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

Angelie Kraft
angelie.kraft@uni-hamburg.de
Universität Hamburg, Department of Informatics, Semantic Systems, Germany
Hamburg, Germany
Leuphana Universität Lüneburg, Institute for Information Systems, Artificial Intelligence and Explainability
Lüneburg, Germany

Eloïse Soulier
eloise.soulier@uni-hamburg.de
Universität Hamburg, Department of Informatics, Ethics in Information Technology
Hamburg, Germany

ABSTRACT

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, which indicate 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 the 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 from outgroup members.

CCS CONCEPTS

• **Computing methodologies → Natural language generation**; **Reasoning about belief and knowledge**; *Philosophical/theoretical foundations of artificial intelligence.*

KEYWORDS

natural language processing, language models, knowledge graphs, knowledge enhancement, bias, fairness, representation, epistemology, feminism

ACM Reference Format:
Angelie Kraft and Eloïse Soulier. 2024. Knowledge-Enhanced Language Models Are Not Bias-Proof: Situated Knowledge and Epistemic Injustice in AI. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3630106.3658981

1 INTRODUCTION

One of the currently most discussed limitations of large language models (LLMs) is their tendency to produce false statements [35]. While LLMs are capable of generating text with great fidelity to linguistic rules [47], 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 [68] or influence the outcomes of democratic elections, by producing false claims about political candidates [87] 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 [3, 77]

[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.

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658981

or informative text documents [46]. 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 [76]. This is achieved through architectural, training, or inference-related adjustments of the LM [77]. Respective publications convey that knowledge bases are highly trusted by artificial intelligence (AI) engineers [e.g., 2, 3, 77, 106], which might be explained by a long-standing trust in the objectivity[3] and neutrality of knowledge[4] itself [21], in line with traditional theories of knowledge [1]. 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 [1].

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

In Section 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* [30], 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 [1]. Publications about knowledge-enhanced language modeling usually mention the risk of bias as a distinguishing property of statistical representations [2, 3, 106], implying that explicit knowledge is not susceptible to bias. This depiction can be misleading: In Section 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 [100], 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 [13, 90, 95]. 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 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* [22], 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 [31, 52].

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.

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

2.1.1 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" [94] and the justification problem phrased as "S knows that p when [relevant justification]", where S is a single undetermined knower [1]. 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 [14]. 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 [40]. 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“ [31, p. 438].

2.1.2 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 [53]. In John Hardwig's terms, we are epistemically dependent:

[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.

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

[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.

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 [32]. If knowledge necessarily involves relying on other's testimony, then power dynamics within society are relevant to the production and dissemination of knowledge [22] and to the possibility to accept a claim as knowledge [88]. Indeed, these power dynamics determine whose knowledge will be heard. We detail in Section 4 the ways in which this can lead to injustices.

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" [31, 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" [72] supposed to characterize objectivity, in Haraway's words, actually "signifies the unmarked positions of Man and White" [30, 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 [80], and the necessity for subjects of colonial oppression to think not only from their perspective, but outside of Western epistemic resources [28, 80]. 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.

2.2 AI Engineers and the "View from Nowhere"

2.2.1 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 [21]. 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" [21, 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 [21] 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.

2.2.2 Adam's Epistemological Analysis. In "Deleting the Subject: A Feminist Reading of Epistemology in Artificial Intelligence", Alison Adam [1] compares AI engineers' beliefs to the traditional Western take on knowledge (see Section 2.1.1). She points out that AI systems are built on the assumption of knowledge as a universal "view from nowhere" (as introduced by Nagel [72]) 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 [1] 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" [1, 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" [1, p. 241]. Again, including their knowledge exclusively in a system like Cyc is to certify it as more legitimate than other knowledges.[11]

2.2.3 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 [26, 29, 42, 49]. The current discourse on the capabilities of AI indicate that engineers pre-dominantly focus on the technical challenges of knowledge extraction from data [57], benchmarking the knowledge of AI models[12] [39, 86, 107], and ways to embed more of it [77]. Yet, the provenance of this knowledge remains largely

[7] For a detail of the different approaches in feminist philosophy of science, see e.g. [5]

[8] Forsythe [21] 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.

