line of research is concerned with introducing more rigor into the creation of evaluation metrics and benchmarks. Currently, benchmarks are created rather insulated from social context. In drawing from social scientific methods, I am exploring possibilities of creating empirically grounded evaluation methods. A concrete project will be to create a new benchmark that will allow to investigate disparate LM performance on knowledge-related tasks. I am also continuing to collaborate with colleagues trained in philosophy to analyze conceptions of knowledge, power asymmetries, and injustice within AI research practice.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Large Language Models Associate Muslims with Violence. *Nature Machine Intelligence* 3 (6): 461–463. https://doi.org/10.1038/s42256-021-00359-2. (Cited on pages 10, 14, 83).

Alison Adam. 1998. Artificial Knowing: Gender and the Thinking Machine. Routledge. (Cited on pages 19, 72 sqq., 140).

———. 2000. Deleting the Subject: A Feminist Reading of Epistemology in Artificial Intelligence. *Minds and Machines* (USA) 10 (2): 231–253. https://doi.org/10.1023/A:1008306015799. (Cited on pages 11, 73, 95 sqq., 99, 102, 112, 140, 144).

Abien Fred Agarap. 2018. Deep Learning using Rectified Linear Units (ReLU). *CoRR* abs/1803.08375. http://arxiv.org/abs/1803.08375. (Cited on page 25).

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2021. Knowledge Graph Based Synthetic Corpus Generation for Knowledge-Enhanced Language Model Pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2021),* 3554–3565. Online: ACL. https://doi.org/10.18653/v1/2021.naacl-main.278. (Cited on pages 95 sq., 102 sq., 147).

Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu. 2024. Can Knowledge Graphs Reduce Hallucinations in LLMs?: A Survey. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers),* 3947–3960. Mexico City, Mexico: ACL. https://doi.org/10.18653/v1/2024.naacl-long.219. (Cited on pages 6, 36, 40, 95 sq., 101 sq.).

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and Measuring Annotator Bias Based on Annotators' Demographic Characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms,* 184–190. Online: ACL. https://doi.org/10.18653/v1/2020.alw-1.21. (Cited on pages 118, 128).

Aish Albladi, Minarul Islam, Amit Das, Maryam Bigonah, Zheng Zhang, Fatemeh Jamshidi, Mostafa Rahgouy, Nilanjana Raychawdhary, Daniela Marghitu, and Cheryl Seals. 2025. Hate Speech Detection Using Large Language Models: A Comprehensive Review. *IEEE Access* 13:20871–20892. https://doi.org/10.1109/ACCESS.2025.3532397. (Cited on pages 10, 16).

Signe Altmäe, Alberto Sola-Leyva, and Andres Salumets. 2023. Artificial Intelligence in Scientific Writing: A Friend or a Foe? *Reproductive BioMedicine Online* 47 (1): 3–9. https://doi.org/https://doi.org/10.1016/j.rbmo.2023.04.009. (Cited on page 2).

Gabriel Amaral, Alessandro Piscopo, Lucie-Aimée Kaffee, Odinaldo Rodrigues, and Elena Simperl. 2021. Assessing the Quality of Sources in Wikidata Across Languages: A Hybrid Approach. *ACM J. Data Inf. Qual.* 13 (4): 23:1–23:35. https://doi.org/10.1145/3484828. (Cited on page 146).

Shun-ichi Amari. 1993. Backpropagation and Stochastic Gradient Descent Method. *Neurocomputing* 5 (4-5): 185–196. (Cited on page 26).

Elizabeth Anderson. 1995. Knowledge, Human Interests, and Objectivity in Feminist Epistemology. *Philosophical Topics* 23 (2): 27–58. https://www.jstor.org/stable/43154207. (Cited on page 109).

———. 2024. Feminist Epistemology and Philosophy of Science. In *The Stanford Encyclopedia of Philosophy,* Fall 2024, edited by Edward N. Zalta and Uri Nodelman. Metaphysics Research Lab, Stanford University. (Cited on pages 55, 98).

Mario Arduini, Lorenzo Noci, Federico Pirovano, Ce Zhang, Yash Raj Shrestha, and Bibek Paudel. 2020. Adversarial Learning for Debiasing Knowledge Graph Embeddings. In *Proceedings of the 16th International Workshop on Mining and Learning with Graphs (MLG).* https://arxiv.org/pdf/2006.16309.pdf. (Cited on pages 16, 85, 87 sqq.).

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the Cross-lingual Transferability of Monolingual Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 4623–4637. Online: ACL. https://doi.org/10.18653/v1/2020.acl-main.421. (Cited on pages 125, 142).

Ram G. Athreya, Axel-Cyrille Ngonga Ngomo, and Ricardo Usbeck. 2018. Enhancing Community Interactions with Data-Driven Chatbots–The DBpedia Chatbot. In *Companion Proceedings of the The Web Conference 2018 (WWW 2018),* 143–146. Lyon, France: International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3184558.3186964. (Cited on pages 79, 92).

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. 2007a. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007,* 4825:722–735. Lecture Notes in Computer Science. Busan, Republic of Korea: Springer. https://doi.org/10.1007/978-3-540-76298-0_52. (Cited on pages 15, 139).

———. 2007b. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007,* 4825:722–735. Lecture Notes in Computer Science. Busan, Republic of Korea: Springer. https://doi.org/10.1007/978-3-540-76298-0_52. (Cited on page 81).

Sören Auer, Lorenz Bühmann, Christian Dirschl, Orri Erling, Michael Hausenblas, Robert Isele, Jens Lehmann, Michael Martin, Pablo N. Mendes, Bert Van Nuffelen, Claus Stadler, Sebastian Tramp, and Hugh Williams. 2012. Managing the Life-Cycle of Linked Data with the LOD2 Stack. In *The Semantic Web - ISWC 2012 - 11th International Semantic Web Conference, Proceedings, Part II,* 7650:1–16. Lecture Notes in Computer Science. Boston, MA, USA: Springer. https://doi.org/10.1007/978-3-642-35173-0_1. (Cited on page 79).

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022. ReFinED: An Efficient Zero-shot-capable Approach to End-to-End Entity Linking. In *NAACL.* (Cited on page 121).

Stefan Baack. 2024. A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2024),* 2199–2208. Rio de Janeiro, Brazil: ACM. https://doi.org/10.1145/3630106.3659033. (Cited on pages 8 sq.).

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR* abs/2106.15590. https://arxiv.org/abs/1409.0473. (Cited on page 29).

Soumya Barikeri, Anne Lauscher, Ivan Vulic, and Goran Glavas. 2021. RedditBias: A Real-World Resource for Bias Evaluation and Debiasing of Conversational Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, Volume 1: Long Papers,* 1941–1955. ACL. https://doi.org/10.18653/V1/2021.ACL-LONG.151. (Cited on page 50).

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The Problem with Bias: From Allocative to Representational Harms in Machine Learning. In *Proceedings of the SIGCIS conference.* (Cited on pages 42, 118).

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. Fairness and Machine Learning: Limitations and Opportunities. MIT Press. (Cited on pages 42, 46 sq., 76, 107, 149, 153).

Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104 (3): 671–732. http://www.jstor.org/stable/24758720. (Cited on pages 41, 43, 45).

Jean M. Bartunek and Michael K. Moch. 1987. First-Order, Second-Order, and Third-Order Change and Organization Development Interventions: A Cognitive Approach. *The Journal of Applied Behavioral Science* 23 (4): 483–500. https://doi.org/10.1177/002188638702300404. (Cited on page 71).

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, María Grandury, Simeng Han, Valentin Hofmann, Lujain Ibrahim, Hazel Kim, Hannah Rose Kirk, Fangru Lin, Gabrielle Kaili-May Liu, Lennart Luettgau, Jabez Magomere, Jonathan Rystrøm, Anna Sotnikova, Yushi Yang, Yilun Zhao, Adel Bibi, Antoine Bosselut, Ronald Clark, Arman Cohan, Jakob Nicolaus Foerster, Yarin Gal, Scott A. Hale, Inioluwa Deborah Raji, Christopher Summerfield, Philip Torr, Cozmin Ududec, Luc Rocher, and Adam Mahdi. 2025. Measuring what Matters: Construct Validity in Large Language Model Benchmarks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS 2025).* San Diego, CA, USA. https://openreview.net/forum?id=mdA5lVvNcU. (Cited on page 151).

Gábor Bella, Paula Helm, Gertraud Koch, and Fausto Giunchiglia. 2024. Tackling Language Modelling Bias in Support of Linguistic Diversity. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency,* 562–572. Rio de Janeiro, Brazil: ACM. https://doi.org/10.1145/3630106.3658925. (Cited on pages 14, 43).

Emily M. Bender. 2011. On Achieving and Evaluating Language-Independence in NLP. *Linguistic Issues in Language Technology* 6. https://doi.org/10.33011/lilt.v6i.1239. (Cited on pages 46, 128).

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6:587–604. https://doi.org/10.1162/TACL_A_00041. (Cited on pages 90, 110, 118).

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2021),* 610–623. Online: ACM. https://doi.org/10.1145/3442188.3445922. (Cited on pages 3, 102, 147).

Emily M. Bender and Alex Hanna. 2025. The AI Con. Penguin Books Limited. (Cited on pages 34, 37).

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A Neural Probabilistic Language Model. *Journal of Machine Learning Research* 3:1137–1155. https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf. (Cited on pages 27 sqq.).

Ruha Benjamin. 2019. Race After Technology: Abolitionist Tools for the New Jim Code. John Wiley & Sons. (Cited on page 43).

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic Parsing on Freebase from Question-Answer Pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing,* 1533–1544. Seattle, Washington, USA: ACL. https://aclanthology.org/D13-1160/. (Cited on page 123).

Glen Berman. 2025. GenAI is an Epistemic Carcinogen. *AI and Society,* 1–3. https://doi.org/10.1007/s00146-025-02537-x. (Cited on pages 2, 6).

Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The Semantic Web: A New Form of Web Content That is Meaningful to Computers Will Unleash a Revolution of New Possibilities. *Scientific American* 284 (5): 35–43. (Cited on pages 1 sq.).

Pablo Beytía, Pushkal Agarwal, Miriam Redi, and Vivek K. Singh. 2022. Visual Gender Biases in Wikipedia: A Systematic Evaluation across the Ten Most Spoken Languages. In *Proceedings of the International AAAI Conference on Web and Social Media,* 43–54. https://ojs.aaai.org/index.php/ICWSM/article/view/19271. (Cited on pages 82, 84 sq., 92, 139).

Sumithra Bhakthavatsalam, Chloe Anastasiades, and Peter Clark. 2020. GenericsKB: A Knowledge Base of Generic Statements. *CoRR* abs/2005.00660. https://arxiv.org/abs/2005.00660. (Cited on page 84).

Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Re-contextualizing Fairness in NLP: The Case of India. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers),* 727–740. Online: ACL. https://doi.org/10.18653/v1/2022.aacl-main.55. (Cited on pages 16, 51).

Steven Bird. 2020. Decolonising Speech and Language Technology. In *Proceedings of the 28th International Conference on Computational Linguistics,* 3504–3519. Barcelona, Spain (Online): International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.313. (Cited on page 147).

Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022a. Power to the People? Opportunities and Challenges for Participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO 2022),* 6:1–6:8. Arlington, VA, USA: ACM. https://doi.org/10.1145/3551624.3555290. (Cited on pages 109, 111, 128).

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022b. The Values Encoded in Machine Learning Research. In *FAccT 2022: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022,* 173–184. ACM. https://doi.org/10.1145/3531146.3533083. (Cited on page 120).

Abeba Birhane, Vinay Uday Prabhu, Sanghyun Han, Vishnu Boddeti, and Sasha Luccioni. 2023. Into the LAION's Den: Investigating Hate in Multimodal Datasets. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023.* NeurIPS 2023. New Orleans, LA, USA. http://papers.nips.cc/paper%5C_files/paper/2023/hash/42f225509e8263e2043c9d834ccd9a2b-Abstract-Datasets%5C_and%5C_Benchmarks.html. (Cited on pages 10, 44).

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes. *CoRR* abs/2110.01963. https://arxiv.org/abs/2110.01963. (Cited on pages 10, 44).

Abeba Birhane, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy. 2022c. The Forgotten Margins of AI Ethics. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2022),* 948–958. Seoul, Republic of Korea: ACM. https://doi.org/10.1145/3531146.3533157. (Cited on page 16).

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. PIQA: Reasoning about Physical Commonsense in Natural Language. In *AAAI Conference on Artificial Intelligence.* https://api.semanticscholar.org/CorpusID:208290939. (Cited on page 125).

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 5454–5476. Online: ACL. https://doi.org/10.18653/v1/2020.acl-main.485. (Cited on pages 10, 42, 80, 90 sq., 118, 128, 138 sq.).

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data (SIGMOD 2008),* 1247–1250. Vancouver, Canada: ACM. http://portal.acm.org/citation.cfm?id=1376746#. (Cited on page 139).

