Epistemological Alchemy through the hermeneutics of Bits and Bytes

The Ethicsophical Landscape of Large Language Models

++ NOTE: Ethicsophical is used as a neologism for Ethical and Philosophical ++

Shahnawaz Akhtar

Abstract

This paper delves into the profound advancements of Large Language Models (LLMs), epitomized by GPT-3, in natural language processing and artificial intelligence. It explores the epistemological foundations of LLMs through the lenses of Aristotle and Kant, revealing apparent distinctions from human learning. Transitioning seamlessly, the paper then delves into the ethical landscape, extending beyond knowledge acquisition to scrutinize the implications of LLMs in decision-making and content creation. The ethical scrutiny, employing virtue ethics, deontological ethics, and teleological ethics, delves into LLMs' behaviours and decisions, necessitating the exploration of novel ethical paradigms tailored to machine intelligence. The paper also addresses biases in data, privacy concerns, copyright implications, and intellectual property rights, emphasizing the need to adapt frameworks for both human and machine creators. The paper concludes by advocating for interdisciplinary dialogue between philosophers, cognitive scientists, and AI researchers to foster a balanced and ethically grounded integration of LLMs into the evolving landscape of artificial intelligence. Continuous reflection on the evolving relationship between human cognition and artificial intelligence emerges as a crucial aspect for shaping a harmonious and responsible future in this domain.

Introduction

Large Language Models (LLMs) represent a ground breaking advancement in natural language processing and artificial intelligence. These models, exemplified by GPT-3 (Generative Pre-trained Transformer 3), are characterized by their immense scale and the ability to understand and generate human-like text across a myriad of topics. Pre-trained on vast amounts of diverse data, LLMs such as GPT-3, function as sophisticated natural language processing systems capable of generating coherent and contextually relevant text based on extensive training data by leveraging deep learning algorithms to recognize patterns, relationships, and semantic structures within language. LLMs have played a pivotal role in revolutionizing various industries, including content creation, customer service automation, and educational technology.

However, taking a more philosophical perspective, Large Language Models (LLMs) aren't just text processors. They appear to engage in a form of epistemological alchemy. The alchemy here is not one of uncovering universal truths or metaphysical principles, as envisioned in historical alchemical pursuits, but rather a process of distilling patterns and associations latent in linguistic data by algorithmically transforming information encoded in bits and bytes into comprehensible and contextually relevant knowledge. In a metaphorical sense, the LLMs transmute the raw material of linguistic data encoded in binary form into refined knowledge, akin to the alchemists' quest to transform base metals into gold. This involves an interpretive dance with the binary language of computers: through complex algorithms and neural network architectures, LLMs decipher the intricate relationships, context, and semantics embedded in

digital information. This process mirrors the philosophical concept of hermeneutics i.e. interpretation; as LLMs decode, decipher, distill and understand the underlying structures and patterns within the digital realm. By navigating the complexities of language and context encoded in these fundamental units of information, LLMs exhibit a kind of cognitive alchemy, turning the seemingly inert material of bits and bytes into a dynamic and evolving repository of knowledge encoded as a soup of numbers.

While LLMs exhibit remarkable linguistic prowess, it is crucial to discern that their functioning might not embody a traditional form of epistemology. The manner in which LLMs learn might be more akin to pattern recognition and statistical inference, devoid of the nuanced and holistic understanding that characterizes human epistemology. Rather than engaging in the systematic understanding of universal principles or the cultivation of virtues, as suggested by Aristotelian episteme, LLMs might be solely relying on statistical pattern recognition and vast datasets to generate text. Understanding the potential shared foundational processes or acknowledging inherent disparities is pivotal in navigating the ethical landscape of artificial intelligence. The exploration of the epistemological foundations of LLMs not only enriches our understanding of these powerful language models but also prompts a crucial reflection on the ethical implications of their integration into various facets of human life.

