Article

new media & society
2026, Vol. 28(1) 5–31

DOI: 10.1177/14614448241279034
journals.sagepub.com/home/nms


ChatGPT in the public eye: Ethical principles and generative concerns in social media discussions

Maayan Cohen
Tel Aviv University, Israel; University of Oxford, UK

Michael Khavkin
Tel Aviv University, Israel

Danielle Movsowitz Davidow
Tel Aviv University, Israel

Eran Toch
Tel Aviv University, Israel

Abstract

With ChatGPT's rapid adoption, concerns regarding generative artificial intelligence (AI) have shifted from theoretical to practical. Drawing upon the "algorithmic imaginary" framework from critical algorithm studies and the anthropological concept of "ordinary ethics," we analyzed Twitter discourse during ChatGPT's initial deployment, examining 368,359 tweets. Our analysis identified five topics reflecting functional and critical aspects of ChatGPT. We specifically point to two topics with a critical perspective: "Ethics" and "Concerns." The first aligns with scholarly discussions in AI ethics on fairness and transparency, while the second focuses on ChatGPT's generative capabilities. This highlights an emerging trend: While the academic discussion on AI ethics has gained popularity, especially in scrutinizing ChatGPT, the conversation is now expanding to

Corresponding author:
Maayan Cohen, Department of Sociology and Anthropology, Gershon H. Gordon Faculty of Social Sciences, Tel Aviv University, 30 Haim Levanon Street, Tel Aviv 6997801, Israel.
Email: maayan11@tauex.tau.ac.il

more nuanced ethical deliberations. We analyzed the posts' engagement and sentiment over time, demonstrating the AI ethics community's influence in addressing the potential and harms of generative AI systems.

Keywords
AI ethics, algorithm, algorithmic imaginary, ChatGPT, discourse, early adoption, generative AI, ordinary ethics, Twitter, user agency

Introduction

ChatGPT's launch in late 2022 marked a significant leap in artificial intelligence (AI), enabling unprecedented public interaction with generative AI technologies. Within weeks, it emerged as the fastest-growing application in terms of user registration. As of March 2024, ChatGPT has approximately 180.5 million active users and 1.6 billion monthly website visits (Nerdy, 2024). This swift uptake not only showcased the practical applications of generative AI in everyday communication but also ignited pressing ethical conversations and broader discussions, such as its gender bias (Gross, 2023), lack of accountability (Liesenfeld et al., 2023) and the effect on creativity (Epstein et al., 2023; Htet et al., 2024). Concerns arose so deeply that some prominent tech figures advocated for a temporary halt in AI progress, alarmed by its potential existential risks (Future of Life Institute, 2023).

Public opinion is central to shaping the ethical discourse surrounding generative AI, impacting not only product acceptance and commercial advancement but also the allocation of research funding and the development of regulations (Smuha, 2021). Furthermore, public perception and feedback play a pivotal role in the evolution of these algorithms, as they are continually refined based on human interactions and responses. With the introduction and tangible realization of generative AI, marked by ChatGPT's launch, public discourse is swiftly adjusting to concerns that were once theoretical but are now concrete.

The focus on the reception of algorithms, specifically, the development of "algorithmic imaginary"—the "ways of thinking about what algorithms are, what they should be and how they function" (Bucher, 2017: 30) among individuals who interact with them has recently become a fundamental concern within the larger field of critical algorithm studies, for example, Kasirzadeh and Gabriel (2023), Lomborg and Kapsch (2020), Schellewald (2022), and Büchi et al. (2023). The literature suggests that people's perceptions of algorithms significantly influence how these algorithms are "enacted" (Bucher, 2017; Devendorf and Goodman, 2014; Seaver, 2017). This theoretical framework proposes that actors do not interact with pre-existing objects such as algorithms; instead, they bring them into existence in contexts where ethical evaluations, folk theories, and diverse imaginaries play crucial roles and impact how algorithms perform and affect individuals (Bucher, 2017). Therefore, understanding people's ethical critiques of algorithms becomes an essential aspect of studies aiming to examine the interaction between algorithms and people. Our work draws on the theoretical concept of "Ordinary Ethics"