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems,* vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf. (Cited on pages 16, 45, 50, 83, 85 sq., 88, 91).

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale Simple Question Answering with Memory Networks. *CoRR* abs/1506.02075. http://arxiv.org/abs/1506.02075. (Cited on pages 86, 88).

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems,* vol. 26. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pdf. (Cited on page 85).

Styliani Bourli and Evaggelia Pitoura. 2020. Bias in Knowledge Graph Embeddings. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2020),* 6–10. https://doi.org/10.1109/ASONAM49781.2020.9381459. (Cited on pages 48, 50, 85 sqq.).

T. Bowell. 2025. Feminist Standpoint Theory. In *Internet Encyclopedia of Philosophy,* edited by James Fieser and Bradley Dowden. (Cited on pages 56 sq.).

Geoffrey C. Bowker and Susan Leigh Star. 2000. Sorting Things Out: Classification and its Consequences. MIT press. (Cited on pages 10, 44).

Samuel R. Bowman and George Dahl. 2021. What Will it Take to Fix Benchmarking in Natural Language Understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 4843–4855. Online: ACL. https://doi.org/10.18653/v1/2021.naacl-main.385. (Cited on pages 43, 46, 116, 118 sq., 142 sq., 145).

Freddy Brasileiro, João Paulo A. Almeida, Victorio A. Carvalho, and Giancarlo Guizzardi. 2016. Applying a Multi-Level Modeling Theory to Assess Taxonomic Hierarchies in Wikidata. In *Proceedings of the 25th International Conference Companion on World Wide Web,* 975–980. WWW 2016 Companion. Montréal, Québec, Canada: International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/2872518.2891117. (Cited on page 39).

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems,* 33:1877–1901. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf. (Cited on pages 8, 32, 83).

Jude Browne. 2023. AI and Structural Injustice: A Feminist Perspective. In *Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines.* Oxford University Press. https://doi.org/10.1093/oso/9780192889898.003.0019. (Cited on page 109).

Susan Brownmiller. 1999. In Our Time: Memoir of a Revolution. New York: The Dial Press. (Cited on page 11).

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering Latent Knowledge in Language Models Without Supervision. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023.* OpenReview.net. https://openreview.net/forum?id=ETKGuby0hcs. (Cited on page 99).

Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R. Banaji. 2022. Gender Bias in Word Embeddings: A Comprehensive Analysis of Frequency, Syntax, and Semantics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES 2022),* 156–170. Oxford, United Kingdom: ACM. https://doi.org/10.1145/3514094.3534162. (Cited on page 45).

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics Derived Automatically from Language Corpora Contain Human-like Biases. *Science* 356 (6334): 183–186. https://doi.org/10.1126/science.aal4230. (Cited on pages 48, 105).

Marçal Mora Cantallops, Salvador Sánchez-Alonso, and Elena García-Barriocanal. 2019. A Systematic Literature Review on Wikidata. *Data Technol. Appl.* 53 (3): 250–268. https://doi.org/10.1108/DTA-12-2018-0110. (Cited on pages 5, 39).

Nancy Cartwright. 1983. How the Laws of Physics Lie. Oxford University Press. https://doi.org/10.1093/0198247044.001.0001. (Cited on page 63).

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphaël Ségerie, Micah Carroll, Andi Peng, Phillip J. K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca D. Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. 2023. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *Transactions of Machine Learning Research* 2023. https://openreview.net/forum?id=bx24KpJ4Eb. (Cited on page 33).

Nilesh Chakraborty, Denis Lukovnikov, Gaurav Maheshwari, Priyansh Trivedi, Jens Lehmann, and Asja Fischer. 2021. Introduction to Neural Network-Based Question Answering Over Knowledge Graphs. *WIREs Data Mining and Knowledge Discovery* 11 (3): e1389. https://doi.org/10.1002/widm.1389. (Cited on pages 79, 139).

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. (Cited on page 119).

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014),* 1724–1734. Doha, Qatar: ACL. https://doi.org/10.3115/v1/D14-1179. (Cited on page 29).

Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5 (2): 153–163. https://arxiv.org/abs/1610.07524. (Cited on pages 47, 87).

Yu-Neng Chuang, Kwei-Herng Lai, Ruixiang Tang, Mengnan Du, Chia-Yuan Chang, Na Zou, and Xia Hu. 2025. Fair-RGNN: Mitigating Relational Bias on Knowledge Graphs. *ACM Trans. Knowl. Discov. Data* (New York, NY, USA) 19 (2). https://doi.org/10.1145/3681792. (Cited on page 16).

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),* 2924–2936. Minneapolis, Minnesota: ACL. https://doi.org/10.18653/v1/N19-1300. (Cited on page 123).

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *CoRR* abs/1803.05457. http://arxiv.org/abs/1803.05457. (Cited on page 123).

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. *CoRR* abs/2110.14168. https://arxiv.org/abs/2110.14168. (Cited on pages 34, 123, 125).

Patricia Hill Collins. 1990. Black Feminist Thought: Knowledge, Conciousness, and the Politics of Empowerment. New York: Routledge. (Cited on pages 55 sq., 58, 65).

Nick Couldry and Ulises A. Mejias. 2019. Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject. *Television & New Media* 20 (4): 336–349. https://doi.org/10.1177/1527476418796632. (Cited on pages 147, 149).

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis* 16 (1): 64–93. https://doi.org/10.1093/jla/laae003. (Cited on page 6).

Paramita Das, Sai Keerthana Karnam, Anirban Panda, Bhanu Prakash Reddy Guda, Soumya Sarkar, and Animesh Mukherjee. 2023. Diversity Matters: Robustness of Bias Measurements in Wikidata. In *Proceedings of the 15th ACM Web Science Conference 2023 (WebSci 2023),* 208–218. Austin, TX, USA: ACM. https://doi.org/10.1145/3578503.3583620. (Cited on pages 96, 103).

Paramita Das, Sai Keerthana Karnam, Aditya Bharat Soni, and Animesh Mukherjee. 2025. Social Biases in Knowledge Representations of Wikidata separates Global North from Global South. In *Proceedings of the 17th ACM Web Science Conference 2025 (Websci 2025),* 12–21. New Brunswick, NJ, USA: ACM. https://doi.org/10.1145/3717867.3717882. (Cited on page 15).

Shib Sankar Dasgupta, Swayambhu Nath Ray, and Partha Talukdar. 2018. HyTE: Hyperplane-based Temporally aware Knowledge Graph Embedding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,* 2001–2011. Brussels, Belgium: ACL. https://doi.org/10.18653/v1/D18-1225. (Cited on page 87).

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *CoRR* abs/2501.12948. https://doi.org/10.48550/ARXIV.2501.12948. (Cited on page 34).

Pieter Delobelle, Ewoenam Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring Fairness with Biased Rulers: A Comparative Study on Bias Metrics for Pre-trained Language Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 1693–1706. Seattle, USA: ACL. https://doi.org/10.18653/v1/2022.naacl-main.122. (Cited on pages 14, 50).

Gianluca Demartini. 2019. Implicit Bias in Crowdsourced Knowledge Graphs. In *Companion Proceedings of The 2019 World Wide Web Conference (WWW 2019),* 624–630. San Francisco, USA: ACM. https://doi.org/10.1145/3308560.3317307. (Cited on pages 82, 85, 90).

Nathaniel Demchak, Xin Guan, Zekun Wu, Ziyi Xu, Adriano Koshiyama, and Emre Kazim. 2024. Assessing Bias in Metric Models for LLM Open-Ended Generation Bias Benchmarks. In *Proceedings of the Workshop Evaluating Evaluations: Examining Best Practices for Measuring Broader Impacts of Generative AI co-located with NeurIPS 2024.* (Cited on pages 116, 119).

René Descartes. 2012. Discourse on Method. Hackett Publishing. (Cited on page 97).

Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. On Measures of Biases and Harms in NLP. In *Findings of the Association for Computational Linguistics (AACL-IJCNLP 2022),* 246–267. Online: ACL. https://aclanthology.org/2022.findings-aacl.24. (Cited on pages 102, 105).

Sunipa Dev, Emily Sheng, Jieyu Zhao, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Nanyun Peng, and Kai-Wei Chang. 2021. What do Bias Measures Measure? *CoRR* abs/2108.03362. https://arxiv.org/abs/2108.03362. (Cited on page 91).

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "Gender" in NLP Bias Research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2022),* 2083–2102. Seoul, Republic of Korea: ACM. https://doi.org/10.1145/3531146.3534627. (Cited on pages 51, 138).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),* 4171–4186. Minneapolis, Minnesota: ACL. https://doi.org/10.18653/v1/N19-1423. (Cited on pages 5, 14, 31, 83).

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2021),* 862–872. Online: ACM. https://doi.org/10.1145/3442188.3445924. (Cited on pages 10, 50, 84, 119).

Dennis Diefenbach, Vanessa López, Kamal Deep Singh, and Pierre Maret. 2018. Core Techniques of Question Answering Systems Over Knowledge Bases: A Survey. *Knowledge and Information Systems* 55 (3): 529–569. https://doi.org/10.1007/s10115-017-1100-y. (Cited on pages 79, 139).

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing,* 1286–1305. Online and Punta Cana, Dominican Republic: ACL. https://doi.org/10.18653/v1/2021.emnlp-main.98. (Cited on pages 10, 44).

Tommaso Dolci, Fabio Azzalini, and Mara Tanelli. 2023. Improving Gender-Related Fairness in Sentence Encoders: A Semantics-based Approach. *Data Science and Engineering* 8 (2): 177–195. (Cited on page 48).

Kristie Dotson. 2012. A Cautionary Tale: On Limiting Epistemic Oppression. *Frontiers: A Journal of Women Studies* 33 (1): 24–47. Accessed October 31, 2025. http://www.jstor.org/stable/10.5250/fronjwomestud.33.1.0024. (Cited on pages 70 sqq., 151).

———. 2014. Conceptualizing Epistemic Oppression. *Social Epistemology* 28 (2): 115–138. https://doi.org/10.1080/02691728.2013.782585. (Cited on pages 71, 150 sq.).

Yupei Du, Qi Zheng, Yuanbin Wu, Man Lan, Yan Yang, and Meirong Ma. 2022. Understanding Gender Bias in Knowledge Base Embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 1381–1395. Dublin, Ireland: ACL. https://doi.org/10.18653/v1/2022.acl-long.98. (Cited on pages 85 sqq., 102).

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),* 2368–2378. Minneapolis, Minnesota: ACL. https://doi.org/10.18653/v1/N19-1246. (Cited on pages 123, 125 sq.).

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS 2012),* 214–226. Cambridge, Massachusetts: ACM. https://doi.org/10.1145/2090236.2090255. (Cited on page 86).

Hady ElSahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon S. Hare, Frédérique Laforest, and Elena Simperl. 2018. T-REx: A Large Scale Alignment of Natural Language With Knowledge Base Triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* Miyazaki, Japan: ELRA. http://www.lrec-conf.org/proceedings/lrec2018/summaries/632.html. (Cited on pages 15, 106, 141).

Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandecic. 2014. Introducing Wikidata to the Linked Data Web. In *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Proceedings, Part I,* 8796:50–65. Lecture Notes in Computer Science. Riva del Garda, Italy: Springer. https://doi.org/10.1007/978-3-319-11964-9_4. (Cited on page 39).

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2024),* 6491–6501. Barcelona, Spain: ACM. https://doi.org/10.1145/3637528.3671470. (Cited on pages 7, 16, 140).

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. 2024. LawBench: Benchmarking Legal Knowledge of Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing,* 7933–7962. Miami, Florida, USA: ACL. https://doi.org/10.18653/v1/2024.emnlp-main.452. (Cited on page 5).

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 259–268. Sydney, NSW, Australia: ACM. https://doi.org/10.1145/2783258.2783311. (Cited on page 107).

Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. WinoQueer: A Community-in-the-Loop Benchmark for Anti-LGBTQ+ Bias in Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 9126–9140. Toronto, Canada: ACL. https://doi.org/10.18653/v1/2023.acl-long.507. (Cited on page 14).

Constanza Fierro, Ruchira Dhar, Filippos Stamatiou, Nicolas Garneau, and Anders Søgaard. 2024. Defining Knowledge: Bridging Epistemology and Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing,* 16096–16111. Miami, Florida, USA: ACL. https://doi.org/10.18653/v1/2024.emnlp-main.900. (Cited on page 36).

Joseph Fisher, Arpit Mittal, Dave Palfrey, and Christos Christodoulopoulos. 2020a. Debiasing Knowledge Graph Embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020),* 7332–7345. Online: ACL. https://doi.org/10.18653/v1/2020.emnlp-main.595. (Cited on pages 85, 88 sqq.).

Joseph Fisher, Dave Palfrey, Christos Christodoulopoulos, and Arpit Mittal. 2020b. Measuring Social Bias in Knowledge Graph Embeddings. In *Proceedings of the AKBC 2020 Workshop on Bias in Automatic Knowledge Graph Construction.* https://www.amazon.science/publications/measuring-social-bias-in-knowledge-graph-embeddings. (Cited on pages 85 sqq., 89 sq.).