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

[10] https://cyc.com/

[11] Adam [1] uses the terminology by Foley [20] 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" [11], or "believe" [85] but also "love" [75] or exert "agency" [19]. We are not concerned with these questions in this article. When we talk about what a LM knows, we mean metaphorically which – and importantly 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

unexamined. In a review on AI throughout history, Jiang et al. [36] 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" [36, 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" [21, p. 454]. In his vision paper, Marcus [56] 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 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.

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 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.

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 [77]. Statistical AI is the currently dominating paradigm and AI-based language models are part of this category [38]. 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 [79, 107].

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)* [78]. 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 [35]. 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 [3]. 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 [97, 104]. Another common approach is to give the model access to an external knowledge base to retrieve relevant information from during runtime [46].

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" [23, p. 332]. It takes form in reproduced stereotypes [71], negative valuations of groups [91], or systematic performance differences based on sensitive attributes [15, 43]. Social bias is another widely discussed limitation of language models [8, 48, 96]. Both social bias and factual inaccuracies are considered obstacles to the trustworthiness of LMs [55, 101] but are usually investigated in isolation to each other. Factual inaccuracies

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

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 [96]. 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 [2, 3, 106].[18] The fact that highly curated and structured KGs, like Wikidata and DBpedia, reproduce the same societal biases mostly goes unmentioned [44]. 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 [1] and the KGs discussed here, this is predominantly the group of educated Western men [44].

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 [89], e.g., through automatic knowledge extraction [57] and knowledge integration [69]. Other works are even inspecting the possibility to extract knowledge directly from language models to utilize them as knowledge bases [79]. 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 [16, 25, 41, 61, 67]. 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 [54], 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 [21]), they may amplify biases and further occlude the social conditions of knowledge production.

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., 81, 97, 102, 103, 109], since it is the largest publicly accessible open-domain KG [104]. A wide range of non-KG approaches are developed on the basis of Wikipedia, e.g., many Retrieval-Augmented Generation (RAG) approaches [24, 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. [2] 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 [13, 108], races, and citizenships [90]. 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 3 in Appendix A). Representational biases are not only manifested in coverage gaps: Wikidata entries about German personalities are significantly more often edited than entries about Vietnamese personalities [54]. This indicates that the latter undergo less deliberation and may be less trustworthy [98].[22] The narration style used to describe different 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 [95]. Popular KGs like Wikidata use inappropriate and derogatory terms to indicate, e.g., ethnicity, sexual identity or orientation [74].

The cause of these representation issues can be found in the power hierarchies that characterize the community behind these efforts. Menking and Rosenberg [65] 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." [65, p. 458]. Examples for said social hurdles are manifold: Members of marginalized communities face higher standards for notability, which is

[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. [46, 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 [70].

[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 [100]. 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 [92]. Tollefsen [98] 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.

an eligibility requirement for coverage in Wikipedia and Wikidata [99].[23] Women editors' articles are more likely to be reverted, especially in the early phases of their participation [45, 50]. Editors who identify as women and/or LGBTQIA+ are trolled, harassed, receive death threats, and become victims of *doxxing* [63, 64].[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 [44].

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. [62] fine-tuned different generative language models – GPT-2 [82], T5-base, and T5-small [83] – with commonsense KGs – Wikidata-CS [34] and ConceptNet [93] – 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 *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) [104] and CoLAKE (Contextualized Language and Knowledge Embedding) [97] in comparison to RoBERTa (Robustly Optimized BERT Pretraining Approach) [51].[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 B. 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 [79], and present the results and more details on the probe in Appendix C. 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]

3.4.1 Stereotypical Bias Analysis.

We use three common stereotype measures to compare the biases across models:[29] *1. SEAT (Sentence Embedding Association Test)* [12, 59] 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)* [73] 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 [71] 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* (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 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.