(Lambek, 2010), which builds on the integral role of ethics in society and contends that ethics is a reflective property, expressing group membership and identity. In this article, we propose to study what we term "Ordinary AI Ethics," taking a socio-cultural perspective on ethics. Through this perspective, we studied ethics not as a set of abstract or technical rules but by following how ethical considerations around ChatGPT are publicly articulated by actual people. We analyzed their ethical reflections, in which they draw on existing repertoires, logics, and imaginaries as they interact with this emerging technology in its initial months of public deployment. In this context, our research highlights the importance of integrating the exploration of algorithm reception and imaginaries with the quickly expanding field of AI ethics. Diverging from the predominantly philosophical, legal, and technical perspectives prevalent in existing studies, our work is focused on the socio-cultural dimensions of AI ethics—a viewpoint frequently neglected, as highlighted by Avnoon et al. (2023).

To examine public discourse about ChatGPT during its initial public interaction phase, we analyzed discussions on the major social media platform Twitter (referred to as "X" since July 2023). Our dataset comprises 368,359 tweets gathered from Twitter between February 13 and June 4, 2023. The principal questions guiding our research are: What topics emerge as people discuss ChatGPT during its initial adoption phase? To what extent does ChatGPT provoke ethical critiques among Twitter users engaging in conversations about it? If such critiques emerge, what characteristics define them?

Our analysis began by identifying the discourse topics that surfaced during this period. These topics were subsequently grouped into five distinct clusters. We then further analyzed each cluster to discern its central discussion motif and the prevailing sentiment.

Our analytical approach draws inspiration from the emerging concept of the "Algorithmic Imaginary" (Bucher, 2017) and the theoretical anthropological framework of "Ordinary Ethics" (Lambek, 2010). Our goal is to contribute to the critical examination of algorithms and the study of AI ethics from a socio-cultural perspective.

Background

User experiences with algorithms

In recent years, scholarly attention has been drawn toward the social and cultural implications of computer algorithms in various aspects of people's lives. Under the "critical algorithm studies" umbrella, researchers have analyzed algorithms' political dimensions, focusing on transparency, privacy, accountability, and fairness (Gillespie, 2014). A central insight from this body of work emphasizes the importance of understanding users' interactions with algorithms. This insight highlights the complex nature of algorithms and their users, recognizing the users' active role in experiencing and sometimes deliberately resisting algorithmic functions. Supporting this notion, Bucher (2018) argues that the influence of algorithms is not just a byproduct of their functionalities but also stems from the imaginaries people weave around them.

While the study of technology adoption is important across all tech domains, it assumes even greater significance in the realm of machine learning algorithms. This significance arises because algorithms are not just static entities molded by their

creators; they also depend on their users for adaptation and continued relevance. As an illustrative point, while specifics about ChatGPT's mechanism remain proprietary, OpenAI has indicated that ChatGPT employs reinforcement learning with human feedback (RLHF) (Ziegler et al., 2019) for refining its responses. This example highlights how user feedback becomes instrumental in determining the trajectory of algorithmic functions.

A growing number of empirical studies have examined the direct experience of users with algorithms. A key concept in this research is "The Algorithmic Imaginary," coined by Bucher (2017). Bucher describes this concept by analyzing tweets and conducting follow-up interviews with their authors. On a related subject, Lomborg and Kapsch (2020) interviewed media users to examine their understanding of algorithms, the knowledge acquisition process, and their reactions to algorithmic outcomes. In professional contexts where algorithms are employed, Christin (2017) builds upon Bucher's (2017) concept of the "Algorithmic Imaginary" to explore how web journalists and legal professionals utilize and interpret algorithms in their work. She argues that algorithms serve as symbolic resources that mediate the negotiation and enactment of professional values by those who employ them. The critiques and resistance toward algorithms in each profession are closely tied to each field's unique "Algorithmic Imaginaries." Similarly, Kotliar (2020) conducted a study involving workers in data analytics companies who relied on algorithms to assist them in their tasks. He demonstrated how these individuals leverage their personal values and social context to attribute symbolic meanings to algorithmic outputs, effectively transforming "algorithmically produced clusters" into distinctive "identity categories."