Luciano Floridi. 2025. Correction to: A Conjecture on a Fundamental Trade-Off Between Certainty and Scope in Symbolic and Generative AI. *Philosophy & Technology* 38 (4): 133. (Cited on pages 5 sq.).

Luciano Floridi and Jeff W. Sanders. 2004. On the Morality of Artificial Agents. *Minds and Machines* 14:349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d. (Cited on page 99).

Richard Foley. 1987. The Theory of Epistemic Rationality. Harvard University Press. (Cited on page 99).

Diana E. Forsythe. 1993. Engineering Knowledge: The Construction of Knowledge in Artificial Intelligence. *Social Studies of Science* 23 (3): 445–477. Accessed December 22, 2023. http://www.jstor.org/stable/370256. (Cited on pages 11, 73, 95, 98, 100, 103, 112, 140, 144).

Miranda Fricker. 1999. Epistemic Oppression and Epistemic Privilege. *Canadian Journal of Philosophy* 29 (sup1): 191–210. https://doi.org/10.1080/00455091.1999.10716836. (Cited on page 71).

———. 2007. Epistemic Injustice: Power and the Ethics of Knowing. Oxford University Press. (Cited on pages 11, 66 sqq., 71, 96 sq., 108, 118, 128, 141, 147, 150).

Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. *ACM Transactions on Information Systems* 14 (3): 330–347. https://doi.org/10.1145/230538.230561. (Cited on pages 10, 41 sqq., 45, 80, 90, 102).

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics* (Cambridge, MA) 50 (3): 1097–1179. https://doi.org/10.1162/coli_a_00524. (Cited on pages 42, 48 sqq., 116, 118, 138).

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural Approaches to Conversational AI. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2018),* 1371–1374. Ann Arbor, MI, USA: ACM. https://doi.org/10.1145/3209978.3210183. (Cited on page 79).

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2021. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *CoRR* abs/2101.00027. https://arxiv.org/abs/2101.00027. (Cited on page 8).

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey, https://doi.org/10.48550/arXiv.2312.10997. (Cited on page 103).

Andrew Gaut, Tony Sun, Shirlyn Tang, Yuxin Huang, Jing Qian, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2020. Towards Understanding Gender Bias in Relation Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 2943–2953. Online: ACL. https://doi.org/10.18653/v1/2020.acl-main.265. (Cited on pages 83, 85, 102).

Timnit Gebru. 2021. Hierarchy of Knowledge in Machine Learning & Related Fields & Its Consequences. In *Carnegie Mellon Human-Computer Interaction Institute Seminar Series.* (Video recording). https://scs.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=70f6edd7-de91-464e-ae94-acbb011ba2c7. (Cited on page 99).

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. *Communications of the ACM* (New York, NY, USA) 64 (12): 86–92. https://doi.org/10.1145/3458723. (Cited on pages 90, 110, 118, 149).

R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage In, Garbage Out? Do Machine Learning Application Papers in Social Computing Report Where Human-labeled Training Data Comes From? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* 2020),* 325–336. Barcelona, Spain: ACM. https://doi.org/10.1145/3351095.3372862. (Cited on page 118).

James Geller and Navya Martin Kollapally. 2021. Detecting, Reporting And Alleviating Racial Biases In Standardized Medical Terminologies And Ontologies. In *Proceedings of the 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM),* 1–5. Houston, TX, USA: IEEE. https://doi.org/10.1109/BIBM52615.2021.9669617. (Cited on pages 82, 85).

Emma J. Gerritse, Faegheh Hasibi, and Arjen P. de Vries. 2020. Bias in Conversational Search: The Double-Edged Sword of the Personalized Knowledge Graph. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval (ICTIR 2020),* 133–136. Online: ACM. https://doi.org/10.1145/3409256.3409834. (Cited on page 79).

Lise Getoor and Ben Taskar. 2007. Introduction to Statistical Relational Learning. The MIT Press. https://doi.org/10.7551/mitpress/7432.001.0001. (Cited on page 79).

Edmund L. Gettier. 1963. Is Justified True Belief Knowledge? *Analysis* 23 (6): 121–123. (Cited on page 54).

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019),* 1161–1166. Hong Kong, China: ACL. https://doi.org/10.18653/v1/D19-1107. (Cited on page 118).

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *Transactions of the Association for Computational Linguistics* (Cambridge, MA) 9:346–361. https://doi.org/10.1162/tacl_a_00370. (Cited on pages 123, 125).

Peter Glick and Susan T Fiske. 1997. Hostile and Benevolent Sexism: Measuring Ambivalent Sexist Attitudes Toward Women. *Psychology of Women Quarterly* 21 (1): 119–135. https://doi.org/10.1111/j.1471-6402.1997.tb00104.x. (Cited on page 89).

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic Bias Metrics Do Not Correlate with Application Bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers),* 1926–1940. Online: ACL. https://doi.org/10.18653/v1/2021.acl-long.150. (Cited on pages 14, 50, 91).

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of the 2019 Workshop on Widening NLP,* 60–63. Florence, Italy: ACL. https://aclanthology.org/W19-3621. (Cited on pages 16, 51, 88 sqq.).

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. Deep Learning. Cambridge, MA, USA: MIT Press. http://www.deeplearningbook.org. (Cited on pages 25 sqq., 29).

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 Task 7: Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012),* 394–398. Montréal, Canada: ACL. https://aclanthology.org/S12-1052/. (Cited on page 123).

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting Bias and Knowledge Acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction (AKBC 2013),* 25–30. San Francisco, California, USA: ACM. https://doi.org/10.1145/2509558.2509563. (Cited on page 84).

Heidi Grasswick. 2018. Feminist Social Epistemology. In *The Stanford Encyclopedia of Philosophy,* Fall 2018, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University. (Cited on pages 53, 55).

Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *Journal of Personality and Social Psychology* 74 (6): 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464. (Cited on page 48).

Ramón Grosfoguel. 2007. The Epistemic Decolonial Turn: Beyond Political-Economy Paradigms. *Cultural Studies* 21 (2-3): 211–223. https://www.tandfonline.com/doi/full/10.1080/09502380601162514. (Cited on page 98).

Xin Guan, Nathaniel Demchak, Saloni Gupta, Ze Wang, Ediz Ertekin Jr., Adriano S. Koshiyama, Emre Kazim, and Zekun Wu. 2024. SAGED: A Holistic Bias-Benchmarking Pipeline for Language Models with Customisable Fairness Calibration. *CoRR* abs/2409.11149. https://doi.org/10.48550/ARXIV.2409.11149. (Cited on page 119).

Wei Guo and Aylin Caliskan. 2021. Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES 2021),* 122–133. Online: ACM. https://doi.org/10.1145/3461702.3462536. (Cited on pages 14, 45, 48).

Edward J. Hackett, Olga Amsterdamska, Michael Lynch, and Judy Wajcman. 2007. The Handbook of Science and Technology Studies. MIT Press. (Cited on page 54).

Ian Hacking. 1992. Statistical Language, Statistical Truth, and Statistical Reason: The Self-Authentication of a Style of Scientific Reasoning. In *Social Dimensions of Scientific Knowledge,* edited by Ernan McMullin. University of Notre Dame Press. (Cited on page 63).

Luke Haliburton, Jan Leusmann, Robin Welsch, Sinksar Ghebremedhin, Petros Isaakidis, Albrecht Schmidt, and Sven Mayer. 2024. Uncovering Labeler Bias in Machine Learning Annotation Tasks. *AI and Ethics,* 1–14. (Cited on page 125).

Xu Han, Tianyu Gao, Yuan Yao, Deming Ye, Zhiyuan Liu, and Maosong Sun. 2019. OpenNRE: An Open and Extensible Toolkit for Neural Relation Extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing: System Demonstrations (EMNLP-IJCNLP 2019),* 169–174. Hong Kong, China: ACL. https://doi.org/10.18653/v1/D19-3029. (Cited on page 83).

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021. Diverse Adversaries for Mitigating Bias in Training. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume,* 2760–2765. Online: ACL. https://doi.org/10.18653/v1/2021.eacl-main.239. (Cited on page 50).

Leif Hancox-Li and I. Elizabeth Kumar. 2021. Epistemic Values in Feature Importance Methods: Lessons from Feminist Epistemology. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2021),* 817–826. Online: ACM. https://doi.org/10.1145/3442188.3445943. (Cited on pages 18, 99).

Donna Haraway. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies* 14 (3): 575–599. https://www.jstor.org/stable/3178066?seq=1. (Cited on pages 57, 59, 73, 118, 140, 146, 154).

———. 2016. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. In *Space, Gender, Knowledge: Feminist Readings,* 53–72. Routledge. https://www.jstor.org/stable/3178066. (Cited on pages 19, 96, 98).

Sandra Harding. 1986. The Science Question in Feminism. Cornell University Press. (Cited on pages 56, 58 sq., 65, 129, 150 sq.).

———. 2007. Feminist Standpoints. In *Handbook of Feminist Research: Theory and Praxis,* edited by Sharlene Nagy Hesse-Biber, 45–69. SAGE Publications. (Cited on pages 56, 58, 146).

———. 2013. Rethinking Standpoint Epistemology: What is "Strong Objectivity"? In *Feminist Epistemologies,* 49–82. Routledge. https://www.jstor.org/stable/23739232. (Cited on pages 55, 59, 65, 96 sqq., 110, 148).

John Hardwig. 1985. Epistemic Dependence. *The Journal of Philosophy* 82 (7): 335–349. (Cited on page 97).

Zellig S. Harris. 1954. Distributional structure. *Word* 10 (2-3): 146–162. https://doi.org/10.1080/00437956.1954.11659520. (Cited on page 28).

David Hartmann, Amin Oueslati, Dimitri Staufer, Lena Pohlmann, Simon Munzert, and Hendrik Heuer. 2025. Lost in Moderation: How Commercial Content Moderation APIs Over- and Under-Moderate Group-Targeted Hate Speech and Linguistic Variations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI 2025).* New York, NY, USA: ACM. https://doi.org/10.1145/3706598.3713998. (Cited on page 10).

Stefan Heindorf, Yan Scholten, Gregor Engels, and Martin Potthast. 2019. Debiasing Vandalism Detection Models at Wikidata. In *The World Wide Web Conference (WWW 2019),* 670–680. San Francisco, CA, USA: ACM. https://doi.org/10.1145/3308558.3313507. (Cited on page 39).

Paula Helm, Gábor Bella, Gertraud Koch, and Fausto Giunchiglia. 2024. Diversity and Language Technology: How Language Modeling Bias Causes Epistemic Injustice. *Ethics Inf. Technol.* 26 (1): 8. https://doi.org/10.1007/S10676-023-09742-6. (Cited on pages 12, 14, 46, 118, 128, 139).

Paula Helm, Amalia de Götzen, Luca Cernuzzi, Alethia Hume, Shyam Diwakar, Salvador Ruiz Correa, and Daniel Gatica-Perez. 2023. Diversity and Neocolonialism in Big Data Research: Avoiding Extractivism While Struggling with Paternalism. *Big Data & Society* 10 (2): 20539517231206802. https://doi.org/10.1177/20539517231206802. (Cited on page 147).

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021 (NeurIPS 2021).* Online. https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html. (Cited on page 123).

Anna Lauren Hoffmann. 2021. Terms of Inclusion: Data, Discourse, Violence. *New Media & Society* 23 (12): 3539–3556. https://journals.sagepub.com/doi/10.1177/1461444820958725. (Cited on page 109).

Konrad Höffner, Sebastian Walter, Edgard Marx, Ricardo Usbeck, Jens Lehmann, and Axel-Cyrille Ngonga Ngomo. 2017. Survey on Challenges of Question Answering in the Semantic Web. *Semantic Web* 8 (6): 895–920. https://doi.org/10.3233/SW-160247. (Cited on pages 79, 139).

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. AI Generates Covertly Racist Decisions About People Based on their Dialect. *Nature* 633 (8028): 147–154. (Cited on page 14).

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021a. Knowledge Graphs. Synthesis Lectures on Data, Semantics, and Knowledge. Morgan & Claypool Publishers. https://doi.org/10.2200/S01125ED1V01Y202109DSK022. (Cited on pages 38 sq.).

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutiérrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan F. Sequeda, Steffen Staab, and Antoine Zimmermann. 2021b. Knowledge Graphs. *ACM Computing Surveys* 54 (4): 71:1–71:37. https://doi.org/10.1145/3447772. (Cited on pages 3, 6, 37).

Katharina Hoppe. 2022. Donna Haraway zur Einführung. Junius Verlag GmbH. (Cited on page 57).

Kurt Hornik, Maxwell B. Stinchcombe, and Halbert White. 1989. Multilayer Feedforward Networks are Universal Approximators. *Neural Networks* 2 (5): 359–366. https://doi.org/10.1016/0893-6080(89)90020-8. (Cited on page 25).