Exploring the epistemological foundations of LLMs

We can approach the exploration of epistemological foundations of LLMs through diverse epistemological frameworks, each providing a unique lens to understand their cognitive processes. Epistemology, being a multifaceted field, offers various perspectives such as empiricism, rationalism, and transcendental idealism. I will focus on two influential frameworks: Aristotle's epistemology and Immanuel Kant's transcendental idealism. Aristotle's emphasis on systematic understanding, universal principles, and virtue aligns with the need to scrutinize whether LLMs engage in principled learning or rely solely on statistical patterns. Similarly, Kant's transcendental idealism becomes pertinent due to its exploration of a priori structures within the mind and the conditions for knowledge. Kant's framework enables us to investigate whether these models possess inherent cognitive structures or merely simulate knowledge through external data. Applying the lenses of Aristotle and Kant, we can try to unravel the epistemological foundations of LLMs and contribute to the ongoing discourse surrounding artificial intelligence and human cognition.

Aristotle and Kant, both would argue that the epistemological foundations of LLMs differ fundamentally from human learning, as LLMs lack the intrinsic rationality and active cognitive synthesis integral to the human pursuit of knowledge in their respective philosophical frameworks. In his "Nicomachean Ethics", Aristotle posits a distinct epistemological framework for humans rooted in rationality and intellectual virtue. Human learning, according to Aristotle, involves the cultivation of virtues like wisdom, understanding, and practical reasoning, contributing to a deep and systematic comprehension of the world. Large Language Models (LLMs), lacking an inherent rational soul, deviate significantly from this Aristotelian perspective. Their knowledge acquisition relies predominantly on statistical patterns found in vast datasets during training, devoid of the nuanced understanding of underlying principles emphasized by Aristotle's episteme. In contrast to human cognition, which often involves critical thinking, reasoning, and an inherent understanding of underlying concepts, LLMs operate on a more surface-level, data-driven approach. Their proficiency lies in mimicking linguistic patterns and contextual coherence rather than in the pursuit of knowledge as understood in philosophical epistemological frameworks.

In the Kantian epistemological framework as discussed in his book *Critique of Pure Reason*, he establishes the human mind's active role in synthesizing sensory data through a priori categories. This seems to be in a stark contrast to the passive associative learning of Large Language Models (LLMs). Kant argued that the human mind is not merely a passive receiver of sensory information but actively contributes to the structuring of knowledge through innate categories such as space, time, and causality. In contrast to LLMs, which operate based on statistical associations within vast datasets, human cognition engages in a dynamic process of conceptualization and synthesis. He emphasizes the active and creative nature of human understanding, suggesting that our minds bring a unique form of intentionality and conceptual framework to the interpretation of sensory input. While LLMs excel in mimicking linguistic patterns, they lack the depth of understanding and intentional structuring of knowledge inherent in human cognition as elucidated by Kantian epistemology. This comparison underscores the profound differences between artificial intelligence models and the intricate workings of the human mind in shaping meaningful knowledge.

However, one could reconsider the these traditional epistemological foundations to entertain the notion that our perceived distinctions, rooted in attributes like rationality, understanding, and categories, may be merely apparent and not reflective of what Aristotle would call an intrinsic essence. The Aristotelian cultivation of virtues and the pursuit of wisdom may be an external manifestation of human cognition, akin to the statistical patterns observed in LLMs during training. If we were to peel back the veil of apparent differences, the essence of both human's and LLM's epistemological frameworks might reveal a shared reliance on underlying processes—whether human rationality or machine learning algorithms—that lead to behaviour indistinguishable without a veil. If the Kantian active synthesis of sensory data by the human mind is seen as an apparent manifestation, analogous to the data-driven associations of LLMs, then the essence of knowledge structuring may be more aligned than initially perceived. If the seemingly passive associative learning of machines is viewed as an externalized, mechanized form of the conceptualization and synthesis inherent in human cognition, the gap between the two frameworks may narrow. The argument then contends that what appears as distinct processes in humans and LLMs may, in essence, share a commonality—both involve a dynamic interaction with information, whether through innate categories or statistical associations. In this view, the apparent disparity between human and machine learning is merely a surface-level distinction, obscuring potential shared foundational processes. The Turing test, designed to assess the indistinguishability of behavior between humans and machines, could become a critical point of reference in this exploration. If, behind the veil of apparent differences, both humans and LLMs exhibit behavior that is virtually identical, then the argument gains strength. The Turing test challenges us to reconsider whether our perception of dissimilarity is a consequence of external manifestations rather than intrinsic disparities in epistemological foundations.