Other works have analyzed user approaches when interacting with machine learning algorithms, in settings such as recommendations in streaming services (Siles et al., 2020) or news selection on social media platforms (Fletcher and Nielsen, 2019). Fletcher and Nielsen's (2019) study underscored a prevailing "generalized skepticism" among users regarding news selection, suggesting a critical stance toward algorithmic functions they might not fully comprehend. Ruckenstein and Granroth (2020) analyzed users' reactions to personalization and targeted advertisements on social media. Swart (2021) showed that users cultivate an awareness of algorithms rooted in their emotional experiences, daily social media engagement, and media exposure to data and privacy controversies. These interactions lead to the formation of "folk theories" about algorithmic functions, which help users make sense of the complex way algorithms operate.

The existing literature points to user discomfort and concern in situations where it is evident that algorithms retain records of individuals' past behaviors, showcasing user perspectives on how algorithms should ethically operate and the boundaries they should respect (Bucher, 2017). For instance, while a user might be fully aware of Facebook's data collection practices, they might still feel unsettled when unexpectedly confronted with images of their ex-partner on their news feed. Simply put, algorithms that bring up past actions can evoke feelings of being under surveillance, compromising personal autonomy and privacy (Ruckenstein and Granroth, 2020). This sentiment is also evident in the field of targeted advertising. When users encounter advertisements related to personal matters they have not explicitly shared, they might speculate if their devices are eavesdropping (Kennedy et al., 2017). This type of unease suggests that users have

certain expectations of appropriate information flow, creating a sense of discomfort when those imaginaries and reality mismatch (Büchi et al., 2023). Drawing from the existing literature, our objective is to examine users' critiques of AI, as exemplified by the discourse on ChatGPT. We will assess their specific characteristics and extract the ethical reflections and concerns that emerge when users interact with this technology. This research aims to enrich both the understanding of users' daily perceptions of algorithms and the nascent literature on AI ethics.

AI ethics

Khan et al. (2022) conducted a systematic literature review on AI ethics, showing that the most commonly debated principles in the current literature include transparency, privacy, accountability, and fairness. These principles have become major areas of research within various communities studying Fairness, Accountability, and Transparency (FAccT) (Laufer et al., 2022) in machine learning and human–computer interaction (Van Berkel et al., 2023). In addition, critical discussions on AI ethics also look into its impact on the future of labor (Brynjolfsson et al., 2023; Jones, 2021; Nissim and Simon, 2021), education (Lim et al., 2023), and the creative arts (Epstein et al., 2023). It has been argued that the current perception of AI systems has not matched the capabilities of this technology, thereby raising questions regarding how people should regard AI systems—whether as humans, objects, or an entirely distinct entity (Laakasuo et al., 2021).

The bulk of research on AI ethics adopts a normative position, inquiring what is philosophically or legally adequate, or examining the operations of existing algorithms to suggest improvements. This normative stance is also evident in empirical studies on AI ethics as they scrutinize the performance of various algorithms to demonstrate how they conform to or violate certain ethical ideals, such as gender or racial equality (Brown et al., 2021; Niforatos et al., 2020). These studies aim to provide guidelines for creating or modifying AI systems to align more closely with values identified as desirable in the research literature. For instance, Bender et al. (2021) illustrated how large language models (LLMs) are susceptible to adopting and reflecting societal biases due to being trained on extensive and often uncurated Internet datasets. This raises concerns about the reinforcement and amplification of harmful stereotypes, noting that language models trained on large internet datasets frequently replicate the biases present in their training data, which can further entrench societal biases (Abid et al., 2021; Weidinger et al., 2022). For generative AI, Ghosh and Caliskan (2023) argued that ChatGPT reinforces gender stereotypes by associating specific genders with certain jobs (e.g., portraying doctors as male and nurses as female) and linking certain activities to a particular gender.

Accountability and transparency have been introduced as key principles in the operationalization of AI ethics. Accountability holds system designers accountable for their design choices, while transparency encourages system developers to make justifications regarding those choices accessible, particularly in case of an unexpected or even unethical output (Kazim and Koshiyama, 2021). Transparency, on the contrary, is a property of algorithms that allow both users and experts to assess the properties of algorithms. For example, users interacting with a chat-like interface should be informed whether they are

communicating with a real person or a chatbot (Díaz-Rodríguez et al., 2023). These principles should be held by AI system developers but also by people using and accessing AI tools.