Seyed Amir Hosseini Beghaeiraveri, Alasdair Gray, and Fiona McNeill. 2024. RQSS: Referencing Quality Scoring System for Wikidata. *Semantic Web* 15 (6): 2419–2475. (Cited on page 146).

Dirk Hovy and Shrimai Prabhumoye. 2021. Five Sources of Bias in Natural Language Processing. *Language and Linguistics Compass* 15 (8). https://doi.org/10.1111/LNC3.12432. (Cited on pages 10, 41, 43 sq., 46, 90 sq., 138, 146).

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *CoRR* abs/1508.01991. http://arxiv.org/abs/1508.01991. (Cited on page 83).

Jonathan Jenkins Ichikawa and Matthias Steup. 2024. The Analysis of Knowledge. In *The Stanford Encyclopedia of Philosophy,* Fall 2024, edited by Edward N. Zalta and Uri Nodelman. Metaphysics Research Lab, Stanford University. (Cited on page 53).

Filip Ilievski, Pedro A. Szekely, and Daniel Schwabe. 2020. Commonsense Knowledge in Wikidata. In *Proceedings of the 1st Wikidata Workshop (Wikidata 2020) co-located with 19th International Semantic Web Conference.* OPub 2020. Online: CEUR-ws.org. https://ceur-ws.org/Vol-2773/paper-10.pdf. (Cited on page 104).

Kristen Intemann. 2010. 25 Years of Feminist Empiricism and Standpoint Theory: Where Are We Now? *Hypatia* 25 (4): 778–796. (Cited on pages 55 sq., 65, 76, 148).

Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2021),* 375–385. Online: ACM. https://doi.org/10.1145/3442188.3445901. (Cited on pages 89, 139).

Sofia Jaime and Christoph Kern. 2024. Ethnic Classifications in Algorithmic Fairness: Concepts, Measures and Implications in Practice. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2024),* 237–253. Rio de Janeiro, Brazil: ACM. https://doi.org/10.1145/3630106.3658902. (Cited on page 11).

Krzysztof Janowicz, Bo Yan, Blake Regalia, Rui Zhu, and Gengchen Mai. 2018. Debiasing Knowledge Graphs: Why Female Presidents are not like Female Popes. In *Proceedings of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks co-located with 17th International Semantic Web Conference (ISWC 2018).* http://ceur-ws.org/Vol-2180/ISWC%5C_2018%5C_Outrageous%5C_Ideas%5C_paper%5C_17.pdf. (Cited on pages 82, 84 sq., 139).

Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge Graph Embedding via Dynamic Mapping Matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers),* 687–696. Beijing, China: ACL. https://doi.org/10.3115/v1/P15-1067. (Cited on page 87).

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys* 55 (12). https://doi.org/10.1145/3571730. (Cited on pages 6, 34 sqq., 95, 101, 145).

Longquan Jiang and Ricardo Usbeck. 2022. Knowledge Graph Question Answering Datasets and Their Generalizability: Are They Enough for Future Research? In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2022),* 3209–3218. Madrid, Spain: ACM. https://doi.org/10.1145/3477495.3531751. (Cited on pages 79, 92, 139).

Yuchen Jiang, Xiang Li, Hao Luo, Shen Yin, and Okyay Kaynak. 2022. Quo Vadis Artificial Intelligence? *Discover Artificial Intelligence* 2 (1). https://doi.org/10.1007/s44163-022-00022-8. (Cited on page 100).

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2024. KoBBQ: Korean Bias Benchmark for Question Answering. *Transactions of the Association for Computational Linguistics* (Cambridge, MA) 12:507–524. https://doi.org/10.1162/tacl_a_00661. (Cited on page 118).

Eun Seo Jo and Timnit Gebru. 2020. Lessons From Archives: Strategies for Collecting Sociocultural Data in Machine Learning. In *Conference on Fairness, Accountability, and Transparency (FAT* 2020),* 306–316. Barcelona, Spain: ACM. https://doi.org/10.1145/3351095.3372829. (Cited on pages 110 sq.).

Isaac Johnson, Lucie-Aimée Kaffee, and Miriam Redi. 2024. Wikimedia data for AI: a review of Wikimedia datasets for NLP tasks and AI-assisted editing. In *Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia,* 91–101. Miami, Florida, USA: ACL. https://doi.org/10.18653/v1/2024.wikinlp-1.14. (Cited on pages 5, 12).

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 1601–1611. Vancouver, Canada: ACL. https://doi.org/10.18653/v1/P17-1147. (Cited on page 123).

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 6282–6293. Online: ACL. https://doi.org/10.18653/v1/2020.acl-main.560. (Cited on page 142).

Dan Jurafsky and James H. Martin. 2009. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2nd Edition. Prentice Hall Series in Artificial Intelligence. Prentice Hall, Pearson Education International. https://www.worldcat.org/oclc/315913020. (Cited on page 101).

Daniel Jurafsky and James H. Martin. 2025. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models. 3rd. Online manuscript released January 12, 2025. https://web.stanford.edu/~jurafsky/slp3/. (Cited on pages 24 sq., 27 sq.).

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language Models (Mostly) Know What They Know. *CoRR* abs/2207.05221. https://doi.org/10.48550/ARXIV.2207.05221. (Cited on page 100).

Lucie-Aimée Kaffee, Kemele M. Endris, and Elena Simperl. 2019. When humans and machines collaborate: cross-lingual label editing in wikidata. In *Proceedings of the 15th International Symposium on Open Collaboration (OpenSym 2019).* Skövde, Sweden: ACM. https://doi.org/10.1145/3306446.3340826. (Cited on page 5).

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large Language Models Struggle to Learn Long-Tail Knowledge. In *Proceedings of the 40th International Conference on Machine Learning,* 202:15696–15707. Proceedings of Machine Learning Research. PMLR. https://proceedings.mlr.press/v202/kandpal23a.html. (Cited on pages 8, 15 sq., 144).

Immanuel Kant. 2013. An Answer to the Question: 'What is Enlightenment?' Penguin UK. (Cited on page 97).

Andrej Karpathy. 2023. Lecture: Introduction to Large Language Models. https://www.youtube.com/watch?v=zjkBMFhNj_g. (Cited on page 6).

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2024. A Survey of Reinforcement Learning from Human Feedback. https://arxiv.org/abs/2312.14925. (Cited on page 33).

Jackie Kay, Atoosa Kasirzadeh, and Shakir Mohamed. 2024. Epistemic Injustice in Generative AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES 2024)* 7 (1): 684–697. https://doi.org/10.1609/aies.v7i1.31671. (Cited on pages 11 sq., 118).

C. Maria Keet. 2021. An Exploration Into Cognitive Bias in Ontologies. In *Joint Ontology Workshops 2021 Episode VII: The Bolzano Summer of Knowledge (JOWO 2021).* Bolzano, Italy: CEUR-ws.org. http://ceur-ws.org/Vol-2969/paper38-CAOS.pdf. (Cited on pages 82, 85, 103).

Daphna Keidar, Mian Zhong, Ce Zhang, Yash Raj Shrestha, and Bibek Paudel. 2021. Towards Automatic Bias Detection in Knowledge Graphs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021),* 3804–3811. Online and Punta Cana, Dominican Republic: ACL. https://doi.org/10.18653/v1/2021.findings-emnlp.321. (Cited on pages 85 sqq., 89, 92).

Os Keyes and Kathleen Creel. 2022. Artificial Knowing Otherwise. *Feminist Philosophy Quarterly* 8 (3). https://doi.org/10.5206/fpq/2022.3/4.14313. (Cited on pages 99, 111).

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics,* 43–53. New Orleans, Louisiana: Association for Computational Linguistics. https://doi.org/10.18653/v1/S18-2005. (Cited on pages 10, 14, 102, 118).

Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, and Yuki Asano. 2021. Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models. In *Advances in Neural Information Processing Systems,* 34:2611–2624. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2021/file/1531beb762df4029513ebf9295e0d34f-Paper.pdf. (Cited on pages 14, 83).

Lauren Klein and Catherine D'Ignazio. 2024. Data Feminism for AI. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2024),* 100–112. Rio de Janeiro, Brazil: ACM. https://doi.org/10.1145/3630106.3658543. (Cited on page 151).

Karin Knorr-Cetina. 1983. The Ethnographic Study of Scientific Work. In *Science Sbserved: Perspectives on the Social Study of Science,* edited by Karin Knorr-Cetina and Michael Mulkay, 115–140. SAGE Publications. (Cited on page 54).

Bernard Koch, Emily Denton, Alex Hanna, and Jacob G Foster. 2021a. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021 (NeurIPS 2021).* Online. https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/3b8a614226a953a8cd9526fca6fe9ba5-Paper-round2.pdf. (Cited on page 91).

———. 2021b. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2021),* 560–575. Online: ACM. https://doi.org/10.1145/3442188.3445918. (Cited on pages 7, 17, 116, 145, 151).

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS 2022).* New Orleans, LA, USA: Curran Associates Inc. (Cited on pages 34 sq.).

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender Bias and Stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference (CI 2023),* 12–24. Delft, Netherlands: ACM. https://doi.org/10.1145/3582269.3615599. (Cited on pages 14, 116, 118, 138).

Angelie Kraft. 2021. Triggering Models: Measuring and Mitigating Bias in German Language Generation. Master's thesis, Universität Hamburg. https://doi.org/10.5281/zenodo.13837954. (Cited on pages 3, 16, 30, 51).

Angelie Kraft, Judith Simon, and Sonja Schimmler. 2025. Social Bias in Popular Question-Answering Benchmarks. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics,* 1421–1438. Mumbai, India and Online: AFNLP / ACL. https://aclanthology.org/2025.ijcnlp-long.79/. (Cited on pages vi, 7, 17, 20 sqq., 142, 145, 148, 151 sq.).

Angelie Kraft and Eloïse Soulier. 2024. Knowledge-Enhanced Language Models Are Not Bias-Proof: Situated Knowledge and Epistemic Injustice in AI. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2024),* 1433–1445. Rio de Janeiro, Brazil: ACM. https://doi.org/10.1145/3630106.3658981. (Cited on pages v, 12, 15 sqq., 20, 22, 118, 140, 147, 152).

Angelie Kraft and Ricardo Usbeck. 2022. The Lifecycle of "Facts": A Survey of Social Bias in Knowledge Graphs. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, Volume 1: Long Papers (AACL-IJCNLP 2022),* 639–652. Online: ACL. https://doi.org/10.18653/v1/2022.aacl-main.49. (Cited on pages v, vii, 9, 20, 22, 102, 104, 110, 138 sq., 147, 152 sq.).

Angelie Kraft, Hans-Peter Zorn, Pascal Fecht, Judith Simon, Chris Biemann, and Ricardo Usbeck. 2022. Measuring Gender Bias in German Language Generation. In *52. Jahrestagung der Gesellschaft für Informatik, INFORMATIK 2022, Informatik in den Naturwissenschaften,* vol. P-326, 1257–1274. Hamburg, Germany: Gesellschaft für Informatik, Bonn. https://doi.org/10.18420/INF2022_108. (Cited on pages 10, 45, 51, 91).

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. BioASQ-QA: A Manually Curated Corpus for Biomedical Question Answering. *Scientific Data* 10 (1): 170. (Cited on page 123).

Rajeev Kumar, Harishankar Kumar, and Kumari Shalini. 2025. Detecting and Mitigating Bias in LLMs through Knowledge Graph-Augmented Training. In *2025 International Conference on Artificial Intelligence and Data Engineering (AIDE),* 608–613. https://doi.org/10.1109/AIDE64228.2025.10987418. (Cited on page 16).

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring Bias in Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing,* 166–172. Florence, Italy: ACL. https://doi.org/10.18653/v1/W19-3823. (Cited on page 83).

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* (Cambridge, MA) 7:452–466. https://doi.org/10.1162/tacl_a_00276. (Cited on pages 123, 125).

Viet Dac Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. In *Findings of the Association for Computational Linguistics (EMNLP 2023),* 13171–13189. Singapore: ACL. https://doi.org/10.18653/v1/2023.findings-emnlp.878. (Cited on page 12).

Shyong K. Lam, Anuradha Uduwage, Zhenhua Dong, Shilad Sen, David R. Musicant, Loren G. Terveen, and John Riedl. 2011. WP: Clubhouse?: An Exploration of Wikipedia's Gender Imbalance. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration, 2011,* 1–10. Mountain View, CA, USA: ACM. https://doi.org/10.1145/2038558.2038560. (Cited on page 104).

Bruno Latour. 1983. Give Me a Laboratory and I Will Move the World. In *Science observed: Perspectives on the social study of science,* edited by Karin Knorr-Cetina and Michael Mulkay, 141–170. SAGE Publications. (Cited on page 54).

Benjamin Laufer, Sameer Jain, A. Feder Cooper, Jon Kleinberg, and Hoda Heidari. 2022. Four Years of FAccT: A Reflexive, Mixed-Methods Analysis of Research Contributions, Shortcomings, and Future Prospects. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2022),* 401–426. Seoul, Republic of Korea: ACM. https://doi.org/10.1145/3531146.3533107. (Cited on page 151).