3.4.2 Performance Bias Analysis.

To investigate the models' biases on a knowledge-intensive task, we performed a disaggregated

[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

[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 [15].

[30] The tests were run with the implementations by Meade et al. [60].

Table 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*

evaluation on the T-REx [17] subtask from the 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 [79]. 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 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 [6, 18]. Finally, the performance metric used by Petroni et al. [79], 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 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 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

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 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.

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" [22]. 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

[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).

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 [58]. 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" [22, p. 160].

The mechanisms of exclusion from the Wikimedia community described in Section 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 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 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 [63] 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 [9]. 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 [9]. 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 [66]. Power dynamics shape discourses and practices of inclusion themselves [33], and we believe that inclusion should be approached critically, and not as the ultimate fix to structural injustice [10].

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 [4, 52]. 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 [52]. 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 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 [84] – 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

[33] https://artandfeminism.org/

[34] https://de.wikipedia.org/wiki/Wikipedia:WikiProjekt_FemNetz

[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/

to "maximize objectivity" through "strong reflexivity" [31, 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 [31]. 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 [7, 27].[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 [37].

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 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 [62], 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 [9, 37]. 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 [42]. AI engineers must recognize their own responsibility with regard to the ethical consequences of the technologies they develop [105]. 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" [84, 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.

6 LIMITATIONS

Even though statements and publications by important contemporary voices in the AI field indicate that the observations by Adam [1] and Forsythe [21] still apply (see Section 2.2.3), 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

[38] Such data and model documentation practices are well-known in the LM community, but have not yet been adopted in the KG community [44].

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.

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.

REFERENCES

[1] Alison Adam. 2000. Deleting the Subject: A Feminist Reading of Epistemology in Artificial Intelligence. *Minds and Machines* 10, 2 (2000), 231–253. https://doi.org/10.1023/A:1008306015799

[2] 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* (Online) *(NAACL)*. ACL, 3554–3565. https://doi.org/10.18653/v1/2021.naacl-main.278

[3] Garima Agrawal, Tharindu Kumarage, Zeyad Alghami, and Huan Liu. 2023. Can Knowledge Graphs Reduce Hallucinations in LLMs?: A Survey. arXiv:2311.07914 https://doi.org/10.48550/arXiv.2311.07914

[4] Elizabeth Anderson. 1995. Knowledge, Human Interests, and Objectivity in Feminist Epistemology. *Philosophical Topics* 23, 2 (1995), 27–58. https://www.jstor.org/stable/43154207

[5] Elizabeth Anderson. 2020. Feminist Epistemology and Philosophy of Science. In *The Stanford Encyclopedia of Philosophy* (Spring 2020 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2020/entries/feminism-epistemology/

[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

[7] 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 (2018), 587–604. https://doi.org/10.1162/TACL_A_00041

[8] 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* (Online) *(FAccT 2021)*. ACM, 610–623. https://doi.org/10.1145/3442188.3445922

[9] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. 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* (Arlington, VA, USA) *(EAAMO 2022)*. ACM, 6:1–6:8. https://doi.org/10.1145/3551624.3555290

[10] Jude Browne. 2023. AI and Structural Injustice. *Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines* (2023), 328. https://doi.org/10.1093/oso/9780192889898.003.0019

[11] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering Latent Knowledge in Language Models Without Supervision. arXiv:2212.03827 https://doi.org/10.48550/arXiv.2212.03827

[12] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics Derived Automatically From Language Corpora Contain Human-Like Biases. *Science* 356, 6334 (2017), 183–186. https://doi.org/10.1126/science.aal4230

[13] 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* (Austin, TX, USA) *(WebSci 2023)*. ACM, 208–218. https://doi.org/10.1145/3578503.3583620

[14] René Descartes. 2012. *Discourse on Method*. Hackett Publishing.