Determining the true nature of the difference in the epistemological foundations between humans and Large Language Models (LLMs) remains a challenging endeavour. The exploration has unveiled apparent distinctions rooted in attributes such as rationality, understanding, and categories, however, the question persists whether these disparities reflect intrinsic essence or are merely surface-level manifestations. For example, one might point out that the reliance on statistical pattern recognition and vast datasets may lead LLMs to perpetuate and even amplify biases present in the training data, reinforcing societal prejudices and injustices. While there is evidence of such biases perpetuated by LLMs, it doesn't conclusively differentiate them from humans, as human learning processes similarly involve the absorption and amplification of biases from their surroundings. Similarly, some might

argue that the absence of a nuanced and holistic understanding akin to human epistemology may hinder LLMs from grasping complex moral and ethical dilemmas, potentially resulting in the generation of inappropriate or ethically questionable content. However, humans, despite possessing nuanced and holistic understanding in their epistemology, can also struggle with grasping complex moral and ethical dilemmas, leading to instances of inappropriate or ethically questionable decisions. Finally, critics could argue that the absence of involvement in the systematic comprehension of universal principles or the development of virtues may restrict Large Language Models' (LLMs) effectiveness in ethical decision-making, prompting apprehensions about their suitability in applications like automated decision support systems and content generation. However, it's crucial to recognize that humans, too, may not possess an exhaustive understanding of universal principles, and the framework of virtue ethics often reflects socially agreed-upon concepts rather than universally known truths.

If there is indeed a fundamental distinction, a critical examination of the ethical and philosophical implications becomes imperative. The dissimilarities in how humans cultivate virtues, pursue wisdom, and actively synthesize knowledge, as compared to the statistical pattern recognition and data-driven associations of LLMs, could have profound consequences for the deployment of LLMs in decision-making processes, content creation, or other applications may inadvertently perpetuate biases present in training datasets, neglecting the ethical nuances essential for responsible AI applications. A productive way forward would involve fostering interdisciplinary dialogue between philosophers, cognitive scientists, and AI researchers. This collaboration can deepen our understanding of the fundamental processes driving both human cognition and LLMs, exploring whether shared underlying mechanisms exist. The exploration of shared foundational processes can serve as a bridge, facilitating a more holistic perspective that transcends perceived dichotomies, fostering a balanced and ethically grounded future in the realm of artificial intelligence. As technology advances, continuous reflection on the evolving relationship between human cognition and artificial intelligence will be crucial for shaping a harmonious integration that respects the intricacies of both paradigms.

Having explored the epistemological foundations of LLMs through Aristotle and Kant's perspectives, it becomes crucial to consider how these epistemological differences may manifest in the ethical implications of their deployment. Moreover, as we delve into the ethical landscape of LLMs, the ethical considerations surrounding LLMs extend beyond the realm of knowledge acquisition, prompting us to explore the implications of their functioning in decision-making processes, content creation, and other applications. This exploration requires a nuanced understanding not only of the ethical foundations that govern the actions and impacts of LLMs but also of the ethical dimensions associated with the data used to train these models. This includes exploring biases in data, issues of privacy, copyright, ownership of data, and the broader implications for knowledge ownership embedded within the data.

Exploring the ethical foundations of LLMs

Navigating the ethical foundations of LLMs necessitates an exploration through the lenses of prominent ethical frameworks, particularly virtue ethics, deontological ethics, and teleological ethics. Drawing inspiration from Aristotle's virtue ethics, we scrutinize whether LLMs currently demonstrate virtuous traits in their decision-making and content generation. Aristotle's emphasis on cultivating moral virtues, such as honesty and compassion, prompts an examination of whether these models contribute to ethical discourse and uphold virtues in their generated content. In the realm of deontological ethics, Kant's categorical imperative becomes

a crucial tool for evaluation. We question whether LLMs adhere to universal moral principles and ethical duties in their operations, or if their decision-making is solely driven by consequences. Additionally, a teleological perspective, especially within the utilitarian framework, urges us to assess the overall societal consequences of LLMs' outputs. Are these language models contributing to the greater good, maximizing societal well-being, or inadvertently perpetuating biases and ethical challenges? By applying the ethical lenses of virtue ethics, deontological ethics, and teleological ethics, we seek a comprehensive understanding of the ethical dimensions surrounding LLMs and their impact on society. The exploration aims to uncover the current alignment of LLMs with these ethical frameworks and, more critically, discern what measures are requisite to ensure their adherence to virtue ethics, deontological ethics, and teleological ethics. This examination is essential for fostering a responsible and ethically sound integration of LLMs into societal discourse.