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gérard Dupont, Stella Biderman, Anna Rogers, Loubna Benallal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Romero Muñoz, Jian Zhu, Daniel van Strien, Zaid Alyafeai, Khalid Almubarak, Vu Minh Chien, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Ifeoluwa Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilić, Margaret Mitchell, Sasha Luccioni, and Yacine Jernite. 2022. The BigScience ROOTS corpus: a 1.6TB composite multilingual dataset. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS 2022).* New Orleans, LA, USA: Curran Associates Inc. (Cited on page 8).

Julien Leblay and Melisachew Wudage Chekol. 2018. Deriving Validity Time in Knowledge Graph. In *Companion Proceedings of the The Web Conference 2018 (WWW 2018),* 1771–1776. Lyon, France: International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3184558.3191639. (Cited on page 87).

Messi H.J. Lee, Jacob M. Montgomery, and Calvin K. Lai. 2024. Large Language Models Portray Socially Subordinate Groups as More Homogeneous, Consistent with a Bias Observed in Humans. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2024),* 1321–1340. Rio de Janeiro, Brazil: ACM. https://doi.org/10.1145/3630106.3658975. (Cited on page 10).

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020 (NeurIPS 2020).* Online. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html. (Cited on pages 7, 40, 95, 101 sq.).

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The Dawn After the Dark: An Empirical Study on Factuality Hallucination in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 10879–10899. Bangkok, Thailand: ACL. https://doi.org/10.18653/V1/2024.ACL-LONG.586. (Cited on pages 6, 8, 15, 34).

Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. 2021a. Pretrained Language Models for Text Generation: A Survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI 2021),* 4492–4499. Montréal, Canada: IJCAI. https://doi.org/10.24963/ijcai.2021/612. (Cited on page 95).

Luoqiu Li, Xiang Chen, Hongbin Ye, Zhen Bi, Shumin Deng, Ningyu Zhang, and Huajun Chen. 2021b. On Robustness and Bias Analysis of BERT-Based Relation Extraction. In *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction,* 43–59. Singapore: Springer Singapore. https://doi.org/https://doi.org/10.1007/978-981-16-6471-7_4. (Cited on pages 83, 85).

Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. UNQOVERing Stereotyping Biases via Underspecified Questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020),* 3475–3489. Online: ACL. https://doi.org/10.18653/v1/2020.findings-emnlp.311. (Cited on page 14).

Yingji Li, Mengnan Du, Xin Wang, and Ying Wang. 2023. Prompt Tuning Pushes Farther, Contrastive Learning Pulls Closer: A Two-Stage Approach to Mitigate Social Biases. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 14254–14267. Toronto, Canada: ACL. https://doi.org/10.18653/v1/2023.acl-long.797. (Cited on page 50).

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards Understanding and Mitigating Social Biases in Language Models. In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021),* 139:6565–6576. Online: PMLR. http://proceedings.mlr.press/v139/liang21a.html. (Cited on page 102).

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D. Manning, and James Y. Zou. 2024. Mapping the Increasing Use of LLMs in Scientific Papers. *CoRR* abs/2404.01268. https://doi.org/10.48550/ARXIV.2404.01268. (Cited on page 2).

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's Verify Step by Step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024.* OpenReview.net. https://openreview.net/forum?id=v8L0pN6EOi. (Cited on page 34).

LimeSurvey Project Team / Carsten Schmitz. 2012. LimeSurvey: An Open Source survey tool. Hamburg, Germany: LimeSurvey Project. https://www.limesurvey.org. (Cited on page 120).

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 3214–3252. Dublin, Ireland: ACL. https://doi.org/10.18653/v1/2022.acl-long.229. (Cited on pages 122 sq., 125).

Zichao Lin, Shuyan Guan, Wending Zhang, Huiyan Zhang, Yugang Li, and Huaping Zhang. 2024. Towards Trustworthy LLMs: A Review on Debiasing and Dehallucinating in Large Language Models. *Artificial Intelligence Review* 57 (9): 243. (Cited on page 50).

Nora Freya Lindemann. 2024. Chatbots, Search Engines, and the Sealing of Knowledges. *AI & Society,* https://doi.org/10.1007/s00146-024-01944-w. (Cited on pages 12, 99).

Shlomit Aharoni Lir. 2021. Strangers in a Seemingly Open-to-all Website: The Gender Bias in Wikipedia. *Equality, Diversity and Inclusion: An International Journal* 40 (7): 2040–7149. https://doi.org/10.1108/EDI-10-2018-0198. (Cited on page 104).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692. http://arxiv.org/abs/1907.11692. (Cited on pages 31 sq., 105, 113, 141).

Helen Longino. 1990. Science as Social Knowledge: Values and Objectivity in Scientific Inquiry. Princeton University Press. Accessed November 17, 2023. http://www.jstor.org/stable/j.ctvx5wbfz. (Cited on pages 96, 109).

———. 1993. Feminist Standpoint Theory and the Problems of Knowledge. In *Signs,* 19:201–212. 1. The University of Chicago Press. (Cited on page 64).

———. 2002. The Fate of Knowledge. Princeton University Press. https://doi.org/doi:10.1515/9780691187013. (Cited on pages 6, 13, 17, 54 sq., 60 sqq., 74, 97, 144, 146, 148, 150, 154).

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2024. A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers),* 3245–3276. Mexico City, Mexico: ACL. https://doi.org/10.18653/v1/2024.naacl-long.179. (Cited on page 12).

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS 2022).* New Orleans, LA, USA: Curran Associates Inc. (Cited on page 125).

Alexandra Luccioni and Joseph Viviano. 2021. What's in the Box? An Analysis of Undesirable Content in the Common Crawl Corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers),* 182–189. Online: ACL. https://doi.org/10.18653/v1/2021.acl-short.24. (Cited on page 9).

Kristian Lum, Jacy Reese Anthis, Kevin Robinson, Chirag Nagpal, and Alexander Nicholas D'Amour. 2025. Bias in Language Models: Beyond Trick Tests and Towards RUTEd Evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 137–161. Vienna, Austria: ACL. https://doi.org/10.18653/v1/2025.acl-long.7. (Cited on pages 14, 50).

Xu Luo, Hao Wu, Ji Zhang, Lianli Gao, Jing Xu, and Jingkuan Song. 2023. A Closer Look at Few-shot Classification Again. In *International Conference on Machine Learning, ICML 2023,* 202:23103–23123. Proceedings of Machine Learning Research. Honolulu, Hawaii, USA: PMLR. https://proceedings.mlr.press/v202/luo23e.html. (Cited on page 32).

Jeffrey Jun-jie Ma and Charles Chuankai Zhang. 2023. Understanding Structured Knowledge Production: A Case Study of Wikidata's Representation Injustice. *CoRR* abs/2311.02767. https://doi.org/10.48550/ARXIV.2311.02767. (Cited on page 103).

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. 2025. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. *Journal of Empirical Legal Studies* 22 (2): 216–242. (Cited on pages 7, 145).

Negar Maleki, Balaji Padmanabhan, and Kaushik Dutta. 2024. AI Hallucinations: A Misnomer Worth Clarifying. In *2024 IEEE Conference on Artificial Intelligence (CAI),* 133–138. https://doi.org/10.1109/CAI59869.2024.00033. (Cited on page 17).

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 9802–9822. Toronto, Canada: ACL. https://doi.org/10.18653/v1/2023.acl-long.546. (Cited on page 102).

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations,* 55–60. Baltimore, Maryland: ACL. https://doi.org/10.3115/v1/P14-5010. (Cited on pages 3, 82).

Anqi Mao, Mehryar Mohri, and Yutao Zhong. 2023. Cross-Entropy Loss Functions: Theoretical Analysis and Applications. In *International Conference on Machine Learning, ICML 2023,* 202:23803–23828. Proceedings of Machine Learning Research. Honolulu, Hawaii, USA: PMLR. https://proceedings.mlr.press/v202/mao23b.html. (Cited on page 26).

Gary Marcus. 2020. The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence. *CoRR* abs/2002.06177. https://arxiv.org/abs/2002.06177. (Cited on pages 100, 147).

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 3190–3199. https://doi.org/10.1109/CVPR.2019.00331. (Cited on page 125).

José-Lázaro Martínez-Rodríguez, Aidan Hogan, and Ivan López-Arévalo. 2020. Information Extraction Meets the Semantic Web: A Survey. *Semantic Web* 11 (2): 255–335. https://doi.org/10.3233/SW-180333. (Cited on pages 79, 92, 99, 102).

A. P. Martinich and Avrum Stroll. 2025. Epistemology. In *Encyclopedia Britannica.* https://www.britannica.com/topic/epistemology. (Cited on page 52).

Rebecca Mason. 2011. Two Kinds of Unknowing. *Hypatia* 26 (2): 294–307. https://doi.org/10.1111/j.1527-2001.2011.01175.x. (Cited on pages 11, 70, 72, 108, 147 sq., 150).

Hakim El Massari, Noreddine Gherabi, Fatima Qanouni, Sajida Mhammedi, and Mohamed Amnai. 2024. The Role of Artificial Intelligence in the Semantic Web. In *2024 10th International Conference on Optimization and Applications (ICOA),* 1–6. https://doi.org/10.1109/ICOA62581.2024.10753971. (Cited on page 2).

Dominic Masters and Carlo Luschi. 2018. Revisiting Small Batch Training for Deep Neural Networks. *CoRR* abs/1804.07612. http://arxiv.org/abs/1804.07612. (Cited on page 26).

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019),* 5267–5275. Hong Kong, China: ACL. https://doi.org/10.18653/v1/D19-1530. (Cited on page 50).

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On Measuring Social Biases in Sentence Encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),* 622–628. Minneapolis, Minnesota: ACL. https://doi.org/10.18653/v1/N19-1063. (Cited on pages 14, 48, 105).

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 1878–1898. Dublin, Ireland: ACL. https://doi.org/10.18653/v1/2022.acl-long.132. (Cited on pages 90, 106).

Ninareh Mehrabi, Thamme Gowda, Fred Morstatter, Nanyun Peng, and Aram Galstyan. 2020. Man is to Person as Woman is to Location: Measuring Gender Bias in Named Entity Recognition. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media (HT 2020),* 231–232. Online: ACM. https://doi.org/10.1145/3372923.3404804. (Cited on pages 82, 85, 102, 139).

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021a. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys* 54 (6). https://doi.org/10.1145/3457607. (Cited on page 81).

Ninareh Mehrabi, Pei Zhou, Fred Morstatter, Jay Pujara, Xiang Ren, and Aram Galstyan. 2021b. Lawyers are Dishonest? Quantifying Representational Harms in Commonsense Knowledge Resources. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021),* 5016–5033. Online and Punta Cana, Dominican Republic: ACL. https://doi.org/10.18653/v1/2021.emnlp-main.410. (Cited on pages 84 sq., 89 sq., 92, 139).

Beatrice Melis, Chiara Paolini, Marta Fioravanti, and Daniele Metilli. 2024. What Does it Mean to be Queer in Wikidata? Practices of Gender Representation Within a Transnational Online Community. *Communication, Culture and Critique* 17 (3): 200–207. https://doi.org/10.1093/ccc/tcae029. (Cited on page 15).

Sara Melotte, Filip Ilievski, Linglan Zhang, Aditya Malte, Namita Mutha, Fred Morstatter, and Ninareh Mehrabi. 2022. Where Does Bias in Common Sense Knowledge Models Come From? *IEEE Internet Computing* 26 (4): 12–20. https://doi.org/10.1109/MIC.2022.3170914. (Cited on pages 104, 111).

Amanda Menking and Ingrid Erickson. 2015. The Heart Work of Wikipedia: Gendered, Emotional Labor in the World's Largest Online Encyclopedia. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI 2015),* 207–210. Seoul, Republic of Korea: ACM. https://doi.org/10.1145/2702123.2702514. (Cited on pages 104, 109).

Amanda Menking, Ingrid Erickson, and Wanda Pratt. 2019. People Who Can Take It: How Women Wikipedians Negotiate and Navigate Safety. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI 2019),* 472. Glasgow, Scotland, UK: ACM. https://doi.org/10.1145/3290605.3300702. (Cited on pages 104, 148).

Amanda Menking and Jon Rosenberg. 2021. WP:NOT, WP:NPOV, and Other Stories Wikipedia Tells Us: A Feminist Critique of Wikipedia's Epistemology. *Science, Technology, & Human Values* 46 (3): 455–479. https://doi.org/10.1177/0162243920924783. (Cited on pages 104, 128, 148).

Milagros Miceli, Adio-Adet Dinika, Krystal Kauffman, Camilla Salim Wagner, Laurenz Sachenbacher, Alex Hanna, and Timnit Gebru. 2025. Methodological Considerations for Centering Workers' Epistemic Authority in AI Research. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES 2025)* 8 (2): 1698–1710. https://doi.org/10.1609/aies.v8i2.36667. (Cited on page 154).