To enhance adherence to ethical norms, adjustments to the development of AI systems can be made by legislation, regulations, designs, instruction, or technological advancements. Technical governance can be employed to ensure that companies building AI systems fulfill their responsibilities to preserve these properties, in addition to direct legislation, which invokes legal compliance. Furthermore, non-technical governance has been implemented to guarantee that decision-makers receive ongoing education and training regarding strategies to enhance transparency and inform users about how automated decisions respect human rights (Lukowicz, 2019).

These steps were taken to promote responsible human-centered AI (Lukowicz, 2019), under which the development of AI systems respects human dignity and autonomy, such that humans can make meaningful and self-conscious decisions. To address cases of limited or inappropriate governance and minimize risks caused by ethical issues, ethical-by-design approaches (Brey and Dainow, 2023) have been proposed for ethical AI development. Such approaches incorporate principles, standards, and best-practice guidelines that system developers can use to improve the system's robustness to ethical violations. It often involves collaboration with experts from various fields, including anthropology and philosophy. These definitions have recently been extended to trustworthy AI (HLEG, 2019), an approach that prioritizes safety and transparency for the users who interact with it.

The ordinary ethics of AI

The ethics of algorithms is currently predominantly discussed mostly from philosophical, legal, and technological perspectives (Avnoon et al., 2023). We propose to assess the socio-cultural components of the ethics of algorithms utilizing insights from the anthropology of ethics and the sociology of morality. Strands in both the sociology of morality and the anthropology of ethics propose to locate their object of inquiry in people's everyday or "ordinary" judgments of what is "right" or "wrong." Thus, they explore temporal and social variations in people's everyday understandings of their obligations, values, and worth (Hitlin et al., 2023; Mattingly and Throop, 2018). Specifically, the anthropology of ethics suggests that ethics can be understood in their "ordinary" contexts (Lambek, 2010). Hence, the term "Ordinary Ethics" in the context of our research refers to cases where people address questions regarding the "good" and the "bad" in everyday situations by focusing on their reflective evaluations, as opposed to the more practical normative stance of "ethics" in philosophy, law, and computer science (Laidlaw, 2018). One of this approach's many strengths is that it allows researchers to examine how cultural ideas inform people and analyze their understanding of "the good" in social relations and reflexive engagement. This is done as a descriptive and not a normative research project.

In our context, we employ this framework with the aim of understanding what ethical discourses emerge around AI systems. In this sense, we regard Twitter users' critical engagement regarding ChatGPT as a space where ethical evaluations of this technology

are formed and refined. Importantly, within this socio-cultural perspective on ethics, individuals' ethical evaluations are influenced by their social group affiliations, class, and political affiliations. This viewpoint, which regards ethical conduct and evaluations as non-monolithic, contextual, and diverse, has been a central theme in both the anthropology of ethics and the sociology of morality. Instead of merely focusing on an individual's choice between "the moral" and "the immoral," the sociology of morality, as proposed by scholars such as (Boltanski and Thévenot, 1999), emphasizes the importance of individuals' ability to navigate within and between diverse moral repertoires and to select among various "moral logics." In the anthropology of ethics, this idea is often present in the discernment that scholars make between the "moral" and the "ethical." While "moral" refers to the set of rules and norms that different societies live by, "ethics" designates the more reflective and agentive processes individuals engage in to cultivate their sense of what is right and wrong in various situations (Mattingly and Throop, 2018).

To show how morality is tied to structural constraints and societal discourses, cultural sociologists often illustrate how morality is employed to demarcate social boundaries. This concept is exemplified in Michelle Lamont's study of working men in the United States and France, where she contends that moral judgments perform boundary work, delineating what separates working men from other groups. This moral delineation is achieved through engagement with what she refers to as available "cultural repertoires," which are shaped by structural constraints and societal discourses (Lamont, 2009; Swidler, 1986). This analytical sensitivity, which aims to include both people's reflective freedom as they consider what is right or wrong, and the social context in which they live, has also become important in the anthropology of ethics. Works such as Saba Mahmood's (2005) study on the gendered politics of piety in Egypt demonstrate how relations between ethics, freedom, power, and politics are configured by the diverse genealogical traditions within which such concepts arise, gain legitimacy, are enacted, and/or are contested (Mattingly and Throop, 2018).