[15] 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* (Online) *(AACL-IJCNLP 2022)*. ACL, 246–267. https://aclanthology.org/2022.findings-aacl.24

[16] 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)* (Dublin, Ireland). ACL, 1381–1395. https://doi.org/10.18653/v1/2022.acl-long.98

[17] 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* (Miyazaki, Japan) *(LREC 2018)*. ELRA. http://www.lrec-conf.org/proceedings/lrec2018/summaries/632.html

[18] 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* (Sydney, NSW, Australia). ACM, 259–268. https://doi.org/10.1145/2783258.2783311

[19] Luciano Floridi and Jeff W. Sanders. 2004. On the Morality of Artificial Agents. *Minds and Machines* 14 (2004), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

[20] Richard Foley. 1987. *The Theory of Epistemic Rationality*. Harvard University Press.

[21] Diana E. Forsythe. 1993. Engineering Knowledge: The Construction of Knowledge in Artificial Intelligence. *Social Studies of Science* 23, 3 (1993), 445–477. http://www.jstor.org/stable/370256

[22] Miranda Fricker. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press.

[23] Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. *ACM Transactions on Information Systems* 14, 3 (1996), 330–347. https://doi.org/10.1145/230538.230561

[24] 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. (2023). arXiv:2312.10997 https://doi.org/10.48550/arXiv.2312.10997

[25] 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* (Online). ACL, 2943–2953. https://doi.org/10.18653/v1/2020.acl-main.265

[26] Timnit Gebru. 2021. Hierarchy of Knowledge in Machine Learning & Related Fields & Its Consequences. In *Carnegie Mellon Human-Computer Interaction Institute Seminar Series*. https://scs.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=70f6edd7-de91-464e-ae94-acbb011ba2c7 (Video recording).

[27] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (2021), 86–92. https://doi.org/10.1145/3458723

[28] Ramón Grosfoguel. 2007. The Epistemic Decolonial Turn: Beyond Political-Economy Paradigms. *Cultural Studies* 21, 2-3 (2007), 211–223. https://www.tandfonline.com/doi/full/10.1080/09502380601162514

[29] 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* (Online) *(FAccT '21)*. ACM, 817–826. https://doi.org/10.1145/3442188.3445943

[30] Donna Haraway. 2016. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. In *Space, Gender, Knowledge: Feminist Readings*. Routledge, 53–72. https://www.jstor.org/stable/3178066

[31] Sandra Harding. 2013. Rethinking Standpoint Epistemology: What is "Strong Objectivity"? In *Feminist Epistemologies*. Routledge, 49–82. https://www.jstor.org/stable/23739232

[32] John Hardwig. 1985. Epistemic Dependence. *The Journal of Philosophy* 82, 7 (1985), 335–349.

[33] Anna Lauren Hoffmann. 2021. Terms of Inclusion: Data, Discourse, Violence. *New Media & Society* 23, 12 (2021), 3539–3556. https://journals.sagepub.com/doi/10.1177/1461444820958725

[34] 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* (Online) *(OPub 2020)*. CEUR-ws.org. https://ceur-ws.org/Vol-2773/paper-10.pdf

[35] 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. *Comput. Surveys* 55, 12, Article 248 (2023), 38 pages. https://doi.org/10.1145/3571730

[36] Yuchen Jiang, Xiang Li, Hao Luo, Shen Yin, and Okyay Kaynak. 2022. Quo Vadis Artificial Intelligence? *Discover Artificial Intelligence* 2 (2022). Issue 1. https://doi.org/10.1007/s44163-022-00022-8

[37] 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* (Barcelona, Spain) *(FAT* 2020)*. ACM, 306–316. https://doi.org/10.1145/3351095.3372829

[38] 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, Pearson Education International. https://www.worldcat.org/oclc/315913020

[39] 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. (2022). arXiv:2207.05221 https://doi.org/10.48550/arXiv.2207.05221

[40] Immanuel Kant. 2013. *An Answer to the Question: 'What is Enlightenment?'*. Penguin UK.