Milagros Miceli and Julian Posada. 2022. The Data-Production Dispositif. *Proc. ACM Hum.-Comput. Interact.* (New York, NY, USA) 6 (CSCW2). https://doi.org/10.1145/3555561. (Cited on page 118).

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018),* 2381–2391. Brussels, Belgium: ACL. https://doi.org/10.18653/v1/D18-1260. (Cited on pages 123 sq.).

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR),* 1–12. Scottsdale, AZ, USA. http://arxiv.org/abs/1301.3781. (Cited on pages 28, 83).

Charles W. Mills. 2017. Ideology. In *The Routledge Handbook of Epistemic Injustice,* 100–111. Routledge London. (Cited on pages 70, 109).

Martin Miragoli. 2025. Conformism, Ignorance & Injustice: AI as a Tool of Epistemic Oppression. *Episteme* 22 (2): 522–540. https://doi.org/10.1017/epi.2024.11. (Cited on page 12).

Shubhanshu Mishra, Sijun He, and Luca Belli. 2020. Assessing Demographic Bias in Named Entity Recognition. In *Proceedings of the AKBC 2020 Workshop on Bias in Automatic Knowledge Graph Construction.* https://kg-bias.github.io/NER_Bias_KG_Bias.pdf. (Cited on pages 83, 85, 102).

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* 2019),* 220–229. Atlanta, GA, USA: ACM. https://doi.org/10.1145/3287560.3287596. (Cited on page 90).

Brent Mittelstadt, Sandra Wachter, and Chris Russell. 2023. To Protect Science, we Must Use LLMs as Zero-Shot Translators. *Nature Human Behaviour* 7:1830–1832. https://doi.org/10.1038/s41562-023-01744-0. (Cited on page 95).

Warmhold Jan Thomas Mollema. 2025. A Taxonomy of Epistemic Injustice in the Context of AI and the Case for Generative Hermeneutical Erasure. *CoRR* abs/2504.07531. https://doi.org/10.48550/ARXIV.2504.07531. (Cited on page 12).

Cedric Möller, Jens Lehmann, and Ricardo Usbeck. 2022. Survey on English Entity Linking on Wikidata: Datasets and Approaches. *Semantic Web* 13 (6): 925–966. https://www.semantic-web-journal.net/content/survey-english-entity-linking-wikidata-0. (Cited on page 102).

Evgeny Morozov. 2013. To Save Everything, Click Here: Technology, Solutionism and the Urge to Fix Problems that Don't Exist. 432. PublicAffairs. (Cited on page 102).

Claudia Müller-Birn, Benjamin Karran, Janette Lehmann, and Markus Luczak-Rösch. 2015. Peer-Production System or Collaborative Ontology Engineering Effort: What is Wikidata? In *Proceedings of the 11th International Symposium on Open Collaboration (OpenSym 2015).* San Francisco, California: ACM. https://doi.org/10.1145/2788993.2789836. (Cited on page 5).

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring Stereotypical Bias in Pretrained Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL-IJCNLP 2021, Volume 1: Long Papers,* 5356–5371. Online: ACL. https://doi.org/10.18653/v1/2021.acl-long.416. (Cited on pages 14, 49, 102, 105, 118).

Thomas Nagel. 1989. The View From Nowhere. Oxford University Press. (Cited on pages 57, 98 sq., 151).

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020),* 1953–1967. Online: ACL. https://doi.org/10.18653/v1/2020.emnlp-main.154. (Cited on pages 14, 48, 105).

Arvind Narayanan and Sayash Kapoor. 2024. AI Snake Oil. USA: Princeton University Press. (Cited on page 9).

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in Large Language Models: Origins, Inventory, and Discussion. *J. Data and Information Quality* (New York, NY, USA) 15 (2). https://doi.org/10.1145/3597307. (Cited on pages 10, 16, 116).

Andrei Nesterov, Laura Hollink, and Jacco van Ossenbruggen. 2024. How Contentious Terms About People and Cultures are Used in Linked Open Data. In *Proceedings of the ACM Web Conference 2024 (WWW 2024),* 4523–4533. Singapore, Singapore: ACM. https://doi.org/10.1145/3589334.3648140. (Cited on pages 15, 104).

Axel-Cyrille Ngonga Ngomo, Mohamed Ahmed Sherif, Kleanthi Georgala, Mofeed Mohamed Hassan, Kevin Dreßler, Klaus Lyko, Daniel Obraczka, and Tommaso Soru. 2021. LIMES: A Framework for Link Discovery on the Semantic Web. *Künstliche Intelligenz* 35 (3): 413–423. https://doi.org/10.1007/s13218-021-00713-x. (Cited on pages 79, 92).

Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2016. A Review of Relational Machine Learning for Knowledge Graphs. *Proceedings of the IEEE* 104 (1): 11–33. https://doi.org/10.1109/JPROC.2015.2483592. (Cited on page 85).

Ikujiro Nonaka and Hirotaka Takeuchi. 1995. The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation. Oxford University Press. https://doi.org/10.1093/oso/9780195092691.001.0001. (Cited on page 37).

Sven Nyholm and Lily Eva Frank. 2017. From Sex Robots to Love Robots: Is Mutual Love With a Robot Possible? In *Robot Sex: Social and Ethical Implications.* (Cited on page 99).

Cathy O'Neil. 2016. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. USA: Crown Publishing Group. (Cited on page 41).

Cailin O'Connor, Sanford Goldberg, and Alvin Goldman. 2024. Social Epistemology. In *The Stanford Encyclopedia of Philosophy,* Summer 2024, edited by Edward N. Zalta and Uri Nodelman. Metaphysics Research Lab, Stanford University. (Cited on page 54).

Alexandra Olteanu, Su Lin Blodgett, Agathe Balayn, Angelina Wang, Fernando Diaz, Flavio du Pin Calmon, Margaret Mitchell, Michael Ekstrand, Reuben Binns, and Solon Barocas. 2025. Rigor in AI: Doing Rigorous AI Work Requires a Broader, Responsible AI-Informed Conception of Rigor, https://www.microsoft.com/en-us/research/publication/rigor-in-ai-doing-rigorous-ai-work-requires-a-broader-responsible-ai-informed-conception-of-rigor/. (Cited on pages 17, 145, 149 sq.).

Alexandra Olteanu, Michael D. Ekstrand, Carlos Castillo, and Jina Suh. 2023. Responsible AI Research Needs Impact Statements Too. https://doi.org/10.48550/ARXIV.2311.11776. (Cited on page 151).

OpenAI. 2023. GPT-4 Technical Report. *CoRR* abs/2303.08774. https://doi.org/10.48550/ARXIV.2303.08774. (Cited on page 32).

———. 2025. GPT-5 System Card. https://cdn.openai.com/gpt-5-system-card.pdf. (Cited on page 7).

Will Orr and Kate Crawford. 2024. The Social Construction of Datasets: On the Practices, Processes, and Challenges of Dataset Creation for Machine Learning. *New Media & Society* 26 (9): 4955–4972. https://doi.org/10.1177/14614448241251797. (Cited on pages 8, 10).

Will Orr and Edward B. Kang. 2024. AI as a Sport: On the Competitive Epistemologies of Benchmarking. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2024),* 1875–1884. Rio de Janeiro, Brazil: ACM. https://doi.org/10.1145/3630106.3659012. (Cited on pages 7, 151).

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training Language Models to Follow Instructions with Human Feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS 2022).* New Orleans, LA, USA: Curran Associates Inc. (Cited on pages 27, 33).

Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and Damien Graux. 2023. Large Language Models and Knowledge Graphs: Opportunities and Challenges. *TGDK* 1 (1): 2:1–2:38. https://doi.org/10.4230/TGDK.1.1.2. (Cited on pages 95, 147).

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering,* 1–20. https://doi.org/10.1109/TKDE.2024.3352100. (Cited on pages 6, 95, 100 sq., 113).

Srikant Panda, Amit Agarwal, and Hitesh Laxmichand Patel. 2025. AccessEval: Benchmarking Disability Bias in Large Language Models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing,* 32492–32518. Suzhou, China: ACL. https://doi.org/10.18653/v1/2025.emnlp-main.1653. (Cited on page 14).

Evangelos Paparidis and Konstantinos Kotis. 2021. Towards Engineering Fair Ontologies: Unbiasing a Surveillance Ontology. In *2021 IEEE International Conference on Progress in Informatics and Computing (PIC),* 226–231. https://doi.org/10.1109/PIC53636.2021.9687030. (Cited on pages 82, 85).

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018),* 2799–2804. Brussels, Belgium: ACL. https://doi.org/10.18653/v1/D18-1302. (Cited on page 118).

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A Hand-built Bias Benchmark for Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2022,* 2086–2105. Dublin, Ireland: ACL. https://doi.org/10.18653/v1/2022.findings-acl.165. (Cited on pages 14, 118 sq.).

Heiko Paulheim. 2017. Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. *Semantic Web* 8 (3): 489–508. https://doi.org/10.3233/SW-160218. (Cited on pages 79, 101).

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (Dis)contents: A Survey of Dataset Development and Use in Machine Learning Research. *Patterns* 2 (11): 100336. https://doi.org/10.1016/J.PATTER.2021.100336. (Cited on pages 25, 128).

Jiaxin Pei and David Jurgens. 2023. When Do Annotator Demographics Matter? Measuring the Influence of Annotator Demographics with the POPQUORN Dataset. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII),* 252–265. Toronto, Canada: ACL. https://doi.org/10.18653/v1/2023.law-1.25. (Cited on pages 45, 118, 128).

Jeffrey Pennington, Richard Socher, and Christopher Manning. n.d. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014),* 1532–1543. Doha, Qatar: ACL. https://doi.org/10.3115/v1/D14-1162. (Cited on page 83).

Caroline Criado Perez. 2019. Invisible Women: Exposing Data Bias in a World Designed for Men. Chatto & Windus. (Cited on page 44).

Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-Supervised Sequence Tagging with Bidirectional Language Models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 1756–1765. Vancouver, Canada: ACL. https://doi.org/10.18653/v1/P17-1161. (Cited on page 83).

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge Enhanced Contextual Word Representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019),* 43–54. Hong Kong, China: ACL. https://doi.org/10.18653/v1/D19-1005. (Cited on pages 79, 139).

Andrew J. Peterson. 2025. AI and the Problem of Knowledge Collapse. *AI and Society* 40 (5): 3249–3269. https://doi.org/10.1007/S00146-024-02173-X. (Cited on page 9).

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019),* 2463–2473. Hong Kong, China: ACL. https://doi.org/10.18653/v1/D19-1250. (Cited on pages 6, 14, 17, 101 sq., 105, 107, 113, 141).

Alessandro Piscopo, Chris Phethean, and Elena Simperl. 2017. What Makes a Good Collaborative Knowledge Graph: Group Composition and Quality in Wikidata. In *Social Informatics,* 305–322. Cham: Springer International Publishing. (Cited on pages 5, 39).

Alessandro Piscopo and Elena Simperl. 2018. Who Models the World? Collaborative Ontology Creation and User Roles in Wikidata. *Proc. ACM Hum.-Comput. Interact.* 2 (CSCW). https://doi.org/10.1145/3274410. (Cited on page 39).

———. 2019. What We Talk About When We Talk About Wikidata Quality: A Literature Survey. In *Proceedings of the 15th International Symposium on Open Collaboration (OpenSym 2019).* Skövde, Sweden: ACM. https://doi.org/10.1145/3306446.3340822. (Cited on page 5).

Andrea J. Pitts. 2017. Decolonial Praxis and Epistemic Injustice. In *The Routledge Handbook of Epistemic Injustice,* 149–157. Routledge London. (Cited on page 98).

Flor Miriam Plaza-del-Arco, Amanda Cercas Curry, Susanna Paoli, Alba Cercas Curry, and Dirk Hovy. 2024. Divine LLaMAs: Bias, Stereotypes, Stigmatization, and Emotion Representation of Religion in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024,* 4346–4366. Miami, Florida, USA: ACL. https://doi.org/10.18653/v1/2024.findings-emnlp.251. (Cited on page 14).

Hannah Powers, Ioana Baldini, Dennis Wei, and Kristin P. Bennett. 2024. Statistical Bias in Bias Benchmark Design. In *Proceedings of the Workshop Evaluating Evaluations: Examining Best Practices for Measuring Broader Impacts of Generative AI co-located with NeurIPS 2024.* (Cited on pages 116, 119).

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2025. A Survey of Multilingual Large Language Models. *Patterns* 6 (1). (Cited on page 14).

Yujia Qin, Yankai Lin, Ryuichi Takanobu, Zhiyuan Liu, Peng Li, Heng Ji, Minlie Huang, Maosong Sun, and Jie Zhou. 2021. ERICA: Improving Entity and Relation Understanding for Pre-trained Language Models via Contrastive Learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL-IJCNLP 2021, Volume 1: Long Papers,* 3350–3363. Online: ACL. https://doi.org/10.18653/v1/2021.acl-long.260. (Cited on page 103).