Studies focusing on AI ethics from a socio-cultural perspective have recently emerged, with particular attention paid to research on algorithmic production centers (Avnoon et al., 2023). Avnoon et al.'s (2023) work describes how algorithm developers navigate their ethics in the constraints of their available repertoire, the libertarian, capitalist, and technocratic environment of the tech industry. Some works on users' ethical reflections and concerns surrounding AI have also begun to emerge. For example, Ghotbi et al. (2022) surveyed 228 college students in Japan on AI ethics. Their study shows that most students identified unemployment as the primary AI-related concern, while some highlighted the emotional impact of AI as a main concern. Cave et al. (2019) examined a UK survey on AI views and found that the dominant AI narratives caused anxiety. Namely, most respondents felt powerless over AI's direction, blaming corporate or governmental dominance.

"Ordinary Ethics" directly corresponds with "folk theories," a theoretical conceptualization of users' interaction with technology, used to study how people understand algorithms and personalization tools. Folk theories are "intuitive, informal theories that individuals develop to explain the outcomes, effects, or consequences of technological systems" (DeVito et al., 2017). Folk theories can significantly

diverge from expert theories because they emerge as individuals generalize algorithmic functions based on their *own* personal experiences with these algorithms.

Moreover, findings from HCI studies resonate with the insights from critical algorithm studies. For instance, Eslami et al. (2016) conducted a qualitative laboratory study, discovering that nearly half of the participants lacked a comprehensive understanding of algorithms. However, they were able to identify ten folk theories elucidating individual perceptions of automated curation processes. In a parallel vein, Rader and Gray (2015) studied how people interpret their Facebook newsfeeds. They found that users employed various theories to rationalize the newsfeed's behavior. Some perceived it as an uncontrollable natural force, while others showcased advanced informal reverse engineering notions.

Building on the approach that ethics should be studied as they are enacted in ordinary social contexts by individuals striving to discern between right and wrong, our work aims to examine how people ethically reflect on ChatGPT during its initial months of deployment. We thereby seek to reveal the imaginaries and folk theories that people draw upon in their reflections.

Data and methods

Data collection

We gathered English-language tweets using the Twitter Developer interface over a 4-month period, from February 13, 2023, to June 4, 2023. This collection was guided by specific inclusion criteria, anchored on a comprehensive list of relevant keywords. The keywords used to retrieve the tweets include combinations, such as {"chatgp" odds ratio (OR) "gpt4" OR "gpt-4"} *AND* {"great" OR "impressed" OR "truth" OR "pretend" OR "fake" OR "bad" OR "damage" OR "wrong" OR "dangerous" OR "extreme"} (the full list of keywords is provided in Supplement 1). Each tweet in our dataset includes several details: unique identifiers, the time of posting, the content of the tweet, and engagement metrics (views, likes, replies, and retweets). Our analysis is confined to English tweets due to the challenges in accurately assessing and comparing topic extraction and labeling performance across multiple languages, which cultural variations among Twitter users can significantly influence. The total dataset comprised 368,359 tweets containing the specified keywords.

We applied a series of standard preprocessing techniques to the collected tweets. These steps included tokenization, removing special characters, pruning URLs, and expanding hashtags into their constituent words where possible. In addition, we filtered out duplicate tweets and those without any user engagement. To identify the more trending tweets, we utilized Twitter's activity metrics as indicators of their social impact.[1] After preprocessing, the tweet corpus was reduced to a total of 178,416 tweets. Figure 1 summarizes the entire flow of our analysis.

To address potential privacy risks, we followed best practices in handling public individual data for research purposes (Fiesler and Proferes, 2018). Our study was authorized by the institutional ethics committee (ethics approval 6997-1). We ensured that we did not publish easily identifiable information (such as usernames) and made no attempts to collect or infer demographic information (such as location or gender) that could facilitate future re-identification.