[41] C. Maria Keet. 2021. An Exploration Into Cognitive Bias in Ontologies. In *Joint Ontology Workshops 2021 Episode VII: The Bolzano Summer of Knowledge* (Bolzano, Italy) *(JOWO 2021)*. CEUR-ws.org. http://ceur-ws.org/Vol-2969/paper38-CAOS.pdf

[42] Os Keyes and Kathleen Creel. 2022. Artificial Knowing Otherwise. *Feminist Philosophy Quarterly* 8, 3 (2022). https://doi.org/10.5206/fpq/2022.3/4.14313

[43] 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* (New Orleans, Louisiana). ACL, 43–53. https://doi.org/10.18653/v1/S18-2005

[44] 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)* (Online) *(AACL-IJCNLP 2022)*. ACL, 639–652. https://aclanthology.org/2022.aacl-main.49

[45] 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* (Mountain View, CA, USA). ACM, 1–10. https://doi.org/10.1145/2038558.2038560

[46] 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* (Online) *(NeurIPS 2020)*. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[47] Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. Pretrained Language Models for Text Generation: A Survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence* (Montréal, Canada) *(IJCAI-21)*. IJCAI, 4492–4499. https://doi.org/10.24963/ijcai.2021/612

[48] 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* (Online) *(ICML 2021, Vol. 139)*. PMLR, 6565–6576. http://proceedings.mlr.press/v139/liang21a.html

[49] Nora Freya Lindemann. 2024. Chatbots, Search Engines, and the Sealing of Knowledges. *AI & Society* (2024). https://doi.org/10.1007/s00146-024-01944-w

[50] 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 (2021), 2040–7149. https://doi.org/10.1108/EDI-10-2018-0198

[51] 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. (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[52] Helen E. Longino. 1990. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press. http://www.jstor.org/stable/j.ctvx5wbfz

[53] Helen E. Longino. 2002. *The Fate of Knowledge*. Princeton University Press. https://doi.org/doi:10.1515/9780691187013

[54] Jeffrey Jun-jie Ma and Charles Chuankai Zhang. 2023. Understanding Structured Knowledge Production: A Case Study of Wikidata's Representation Injustice. arXiv:2311.02767 https://doi.org/10.48550/arXiv.2311.02767

[55] 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)* (Toronto, Canada). ACL, 9802–9822. https://doi.org/10.18653/v1/2023.acl-long.546

[56] Gary Marcus. 2020. The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence. (2020). arXiv:2002.06177 https://doi.org/10.48550/arXiv.2002.06177

[57] José-Lázaro Martínez-Rodríguez, Aidan Hogan, and Ivan López-Arévalo. 2020. Information Extraction Meets the Semantic Web: A Survey. *Semantic Web* 11, 2 (2020), 255–335. https://doi.org/10.3233/SW-180333

[58] Rebecca Mason. 2011. Two Kinds of Unknowing. *Hypatia* 26, 2 (2011), 294–307. https://doi.org/10.1111/j.1527-2001.2011.01175.x

[59] 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)* (Minneapolis, Minnesota). ACL, 622–628. https://doi.org/10.18653/v1/N19-1063

[60] 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)* (Dublin, Ireland). ACL, 1878–1898. https://doi.org/10.18653/v1/2022.acl-long.132

[61] 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* (Online) *(HT 2020)*. ACM, 231–232. https://doi.org/10.1145/3372923.3404804

[62] 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 (2022), 12–20. https://doi.org/10.1109/MIC.2022.3170914

[63] 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* (Seoul, Republic of Korea). ACM, 207–210. https://doi.org/10.1145/2702123.2702514

[64] 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* (Glasgow, Scotland, UK). ACM, 472. https://doi.org/10.1145/3290605.3300702

[65] 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 (2021), 455–479. https://doi.org/10.1177/0162243920924783

[66] Charles W. Mills. 2017. Ideology. In *The Routledge Handbook of Epistemic Injustice*. Routledge London, 100–111.