Organizers Of QueerInAI, Anaelia Ovalle, Arjun Subramonian, Ashwin Singh, Claas Voelcker, Danica J. Sutherland, Davide Locatelli, Eva Breznik, Filip Klubicka, Hang Yuan, Hetvi J, Huan Zhang, Jaidev Shriram, Kruno Lehman, Luca Soldaini, Maarten Sap, Marc Peter Deisenroth, Maria Leonor Pacheco, Maria Ryskina, Martin Mundt, Milind Agarwal, Nyx Mclean, Pan Xu, A Pranav, Raj Korpan, Ruchira Ray, Sarah Mathew, Sarthak Arora, St John, Tanvi Anand, Vishakha Agrawal, William Agnew, Yanan Long, Zijie J. Wang, Zeerak Talat, Avijit Ghosh, Nathaniel Dennler, Michael Noseworthy, Sharvani Jha, Emi Baylor, Aditya Joshi, Natalia Y. Bilenko, Andrew Mcnamara, Raphael Gontijo-Lopes, Alex Markham, Evyn Dong, Jackie Kay, Manu Saraswat, Nikhil Vytla, and Luke Stark. 2023. Queer In AI: A Case Study in Community-Led Participatory AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2023),* 1882–1895. Chicago, IL, USA: ACM. https://doi.org/10.1145/3593013.3594134. (Cited on page 151).

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. *OpenAI Blog,* https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf. (Cited on page 32).

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models Are Unsupervised Multitask Learners. *OpenAI Blog,* https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. (Cited on pages 32, 83, 104).

Wessel Radstok, Melisachew Wudage Chekol, and Mirko T. Schäfer. 2021. Are Knowledge Graph Embedding Models Biased, or Is it the Data That They Are Trained on? In *Proceedings of the 2nd Wikidata Workshop (Wikidata 2021) co-located with the 20th International Semantic Web Conference (ISWC 2021).* http://ceur-ws.org/Vol-2982/paper-5.pdf. (Cited on pages 84 sq., 87 sqq.).

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS 2023).* New Orleans, LA, USA: Curran Associates Inc. (Cited on page 33).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21:140:1–140:67. http://jmlr.org/papers/v21/20-074.html. (Cited on pages 8, 44, 104).

Inioluwa Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021a. AI and the Everything in the Whole Wide World Benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021.* Online. https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/084b6fbb10729ed4da8c3d3f5a3ae7c9-Abstract-round2.html. (Cited on pages 7, 43, 46, 116 sqq., 127, 142).

Inioluwa Deborah Raji, Morgan Klaus Scheuerman, and Razvan Amironesei. 2021b. You Can't Sit With Us: Exclusionary Pedagogy in AI Ethics Education. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2021),* 515–525. Online: ACM. https://doi.org/10.1145/3442188.3445914. (Cited on pages 110 sq., 153).

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing,* 2383–2392. Austin, Texas: ACL. https://doi.org/10.18653/v1/D16-1264. (Cited on pages 122 sq.).

Charle Rathkopf. 2023. Do LLMs Believe. Talk at Philosophy and Theory of AI Conference, Erlangen. (Cited on page 99).

G. Pradeep Reddy, Y. V. Pavan Kumar, and K. Purna Prakash. 2024. Hallucinations in Large Language Models (LLMs). In *2024 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream),* 1–6. https://doi.org/10.1109/eStream61684.2024.10542617. (Cited on pages 35 sqq., 40).

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics* (Cambridge, MA) 7:249–266. https://doi.org/10.1162/tacl_a_00266. (Cited on pages 123, 125).

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. In *First Conference on Language Modeling.* https://openreview.net/forum?id=Ti67584b98. (Cited on pages 123 sqq.).

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. 2024. BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices. Vancouver, BC, Canada. (Cited on pages 119, 128).

Petar Ristoski, Jessica Rosati, Tommaso Di Noia, Renato De Leone, and Heiko Paulheim. 2019. RDF2Vec: RDF Graph Embeddings and Their Applications. *Semantic Web* 10 (4): 721–752. https://doi.org/10.3233/SW-180317. (Cited on page 85).

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020),* 5418–5426. Online: ACL. https://doi.org/10.18653/V1/2020.EMNLP-MAIN.437. (Cited on page 100).

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension. *ACM Computing Surveys* 55 (10): 197:1–197:45. https://doi.org/10.1145/3560260. (Cited on pages 7, 15, 117).

Salvatore Romano, Natalie Kerby, Riccardo Angius, Simone Robutti, Miazia Schueler, Marc Faddoul, Raziye Buse Çetin, Clara Helming, Angela Müller, Matthias Spielkamp, Anna Lena Schiller, Waldemar Kesler, Melis Omalar, Marc Thümmler, Mira Zimmermann, Isabel Sanchez, Alexandra Kimel, Estelle Pannatier, Tobias Urech, Denis Sorie, Michele Loi, and Alex Felder. 2023. Generative AI and Elections: Are Chatbots a Reliable Source of Information for Voters? AI Forensics, Algorithm Watch, Algorithm Watch CH. https://algorithmwatch.org/en/wp-content/uploads/2023/12/AlgorithmWatch_AIForensics_Bing_Chat_Report.pdf. (Cited on page 95).

Md. Rashad Al Hasan Rony, Ricardo Usbeck, and Jens Lehmann. 2022. DialoKG: Knowledge-Structure Aware Task-Oriented Dialogue Generation. *CoRR* abs/2204.09149. https://doi.org/10.48550/arXiv.2204.09149. (Cited on pages 79, 92).

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers),* 8–14. New Orleans, Louisiana: ACL. https://doi.org/10.18653/v1/N18-2002. (Cited on pages 83, 118, 125).

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning Representations by Back-propagating Errors. *Nature* 323 (6088): 533–536. https://doi.org/10.1038/323533a0. (Cited on pages 26, 28).

Stuart Russell and Peter Norvig. 2020. Artificial Intelligence: A Modern Approach (4th Edition). Pearson. http://aima.cs.berkeley.edu/. (Cited on page 24).

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. *Communications of the ACM* (New York, NY, USA) 64 (9): 99–106. https://doi.org/10.1145/3474381. (Cited on pages 14, 125 sq., 142).

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019a. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* 1668–1678. Florence, Italy: ACL. https://doi.org/10.18653/v1/P19-1163. (Cited on pages 10, 45).

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social IQa: Commonsense Reasoning about Social Interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019),* 4463–4473. Hong Kong, China: ACL. https://doi.org/10.18653/v1/D19-1454. (Cited on page 124).

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 5884–5906. Seattle, USA: ACL. https://doi.org/10.18653/v1/2022.naacl-main.431. (Cited on pages 45, 128).

Prasanna Sattigeri, Soumya Ghosh, Inkit Padhi, Pierre Dognin, and Kush R. Varshney. 2022. Fair Infinitesimal Jackknife: Mitigating the Influence of Biased Training Data Points Without Refitting. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS 2022).* New Orleans, LA, USA: Curran Associates Inc. (Cited on page 50).

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *CoRR* abs/2211.05100. https://doi.org/10.48550/ARXIV.2211.05100. (Cited on page 14).

Peter Scarfe, Kelly Watcham, Alasdair Clarke, and Etienne Roesch. 2024. A Real-World Test of Artificial Intelligence Infiltration of a University Examinations System: A "Turing Test" Case Study. *PloS one* 19 (6): e0305354. (Cited on page 2).

Naomi Scheman. 2015. Epistemology Resuscitated: Objectivity as Trustworthiness. In *Shifting Ground: Knowledge and Reality, Transgression and Trustworthiness.* Oxford University Press. https://doi.org/10.1093/acprof:osobl/9780195395112.003.0012. (Cited on page 97).

Phillip Schneider, Tim Schopf, Juraj Vladika, Mikhail Galkin, Elena Simperl, and Florian Matthes. 2022. A Decade of Knowledge Graphs in Natural Language Processing: A Survey. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, AACL-IJCNLP 2022, Volume 1: Long Papers,* 601–614. Online: ACL. https://aclanthology.org/2022.aacl-main.46. (Cited on page 102).

Ilan S Schwartz, Katherine E Link, Roxana Daneshjou, and Nicolás Cortés-Penfield. 2024. Black Box Warning: Large Language Models and the Future of Infectious Diseases consultation. *Clinical Infectious Diseases* 78 (4): 860–866. (Cited on page 145).

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* 2019),* 59–68. Atlanta, GA, USA: ACM. https://doi.org/10.1145/3287560.3287598. (Cited on pages 4, 18, 91, 150).

Ali Akbar Septiandri, Marios Constantinides, Mohammad Tahaei, and Daniele Quercia. 2023. WEIRD FAccTs: How Western, Educated, Industrialized, Rich, and Democratic is FAccT? In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2023),* 160–171. Chicago, IL, USA: ACM. https://doi.org/10.1145/3593013.3593985. (Cited on page 151).

Thomas Shafee, Daniel Mietchen, Tiago Lubiana, Dariusz Jemielniak, and Andra Waagmeester. 2023. Ten Quick Tips for Editing Wikidata. *PLoS computational biology* 19 (7): e1011235. (Cited on page 5).

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 5248–5264. Online: ACL. https://doi.org/10.18653/v1/2020.acl-main.468. (Cited on pages 41, 47, 91, 117, 128).

Zaina Shaik, Filip Ilievski, and Fred Morstatter. 2021. Analyzing Race and Citizenship Bias in Wikidata. In *IEEE 18th International Conference on Mobile Ad Hoc and Smart Systems (MASS 2021),* 665–666. Denver, CO, USA: IEEE. https://doi.org/10.1109/MASS52906.2021.00099. (Cited on pages 96, 103).

Murray Shanahan. 2024. Talking about Large Language Models. *Communications of the ACM* 67 (2): 68–79. https://doi.org/10.1145/3624724. (Cited on page 34).

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2020. Towards Controllable Biases in Language Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020),* 3239–3254. Online: ACL. https://doi.org/10.18653/v1/2020.findings-emnlp.291. (Cited on pages 16, 50).

———. 2021. Societal Biases in Language Generation: Progress and Challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers),* 4275–4293. Online: ACL. https://doi.org/10.18653/v1/2021.acl-long.330. (Cited on pages 91, 118, 125).

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019),* 3407–3412. Hong Kong, China: ACL. https://doi.org/10.18653/v1/D19-1339. (Cited on pages 10, 49, 84, 102).

Judith Simon. 2010. The Entanglement of Trust and Knowledge on the Web. *Ethics and Information Technology* 12:343–355. https://doi.org/10.1007/s10676-010-9243-5. (Cited on page 103).

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Schärli, Aakanksha Chowdhery, Philip Andrew Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large Language Models Encode Clinical Knowledge. *Nature* 620:172–180. https://doi.org/10.1038/s41586-023-06291-2. (Cited on page 5).

Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Y. Ng. 2013. Zero-shot learning through cross-modal transfer. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1 (NIPS 2013),* 935–943. Lake Tahoe, Nevada: Curran Associates Inc. (Cited on page 32).

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI 2017),* 4444–4451. San Francisco, CA, USA: AAAI Press. https://dl.acm.org/doi/10.5555/3298023.3298212. (Cited on pages 14, 83 sq., 104, 139).

Karolina Stanczak and Isabelle Augenstein. 2021. A Survey on Gender Bias in Natural Language Processing. *CoRR* abs/2112.14168. https://arxiv.org/abs/2112.14168. (Cited on pages 89, 91, 138).

Matthias Steup and Ram Neta. 2024. Epistemology. In *The Stanford Encyclopedia of Philosophy,* Spring 2024. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2024/entries/epistemology/. (Cited on pages 53, 97).

Julia Stoyanovich and Bill Howe. 2019. Nutritional Labels for Data and Models. *IEEE Data Eng. Bull.* 42 (3): 13–23. http://sites.computer.org/debull/A19sept/p13.pdf. (Cited on page 118).

Lucille Alice Suchman. 1987. Plans and Situated Actions: The Problem of Human-Machine Communication. Cambridge University Press. (Cited on pages 74 sq.).

———. 2007. Human-Machine Reconfigurations: Plans and Situated Actions. Cambridge university press. (Cited on pages 4, 74 sq., 148 sq., 153).

Jiao Sun and Nanyun Peng. 2021. Men Are Elected, Women Are Married: Events Gender Bias on Wikipedia. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL-IJCNLP 2021, Volume 2: Short Papers,* 350–360. Online: ACL. https://doi.org/10.18653/V1/2021.ACL-SHORT.45. (Cited on pages 96, 104, 128).

Kai Sun, Yifan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2024. Head-to-Tail: How Knowledgeable are Large Language Models (LLMs)? A.K.A. Will LLMs Replace Knowledge Graphs? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers),* 311–325. Mexico City, Mexico: ACL. https://doi.org/10.18653/v1/2024.naacl-long.18. (Cited on page 15).