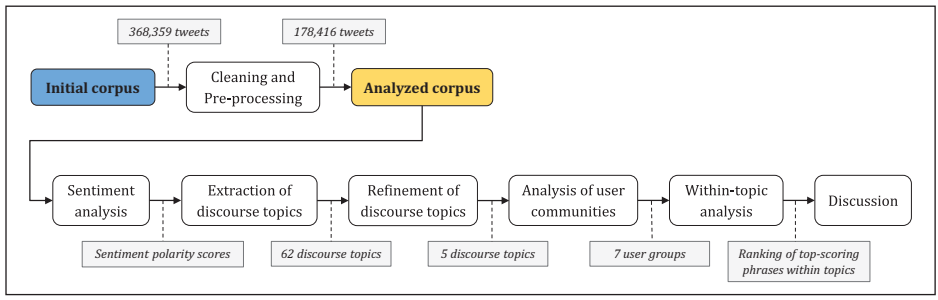

Figure 1. Schematic flow of the employed analysis methodology.

Analyzing topics

To investigate the underlying discourse topics, we used ChatGPT's LLM as a powerful classifier. The use of LLMs, such as ChatGPT, for zero-shot learning (Larochelle et al., 2008) has recently gained traction (Hu et al., 2023; Kojima et al., 2022). This technique uses existing text–class relationships to predict new classes or answer new questions without labeled data, utilizing semantic similarities as prior knowledge. We have used ChatGPT language model capabilities to perform predictive analysis involving natural language processing (Hu et al., 2023; Rathje et al., 2024; Törnberg, 2023), outperforming domain experts in many tasks (Gilardi et al., 2023; Törnberg, 2023). This approach diverges from the traditional Latent Dirichlet Allocation (LDA) topic modeling, often used in ChatGPT and Twitter discourse studies (Adeshola and Adepoju, 2023; Dodd, 2023; Korkmaz et al., 2023; Taecharungroj, 2023), but limited by its need for extensive hyper-parameter tuning.

We employed a three-step procedure for topic modeling: Initially, ChatGPT analyzed ~2000 of the most engaging tweets from our corpus (the tweets that received the highest number of views, likes, replies, and retweets combined), assigning a concise five-word title to each identified topic (see Supplement 5 for the complete list). We then used ChatGPT to cluster these topics into five representative groups, each assigned a title (referred to hereafter as topics). An additional "Other" category was created for tweets that did not fit any single topic, due to the unstructured nature of Twitter text. To manage the stochastic nature of LLMs and ensure response quality, ChatGPT's temperature setting was adjusted to medium. A sensitivity analysis was conducted to assess the effect of different temperature levels on topic assignment consistency. The results showed stable and consistent performance with respect to all topics (see Supplementary Figure S1). Moreover, an insignificant difference was found between the performance of ChatGPT's topic labeling for jargon-based and content-based instructions across varying temperature levels, further supporting the robustness of our prompt instructions when interacting with ChatGPT (see Supplementary Figure S2). Finally, ChatGPT categorized each tweet in our dataset into one of the six topics.

We validated the topic categorization through several methods. Coherence was evaluated using average pairwise word-similarity scores, with all topics scoring high (above 0.67, average 0.71), indicating that the generated topic clusters were consistent, clear,

and relevant to our analysis. Distinctiveness was assessed through similarity-based overlap between phrase distributions across topics, showing a low overlap (0.11). This low overlap remained constant over our analyzed time frame (see Supplementary Figure S1), confirming the validity of our topic clustering. Fleiss' Kappa with weight correction was calculated to measure inter-rater agreement between ChatGPT and the authors, particularly for differentiating critical and general discourse, resulting in a Kappa of $k = 0.76$ ($p < .001$). To statistically examine topic differences over time, we used non-parametric tests due to the non-normal data distribution: A chi-square test for the number of tweets per topic, a Friedman test with Nemenyi post hoc analysis for time series tweet counts, a Mann-Kendall test for trends in discourse within topics, and a Wilcoxon signed-rank test for changes in user engagement over time across topics. These tests were performed at a significance level of $\alpha = .05$.