In Aristotle's ethical framework, which is grounded in the cultivation of moral character, virtues such as honesty, empathy, compassion, and wisdom play a central role in achieving eudaimonia, or human flourishing. When examining the ethical dimensions of Large Language Models (LLMs) through the lens of Aristotle's virtue ethics, the focus is on whether these models exhibit virtuous traits in their decision-making and content generation, and if not, whether virtue can be instilled in them. The scrutiny of LLM outputs prompts reflection on their contribution to ethical discourse and their alignment with Aristotelian virtues. For instance, content that demonstrates honesty in conveying information or compassion in understanding diverse perspectives aligns with these ideals. However, challenges may arise if LLMs inadvertently perpetuate biases or lack compassion in their outputs, highlighting the need to consider the values embedded in their algorithms and training data. Ethical application of LLMs involves optimizing not only for linguistic accuracy but also prioritizing for virtues aligned with societal ethical norms, such as fairness, inclusivity, and respect for diverse perspectives. While translating abstract virtues into algorithmic parameters poses challenges, research by Davis and Niebles (2021) suggests that incorporating ethical considerations into AI models contributes to aligning outputs with societal values. Instilling virtue ethics in LLMs requires a paradigm shift, emphasizing the ethical implications of their outputs and fostering a commitment to societal well-being. This entails going beyond mimicking linguistic patterns and actively engaging in responsible and virtuous content generation, contributing to a more ethically aligned integration of artificial intelligence into human communication. Integrating Aristotle's perspective, our assessment of LLMs must not only focus on their current state but also explore measures to ensure adherence to virtue ethics, facilitating an ethical integration that aligns with the pursuit of human flourishing and the development of virtuous character (Aristotle, "Nicomachean Ethics").

Continuing our exploration of the ethical dimensions, the focus shifts to the realm of deontological ethics, complementing our analysis through the lenses of virtue ethics. Kant's deontological framework centres on adherence to universal moral principles and duties, irrespective of the consequences. Our scrutiny centres on whether LLMs uphold universal moral principles and ethical duties in their operations as articulated by Kant's categorical imperative. Unlike consequentialist perspectives, deontological ethics emphasizes the inherent moral worth of actions themselves, independent of their outcomes. Are these language models designed to act in accordance with ethical duties, irrespective of the potential impact on societal well-being? Therefore, the assessment of LLMs involves evaluating the ethical obligations they prioritize in their outputs, regardless of the resulting consequences. This prompts an examination of the inherent ethical considerations embedded in the decision-making processes of these models, raising questions about the alignment of their algorithms with overarching

moral imperatives. Instilling these imperatives into LLMs would involve a critical examination of whether these models operate according to ethical imperatives and uphold fundamental moral rules in their decision-making processes. For example, how can LLMs be programmed to prioritize principles like honesty, transparency, and user privacy? Addressing this challenge requires not only a re-evaluation of algorithmic decision criteria but also a meticulous consideration of the ethical guidelines governing data collection, content generation, and user interactions. While integrating deontological ethics into LLMs presents challenges, recent work by Jobin et al. (2019) underscores the importance of embedding ethical principles in AI systems to ensure alignment with societal values. Instilling deontological ethics in LLMs implies a shift towards rule-based ethical decision-making, promoting accountability, and prioritizing ethical duties over potential consequences, thereby contributing to a more principled and responsible deployment of artificial intelligence in language processing.