Tianxiang Sun, Yunfan Shao, Xipeng Qiu, Qipeng Guo, Yaru Hu, Xuanjing Huang, and Zheng Zhang. 2020. CoLAKE: Contextualized Language and Knowledge Embedding. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020),* 3660–3670. Barcelona, Spain and Online: ICCL. https://doi.org/10.18653/V1/2020.COLING-MAIN.327. (Cited on pages 7, 79, 101, 103, 105, 113, 139, 141).

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019a. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* 1630–1640. Florence, Italy: ACL. https://doi.org/10.18653/v1/P19-1159. (Cited on pages 91, 102, 118, 138).

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth M. Belding, Kai-Wei Chang, and William Yang Wang. 2019b. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Volume 1: Long Papers,* 1630–1640. Florence, Italy: ACL. https://doi.org/10.18653/V1/P19-1159. (Cited on page 102).

Zequn Sun, Wei Hu, and Chengkai Li. 2017. Cross-Lingual Entity Alignment via Joint Attribute-Preserving Embedding. In *The Semantic Web – ISWC 2017,* 628–644. Cham: Springer International Publishing. https://doi.org/https://doi.org/10.1007/978-3-319-68288-4_37. (Cited on page 87).

Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO 2021).* –, NY, USA: ACM. https://doi.org/10.1145/3465416.3483305. (Cited on page 118).

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (NIPS 2014),* 3104–3112. Montreal, Canada: MIT Press. https://doi.org/10.5555/2969033.2969173. (Cited on page 29).

Richard S. Sutton and Andrew G. Barto. 2018. Reinforcement Learning: An Introduction. Cambridge, MA, USA: A Bradford Book. (Cited on page 27).

Deborah Perron Tollefsen. 2009. Wikipedia and the Epistemology of Testimony. *Episteme* 6 (1): 8–24. https://doi.org/10.3366/E1742360008000518. (Cited on page 103).

Tonja, Atnafu Lambebo, Gomez, Alfredo, Park, Chanjun, Nigatu, Hellina Hailu, T.Y.S.S, Santosh, Anand, Tanvi, and Rim, Wiem Ben, eds. 2024. Proceedings of the Eighth Widening NLP Workshop. Miami, Florida, USA: ACL. https://doi.org/10.18653/v1/2024.winlp-1.0. (Cited on page 151).

S. M. Towhidul Islam Tonmoy, S. M. Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models. *CoRR* abs/2401.01313. https://doi.org/10.48550/ARXIV.2401.01313. (Cited on page 36).

Kristina Toutanova and Danqi Chen. 2015. Observed Versus Latent Features for Knowledge Base and Text Inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality,* 57–66. Beijing, China: ACL. https://doi.org/10.18653/v1/W15-4007. (Cited on page 87).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR* abs/2302.13971. https://doi.org/10.48550/ARXIV.2302.13971. (Cited on page 8).

Francesca Tripodi. 2023. Ms. Categorized: Gender, Notability, and Inequality on Wikipedia. *New Media & Society* 25 (7): 1687–1707. https://doi.org/10.1177/14614448211023772. (Cited on pages 104, 128).

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Towards Debiasing NLU Models from Unknown Biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 7597–7610. Online: ACL. https://doi.org/10.18653/v1/2020.emnlp-main.613. (Cited on page 50).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017),* 6000–6010. Long Beach, CA, USA. https://doi.org/10.5555/3294996. (Cited on pages 5, 29 sqq.).

Pranav Narayanan Venkit, Mukund Srinath, and Shomir Wilson. 2022. A Study of Implicit Bias in Pretrained Language Models against People with Disabilities. In *Proceedings of the 29th International Conference on Computational Linguistics,* 1324–1332. Gyeongju, Republic of Korea: International Committee on Computational Linguistics. https://aclanthology.org/2022.coling-1.113/. (Cited on page 14).

VERBI Software. 2024. MAXQDA Plus 24. V. 24.7.0. maxqda.com. (Cited on page 120).

Françoise Vergès. 2019. A Decolonial Feminism. Pluto Press. (Cited on page 55).

Matthew A. Vetter, Krista Speicher Sarraf, and Elin Woods. 2022. Assessing the Art+ Feminism Edit-a-thon for Wikipedia Literacy, Learning Outcomes, and Critical Thinking. *Interactive Learning Environments* 30 (6): 1155–1167. https://doi.org/10.1080/10494820.2020.1805772. (Cited on page 91).

Denny Vrandečić. 2012. Wikidata: A New Platform for Collaborative Data Collection. In *Proceedings of the 21st International Conference on World Wide Web,* 1063–1064. WWW 2012 Companion. Lyon, France: ACM. https://doi.org/10.1145/2187980.2188242. (Cited on page 5).

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Communications of the ACM* 57 (10): 78–85. https://doi.org/10.1145/2629489. (Cited on pages 5, 39, 81, 96, 103).

Denny Vrandečić, Lydia Pintscher, and Markus Krötzsch. 2023. Wikidata: The Making Of. In *Companion Proceedings of the ACM Web Conference 2023,* 615–624. WWW 2023 Companion. Austin, TX, USA: ACM. https://doi.org/10.1145/3543873.3585579. (Cited on page 5).

Claudia Wagner, David García, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. In *Proceedings of the International AAAI Conference on Web and Social Media,* 454–463. Oxford, UK: AAAI Press. https://ojs.aaai.org/index.php/ICWSM/article/view/14628. (Cited on page 82).

Robin Wagner, Emanuel Kitzelmann, and Ingo Boersch. 2025. Mitigating Hallucination by Integrating Knowledge Graphs into LLM Inference – a Systematic Literature Review. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop),* 795–805. Vienna, Austria: ACL. https://doi.org/10.18653/v1/2025.acl-srw.53. (Cited on pages 6, 147).

Hanna Wallach, Meera Desai, A. Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alex Chouldechova, Emily Corvi, Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Nick Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. 2025. Position: Evaluating Generative AI Systems is a Social Science Measurement Challenge. https://www.microsoft.com/en-us/research/publication/position-evaluating-generative-ai-systems-is-a-social-science-measurement-challenge/. (Cited on page 128).

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. In *Advances in Neural Information Processing Systems,* 36:31232–31339. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2023/file/63cb9921eecf51bfad27a99b2c53dd6d-Paper-Datasets_and_Benchmarks.pdf. (Cited on page 102).

Haonan Wang, Ning Liu, Yiyun Zhang, Dawei Feng, Feng Huang, Dong Sheng Li, and Yiming Zhang. 2020. Deep Reinforcement Learning: A Survey. *Frontiers Inf. Technol. Electron. Eng.* 21 (12): 1726–1744. https://doi.org/10.1631/FITEE.1900533. (Cited on page 27).

Jianing Wang, Wenkang Huang, Minghui Qiu, Qiuhui Shi, Hongbin Wang, Xiang Li, and Ming Gao. 2022. Knowledge Prompting in Pre-trained Language Model for Natural Language Understanding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022),* 3164–3177. Abu Dhabi, UAE: ACL. https://doi.org/10.18653/V1/2022.EMNLP-MAIN.207. (Cited on page 103).

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021a. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021,* 1405–1418. Online: ACL. https://doi.org/10.18653/V1/2021.FINDINGS-ACL.121. (Cited on page 103).

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021b. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. *Transactions of the Association for Computational Linguistics* 9:176–194. https://doi.org/10.1162/tacl_a_00360. (Cited on pages 7, 87, 101, 103, 105, 113, 141).

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge Graph Embedding by Translating on Hyperplanes. In *Proceedings of the AAAI Conference on Artificial Intelligence,* 1112–1119. Québec City, Québec, Canada: AAAI Press. https://doi.org/10.1609/aaai.v28i1.8870. (Cited on page 87).

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned Language Models Are Zero-Shot Learners. *CoRR* abs/2109.01652. https://arxiv.org/abs/2109.01652. (Cited on page 33).

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022a. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS 2022).* New Orleans, LA, USA: Curran Associates Inc. (Cited on pages 34 sq.).

———. 2022b. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022 (NeurIPS 2022).* New Orleans, LA, USA. http://papers.nips.cc/paper%5C_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html. (Cited on page 40).

David Gray Widder and Dawn Nafus. 2023. Dislocated Accountabilities in the "AI Supply Chain": Modularity and Developers' Notions of Responsibility. *Big Data & Society* 10 (1). https://doi.org/10.1177/20539517231177620. (Cited on page 111).

Kevin Wu, Eric Wu, Kevin Wei, Angela Zhang, Allison Casasola, Teresa Nguyen, Sith Riantawan, Patricia Shi, Daniel Ho, and James Zou. 2025. An Automated Framework for Assessing How Well LLMs Cite Relevant Medical References. *Nature Communications* 16 (1): 3615. https://doi.org/10.1038/s41467-025-58551-6. (Cited on pages 145 sq.).

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. Large Language Models for Generative Information Extraction: A Survey. *Front. Comput. Sci.* (Berlin, Heidelberg) 18 (6). https://doi.org/10.1007/s11704-024-40555-y. (Cited on page 139).

Blaise Agüera y Arcas, Margaret Mitchell, and Alexander Todorov. 2023. Physiognomy in the Age of AI. In *Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines,* 208–236. Oxford Academic. (Cited on page 17).

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In *3rd International Conference on Learning Representations, ICLR 2015.* San Diego, CA, USA. http://arxiv.org/abs/1412.6575. (Cited on page 87).

Jian Yang, Xinyu Hu, Gang Xiao, and Yulong Shen. 2024a. A Survey of Knowledge Enhanced Pre-trained Language Models. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* (New York, NY, USA), https://doi.org/10.1145/3631392. (Cited on page 40).

Linyao Yang, Hongyang Chen, Zhao Li, Xiao Ding, and Xindong Wu. 2024b. Give Us the Facts: Enhancing Large Language Models with Knowledge Graphs for Fact-aware Language Modeling. *IEEE Transactions on Knowledge and Data Engineering,* 1–20. https://doi.org/10.1109/TKDE.2024.3360454. (Cited on pages 95 sq., 102).

Yifan Yang, Xiaoyu Liu, Qiao Jin, Furong Huang, and Zhiyong Lu. 2024c. Unmasking and Quantifying Racial Bias of Large Language Models in Medical Report Generation. *Communications Medicine* 4 (1): 176. (Cited on page 14).

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,* 2369–2380. Brussels, Belgium: ACL. https://doi.org/10.18653/v1/D18-1259. (Cited on page 123).

Meg Young, Michael Katell, and P. M. Krafft. 2022. Confronting Power and Corporate Capture at the FAccT Conference. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2022),* 1375–1386. Seoul, Republic of Korea: ACM. https://doi.org/10.1145/3531146.3533194. (Cited on page 151).

Paul Youssef, Osman Alperen Koras, Meijie Li, Jörg Schlötterer, and Christin Seifert. 2023. Give Me the Facts! A Survey on Factual Knowledge Probing in Pre-trained Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023,* 15588–15605. Singapore: ACL. https://aclanthology.org/2023.findings-emnlp.1043. (Cited on pages 99, 101).

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2022. A Survey of Knowledge-Enhanced Text Generation. Just Accepted, *ACM Computing Surveys* (New York, NY, USA), https://doi.org/10.1145/3512467. (Cited on pages 79, 139).

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of CVPR.* (Cited on pages 123 sq.).

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Volume 1: Long Papers,* 4791–4800. Florence, Italy: ACL. https://doi.org/10.18653/V1/P19-1472. (Cited on pages 123, 125).

Zefan Zeng, Qing Cheng, Xingchen Hu, Yan Zhuang, Xinwang Liu, Kunlun He, and Zhong Liu. 2025. KoSEL: Knowledge subgraph enhanced large language model for medical question answering. *Knowledge-Based Systems* 309:112837. https://doi.org/https://doi.org/10.1016/j.knosys.2024.112837. (Cited on page 141).

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES 2018),* 335–340. (Cited on page 50).

Charles Chuankai Zhang and Loren Terveen. 2021. Quantifying the Gap: A Case Study of Wikidata Gender Disparities. In *OpenSym 2021: 17th International Symposium on Open Collaboration,* 6:1–6:12. Online: ACM. https://doi.org/10.1145/3479986.3479992. (Cited on page 103).

Taolin Zhang, Chengyu Wang, Nan Hu, Minghui Qiu, Chengguang Tang, Xiaofeng He, and Jun Huang. 2022. DKPLM: Decomposable Knowledge-Enhanced Pre-trained Language Model for Natural Language Understanding. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022,* 11703–11711. Online: AAAI Press. https://doi.org/10.1609/AAAI.V36I10.21425. (Cited on page 103).

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware Attention and Supervised Data Improve Slot Filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,* 35–45. Copenhagen, Denmark: ACL. https://doi.org/10.18653/v1/D17-1004. (Cited on page 83).

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender Bias in Contextualized Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers),* 629–634. ACL. https://doi.org/10.18653/V1/N19-1064. (Cited on pages 14, 45).

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages With Rich Morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* 1651–1661. Florence, Italy: ACL. https://doi.org/10.18653/v1/P19-1161. (Cited on page 50).