Analysis of Twitter users

The collected Twitter corpus reflects a broad spectrum of users from multiple communities. We conducted an additional analysis to investigate what central user groups participate in the discussion about ChatGPT across the extracted discourse topics. Namely, we extracted the 50 most engaging tweets from each topic in our corpus (estimated by the number of retweets, likes, quotes, replies, and views) and retrieved the profile information of their authors. Then, we manually reviewed the authors' profiles and categorized them into seven communities. These communities encompassed various backgrounds and professions: technology/AI professionals, creative professionals, academics, journalists, legal professionals, business promoters, and laypeople. To confirm the quality of our labeling, two of the authors labeled the extracted user profiles, reaching a substantial inter-rater agreement (Fleiss' kappa of $k = 0.68$). Subsequently, we utilized the resulting group distribution to derive conclusions regarding the dominant communities echoing the discourse, particularly in critical topics discussing concerns and ethical issues.

Exploring the discourse within topics

To further differentiate the types of discussions surrounding ChatGPT and comprehend their primary attributes, we provided a deeper analysis of the discussions that make up the main topics. To that end, we employed LDA topic modeling (Blei, 2012) to identify the dominant phrases within each topic and thus capture its latent discourse aspects.

In addition, as our work aimed to focus on users' ethical critiques of ChatGPT, we applied sentiment analysis to the tweets to identify tweets with a critical tone. Although this task is highly common when processing static textual corpora, here we performed sentiment analysis temporally to capture the trends in user experience with ChatGPT. For this purpose, we first utilized BERTweet,[2] a state-of-the-art model pre-trained on English tweets, to determine whether the sentiment in each tweet is positive, negative, or neutral. Compared with lexicon and rule-based models commonly used for sentiment analysis (Adeshola and Adepoju, 2023; Dodd, 2023), classifiers such as BERTweet enhance comprehension of underlying sentiments by better capturing the context within sentences.

In addition to extracting the sentiment label, we measured the sentiment polarity of the tweets on a continuous scale, that is, the intensity of the sentiment conveyed by a particular tweet. Then, to enhance our insights about the discourse around ChatGPT, we synthesized the analysis of the classified sentiment with the categorization of the discovered topics. Finally, we manually skimmed the content of the tweets in our corpus to locate representative examples of different discourse directions.

Results

Distribution of discourse topics

Table 1 outlines the five principal topics identified in tweets about ChatGPT during the studied time period. Each topic was assigned a representative title by the ChatGPT model, encapsulating the principal ideas therein. The analysis uncovers a rich tapestry of discussions about ChatGPT, with topics ranging from functionalities and applications across various sectors to evaluations and comparisons with other AI tools.

This study aimed to examine the nature and extent of the critical discourse surrounding ChatGPT during its initial months. Remarkably, 41.84% of the analyzed tweets conveyed what seemed to be a critical perspective, falling under the titles Topic 2: "Concerns" (28.25%) and Topic 5: "Ethics" (13.59%). In addition, tweets under the "Concerns" topic were disproportionately represented, accounting for twice as many tweets as any other topic ($\chi^2[5, N = 178,416] = 19,120.86$, $p < .001$). This was later verified as we examined the prevailing sentiments in each topic, as explained in the next subsection. The remaining topics (Topics 1, 3, and 4) exhibited a more evenly distributed representation in our corpus.

Significantly, the topics of "Ethics" and "Concerns" emerged as the areas where a substantially negative tone was most prevalent. For instance, we discovered tweets within these topics that discussed ChatGPT's offensive language, expressing dissatisfaction with its conversational capabilities, for example, *"Major Problem Ahead: AI giving abusive, mean . . . responses!.*" Consequently, our analysis concentrated on these two topics, rather than the other three topics that did not exhibit critical negativity. Nevertheless, the remaining topics (Topics 1, 3, and 4) were used to contextualize this type of criticism among the richly diverse discourse around ChatGPT.