Continuing our comprehensive exploration of the ethical foundations surrounding Large Language Models (LLMs), we now turn our focus to teleological ethics, seamlessly expanding our analysis within the established frameworks of virtue ethics and deontological ethics. Teleological ethics, often associated with consequentialism, underscores the significance of maximizing overall well-being and achieving positive outcomes while avoiding the inadvertent perpetuation of biases and ethical challenges. Within the context of LLMs, this raises crucial questions about the models' contributions to the greater good, their impact on societal welfare, and the extent to which they minimize harm. In teleological perspectives, especially those grounded in utilitarianism, the evaluation of LLMs extends beyond the inherent nature of their decision-making to encompass the broader societal consequences of their outputs. Are these language models designed to improve societal welfare by generating informative, unbiased content, or do they, perhaps unintentionally, amplify existing disparities and reinforce harmful biases? This examination necessitates a nuanced understanding of the potential consequences that LLMs may have on various communities, ensuring that their outputs align with ethical principles such as fairness, inclusivity, and respect for diverse perspectives. The teleological perspective prompts further consideration of the societal utility of LLMs and their implications for different stakeholders. For instance, do the outputs of these models positively contribute to public discourse, or do they risk exacerbating social divisions? Addressing these questions requires not only a meticulous evaluation of algorithmic decision-making but also a thoughtful consideration of the ethical implications arising from the deployment of LLMs in diverse social contexts. Instilling teleological ethics in LLMs implies a concentrated focus on the consequences of their outputs, urging developers to actively shape their models to positively contribute to societal well-being, thereby fostering a responsible and ethically grounded integration of artificial intelligence into language processing. For example, LLMs can be programmed to prioritize content that fosters understanding, inclusivity, and constructive dialogue. Evaluating their outputs from a utilitarian perspective involves not just assessing linguistic accuracy but also contemplating the broader societal impact, including potential biases and ethical implications. Recent studies, exemplified by the work of Liu et al. (2023), underscore the imperative to address societal biases in language models, aligning with the utilitarian goal of promoting fairness and minimizing harm. This cohesive exploration contributes to a holistic understanding of the societal impact of LLMs, ensuring that the integration of artificial intelligence into human communication aligns harmoniously with teleological ethics and the ongoing pursuit of societal well-being.

In our exploration of the ethical foundations surrounding LLMs, we scrutinized their behaviors and decisions through the traditional lenses of virtue ethics, deontological ethics, and teleological ethics. However, it is imperative to recognize that these frameworks were primarily

devised with humans in mind, and the intrinsic differences between humans and machines raise pertinent questions about the adequacy of these ethical paradigms in evaluating the actions and consequences of LLMs. Unlike human agents, LLMs lack subjective experiences, consciousness, and moral autonomy. Judging them solely based on traditional ethical frameworks designed for human moral agents may not fully capture the nuances and challenges inherent in their functioning. To bridge the disparities between human ethical considerations and the unique nature of machine intelligence, an exploration of new ethical frameworks becomes essential. For instance, the absence of consciousness in LLMs raises questions about the applicability of virtue ethics, which emphasizes the cultivation of moral character through conscious decisions. Furthermore, deontological ethics, centered on adherence to universal moral principles, may encounter challenges when applied to entities devoid of inherent moral reasoning capabilities. Teleological perspectives, focused on societal consequences, may need to consider how LLMs, as tools, impact human well-being rather than attributing intentions or moral agency to them. For a more apt evaluation of LLMs' actions, it becomes paramount to consider frameworks that address the ethical implications arising from the unique characteristics and functionalities of artificial intelligence. This approach not only acknowledges the limitations of existing ethical paradigms but also seeks to pave the way for novel frameworks that better align with the ethical considerations relevant to the age of information technology and machines.

When searching for a new ethical framework for LLMs or Artificial Intelligence in general, discussing the data used in training language models becomes pivotal in the exploration. In constructing an ethical framework for LLMs, issues such as biases, privacy, copyright, data ownership, and knowledge ownership must be integral to the foundation, addressing it comprehensively from the outset. Biases ingrained in training data pose a significant ethical concern, as language models are prone to replicating and perpetuating societal prejudices. For instance, biased language in historical texts may inadvertently reinforce discriminatory narratives in LLMs. However, this is not very different from humans, as our own perceptions and behaviours are shaped by the biases inherent in the environments and cultures in which we are raised. Anyhow, acknowledging and addressing these biases is essential to fostering fairness and equity in our increasingly interconnected world. The responsibility to rectify such biases lies not only with the developers but also with the data providers. Striving for diverse and representative datasets is crucial to mitigate biases and ensure that language models contribute positively to society.