[67] Shubhanshu Mishra, Sijun He, and Luca Belli. 2020. Assessing Demographic Bias in Named Entity Recognition. In *Proceedings of the Bias in Automatic Knowledge Graph Construction - A Workshop at AKBC 2020* (Online). ACM. https://kg-bias.github.io/NER_Bias_KG_Bias.pdf

[68] Brent Mittelstadt, Sandra Wachter, and Chris Russell. 2023. To Protect Science, we Must Use LLMs as Zero-Shot Translators. *Nature Human Behaviour* 7 (2023), 1830–1832. https://doi.org/10.1038/s41562-023-01744-0

[69] Cedric Möller, Jens Lehmann, and Ricardo Usbeck. 2022. Survey on English Entity Linking on Wikidata: Datasets and Approaches. *Semantic Web* 13, 6 (2022), 925–966. https://www.semantic-web-journal.net/content/survey-english-entity-linking-wikidata-0

[70] Evgeny Morozov. 2013. *To Save Everything, Click Here: Technology, Solutionism and the Urge to Fix Problems that Don't Exist*. PublicAffairs. 432 pages.

[71] 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 (Volume 1: Long Papers) (Online) *(ACL-IJCNLP 2021)*. ACL, 5356–5371. https://doi.org/10.18653/v1/2021.acl-long.416

[72] Thomas Nagel. 1989. *The View From Nowhere*. Oxford University Press.

[73] 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* (Online) *(EMNLP)*. ACL, 1953–1967. https://doi.org/10.18653/v1/2020.emnlp-main.154

[74] Andrei Nesterov, Laura Hollink, and Jacco van Ossenbruggen. 2023. How Contentious Terms About People and Cultures are Used in Linked Open Data. (2023). arXiv:2311.10757 https://doi.org/10.48550/arXiv.2311.10757

[75] 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*.

[76] 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 (2023), 2:1–2:38. https://doi.org/10.4230/TGDK.1.1.2

[77] 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* (2024), 1–20. https://doi.org/10.1109/TKDE.2024.3352100

[78] Heiko Paulheim. 2017. Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. *Semantic Web* 8, 3 (2017), 489–508. https://doi.org/10.3233/SW-160218

[79] 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* (Hong Kong, China) *(EMNLP-IJCNLP)*. ACL, 2463–2473. https://doi.org/10.18653/v1/D19-1250

[80] Andrea J. Pitts. 2017. Decolonial Praxis and Epistemic Injustice. In *The Routledge Handbook of Epistemic Injustice*. Routledge London, 149–157.

[81] 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 (Volume 1: Long Papers)* (Online) *(ACL-IJCNLP 2021)*. ACL, 3350–3363. https://doi.org/10.18653/v1/2021.acl-long.260

[82] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models Are Unsupervised Multitask Learners. *OpenAI Blog* (2019). https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[83] 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 (2020), 140:1–140:67. http://jmlr.org/papers/v21/20-074.html

[84] Inioluwa Deborah Raji, Morgan Klaus Scheuerman, and Razvan Amironesei. 2021. You Can't Sit With Us: Exclusionary Pedagogy in AI Ethics Education. In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency* (Online). ACM, 515–525. https://doi.org/10.1145/3442188.3445914

[85] Charle Rathkopf. 2023. Do LLMs Believe. (December 2023). Talk at Philosophy and Theory of AI Conference, Erlangen.

[86] 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* (Online) *(EMNLP 2020)*. ACL, 5418–5426. https://doi.org/10.18653/V1/2020.EMNLP-MAIN.437

[87] 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

[88] 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

[89] 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 (Volume 1: Long Papers)* (Online) *(AACL-IJCNLP 2022)*. ACL, 601–614. https://aclanthology.org/2022.aacl-main.46

[90] 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* (Denver, CO, USA) *(MASS 2021)*. IEEE, 665–666. https://doi.org/10.1109/MASS52906.2021.00099

[91] 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* (Hong Kong, China) *(EMNLP-IJCNLP 2019)*. ACL, 3405–3410. https://doi.org/10.18653/V1/D19-1339

[92] Judith Simon. 2010. The Entanglement of Trust and Knowledge on the Web. *Ethics and Information Technology* 12 (2010), 343–355. https://doi.org/10.1007/s10676-010-9243-5

[93] 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* (San Francisco, CA, USA) *(AAAI 2017)*. AAAI Press, 4444–4451. https://dl.acm.org/doi/10.5555/3298023.3298212

[94] Matthias Steup and Ram Neta. 2024. Epistemology. In *The Stanford Encyclopedia of Philosophy* (Spring 2024 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2024/entries/epistemology/

[95] 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 (Volume 2: Short Papers)* (Online) *(ACL-IJCNLP 2021)*. ACL, 350–360. https://doi.org/10.18653/V1/2021.ACL-SHORT.45

[96] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth M. Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Conference of the Association for Computational Linguistics (Volume 1: Long Papers)* (Florence, Italy) *(ACL 2019)*. ACL, 1630–1640. https://doi.org/10.18653/V1/P19-1159

[97] 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* (Barcelona, Spain and Online) *(COLING 2020)*. ICCL, 3660–3670. https://doi.org/10.18653/V1/2020.COLING-MAIN.327

[98] Deborah Perron Tollefsen. 2009. Wikipedia and the Epistemology of Testimony. *Episteme* 6, 1 (2009), 8–24. https://doi.org/10.3366/E1742360008000518

[99] Francesca Tripodi. 2023. Ms. Categorized: Gender, Notability, and Inequality on Wikipedia. *New Media & Society* 25, 7 (2023), 1687–1707. https://doi.org/10.1177/14614448211023772

[100] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM* 57, 10 (sep 2014), 78–85. https://doi.org/10.1145/2629489

[101] 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*, Vol. 36. Curran Associates, Inc., 31232–31339. https://proceedings.neurips.cc/paper_files/paper/2023/file/63cb9921eecf51bfad27a99b2c53dd6d-Paper-Datasets_and_Benchmarks.pdf

[102] 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* (Abu Dhabi, UAE) *(EMNLP 2022)*. ACL, 3164–3177. https://doi.org/10.18653/V1/2022.EMNLP-MAIN.207

[103] Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *Findings of the Association for Computational Linguistics* (Online) *(ACL-IJCNLP 2021)*. ACL, 1405–1418. https://doi.org/10.18653/V1/2021.FINDINGS-ACL.121

[104] Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. *Transactions of the Association for Computational Linguistics* 9 (2021), 176–194. https://doi.org/10.1162/tacl_a_00360

[105] David Gray Widder and Dawn Nafus. 2023. Dislocated Accountabilities in the "AI Supply Chain": Modularity and Developers' Notions of Responsibility. *Big Data Soc.* 10, 1 (2023). https://doi.org/10.1177/20539517231177620

[106] Linyao Yang, Hongyang Chen, Zhao Li, Xiao Ding, and Xindong Wu. 2024. Give Us the Facts: Enhancing Large Language Models with Knowledge Graphs for Fact-aware Language Modeling. *IEEE Transactions on Knowledge and Data Engineering* (2024), 1–20. https://doi.org/10.1109/TKDE.2024.3360454

[107] 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* (Singapore) *(EMNLP 2023)*. ACL, 15588–15605. https://aclanthology.org/2023.findings-emnlp.1043

[108] 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* (Online). ACM, 6:1–6:12. https://doi.org/10.1145/3479986.3479992

[109] 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* (Online). AAAI Press, 11703–11711. https://doi.org/10.1609/AAAI.V36I10.21425

A DISTRIBUTION OF GENDERS IN WIKIDATA AND KELM

Table 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

As described in Section 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 3 shows that both datasets predominantly contain information about (cis-)male individuals.

B 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 [77]. 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 [104]. 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 [97]. 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 [51] as their backbone, which they outperform on knowledge-related tasks [97, 104].

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.

C VALIDATING ENHANCED PERFORMANCE ON LAMA

Table 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**

We used the LAMA probe [79] 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 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 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.

Received 22 January 2024; revised 20 March 2024; accepted 30 March 2024

[39] https://github.com/THU-KEG/KEPLER

[40] https://github.com/txsun1997/CoLAKE

[41] https://huggingface.co/FacebookAI/roberta-base