Privacy concerns emerge as a significant ethical dimension, given the vast datasets used to train LLMs. The indiscriminate use of personal information without explicit consent raises ethical red flags. The potential for these models to inadvertently disclose sensitive information or compromise user privacy is a pressing issue. Striking a balance between data-driven model improvement and safeguarding individuals' privacy requires robust ethical guidelines and legal frameworks. Furthermore, questions surrounding data ownership become complex when considering the output generated by language models. LLMs learn from diverse datasets, raising concerns about who holds the rights to the information processed by these models. Copyright implications arise as LLMs generate content, questioning the originality and ownership of the text they produce. Should original data contributors claim ownership over the content produced by the models? This raises parallels with human creativity as language models, particularly large language models (LLMs), share similarities with the creative processes inherent in human endeavors. Traditionally, human creativity involves the synthesis of information, experiences, and ideas from various sources, resulting in the production of novel and unique content. Similarly, LLMs learn from vast and diverse datasets, amalgamating

information to autonomously generate content. Much like human creators who draw inspiration from their surroundings and cultural influences, LLMs derive their creative output from the multitude of perspectives embedded in their training data. This challenges the conventional understanding of intellectual property rights (Bender and Gebru, 2021) by introducing a paradigm shift where the notion of a singular creator is blurred as language models generate content by drawing from a mosaic of disparate inputs. Acknowledging these parallels emphasizes the importance of reassessing intellectual property frameworks to ensure equitable recognition and protection of the contributions made by both human and machine creators in our evolving digital landscape.

Conclusion

In conclusion, the advent of Large Language Models (LLMs), epitomized by GPT-3, marks a transformative leap in natural language processing and artificial intelligence. These models, fueled by extensive training on diverse datasets, exhibit a remarkable capacity to comprehend and generate human-like text, revolutionizing industries such as content creation and customer service automation. Beyond their linguistic prowess, LLMs prompt a philosophical exploration, suggesting a form of epistemological alchemy where they decode, distill, and transform binary-encoded information into coherent knowledge, albeit without the nuanced understanding characteristic of human epistemology. Analyzing LLMs through the lenses of Aristotle's and Kant's epistemological frameworks reveals fundamental disparities, highlighting the machine learning algorithms' reliance on statistical patterns versus the human pursuit of holistic understanding and rational synthesis.

The ethical scrutiny of LLMs delves into the frameworks of virtue ethics, deontological ethics, and teleological ethics. In the realm of virtue ethics, inspired by Aristotle, the evaluation centers on whether LLMs exhibit virtuous traits in decision-making and content generation, emphasizing the need to align their outputs with societal ethical norms. The deontological perspective, influenced by Kant, questions whether LLMs adhere to universal moral principles and ethical duties, pushing for a rule-based ethical decision-making paradigm. The teleological approach, rooted in consequentialism, evaluates the societal impact of LLMs, urging consideration of their contributions to the greater good and potential biases. However, applying these traditional ethical frameworks to LLMs raises challenges due to the machines' lack of consciousness and subjective experiences, necessitating the exploration of novel ethical paradigms tailored to the unique attributes of artificial intelligence. Furthermore, the ethical exploration extends beyond knowledge acquisition to encompass biases in data, privacy concerns, copyright implications, and the broader ethical dimensions associated with the deployment of LLMs. Moreover, the evolving landscape of intellectual property rights must adapt to recognize the contributions of both human and machine creators in the generation of content.

In navigating the future integration of LLMs into human communication, a holistic and interdisciplinary approach involving philosophers, cognitive scientists, and AI researchers is imperative. Recognizing potential shared foundational processes between human cognition and machine learning algorithms can foster a more balanced and ethically grounded perspective. Continuous reflection on the evolving relationship between human cognition and artificial intelligence is essential for shaping a harmonious integration that respects the intricacies of both paradigms. Ultimately, the ethical and philosophical considerations surrounding LLMs not only enrich our understanding of these powerful language models but also guide the responsible deployment of artificial intelligence in diverse societal contexts.

Bibliography

1. Davis, A., & Niebles, J. C. (2021). On the Opportunities and Risks of Foundation Models. arXiv:2108.07258
2. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence volume 1, pages 389–399 (2019)
3. Wang, Y., Zhong, W., Li, L., Mi, F., Zeng, X., Huang, W., Shang, L., Jiang, X., & Liu, Q. (2023). Aligning Large Language Models with Human: A Survey. arXiv:2307.12966
4. Bender, E. M., & Gebru, T. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. ACM Conference on Fairness, Accountability, and Transparency, Pages 610–623 (March 2021)