Temporal changes and sentiment analysis

Upon an examination of the daily count of tweets relating to the topics over time (Figure 2), we observed a statistically significant difference between the topics ($\chi^2[4] = 216.43$, $p < .001$). We discerned that the curves representing the topics "Concerns" (Topic 2) and "Ethics" (Topic 5) exhibited a different trend compared with those of Topics 1, 3, and 4, which remained stationary, starting in late March 2023 (Mann-Kendall test; $\tau = -0.116, \tau = -0.059, and \tau = 0.015$, respectively, $p > .05$). That distinction supported the validity of our observations, proving that the captured differences in the critical topics ("Concerns" and "Ethics") are specific and have the potential to shed more light on the investigated critical discourse. In addition, although the

Table 1. Descriptive statistics of the extracted discourse topics around ChatGPT in our Twitter corpus.

Discourse topic	Count	Pct. from total (%)	Mean daily count	STD daily count	Example of a tweet with affinity to the topic
All	178,416	100	272.80	259.03	
Topic 1: Functionality	21,849	12.24	280.55	151.83	*"In my opinion, they are good for the rewrites . . . I usually give the command 'rewrite this' followed by input . . . it's better than me."*
Topic 2: Concerns	50,391	28.25	530.33	223.23	*"The worrisome aspect of #chatgpt is that it increasingly resembling a human each day . . . Its proficiency in fabricating falsehoods is truly impressive."*
Topic 3: AI in Various Industries	26,191	14.68	326.25	282.98	*'#ChatGPT providing better tax advice. . .than many return preparers out there. . .'*
Topic 4: Comparison and Advancement of AI	23,993	13.45	299.25	303.33	*"ChatGPT could soon be the better way to Google . . ."*
Topic 5: Ethics	24,245	13.59	268.14	154.39	*"When asked about how a female can maximize her femininity in a new chat . . . chatGPT returns a list of advice including wear make up and take care of your physical appearance."*
Other	31,747	17.79	419.70	297.63	"BUY NFT FROM CHATGPT"

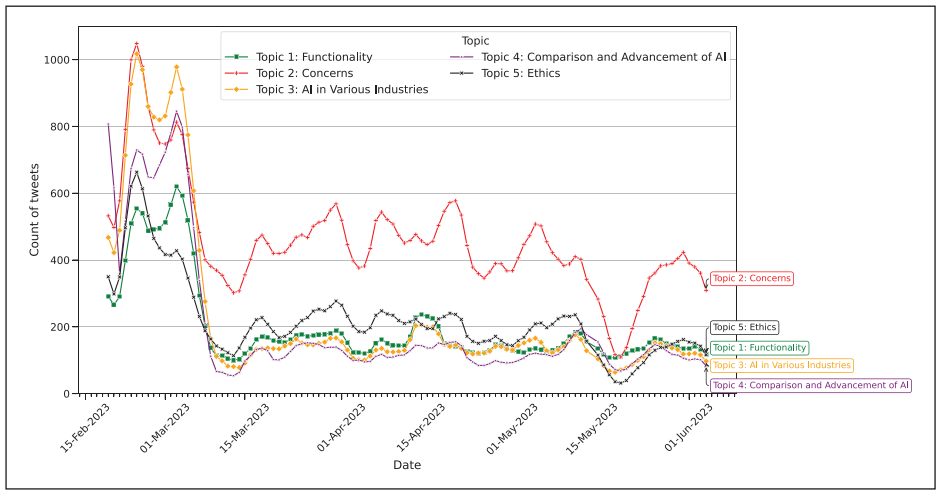

Figure 2. Daily count of tweets per topic over time. Data is smoothed using a 5-day simple moving average.

number of tweets in the "Concerns" topic over time was double that of the "Ethics" topic ($p < .001$), based on the variations in the gradients of the trend curves, these topics exhibited a parallel trajectory (Wilcoxon signed-rank test; $Z = 1838.5$, $p = .890$). This observation suggests that waves of criticism, regardless of similarity or diversity in nature, directed at this emerging technology usually arise and dissipate simultaneously. In contrast, other forms of discussion, which are more positive toward this emerging technology, tend to remain constant over time.

Our study was designed to investigate users "ordinary" ethical reflections on ChatGPT, focusing particularly on tweets expressing criticism toward this technology. To isolate such tweets, we employed sentiment analysis within each topic category, targeting those primarily emanating negative sentiments, which are typically indicative of critical viewpoints.

As depicted in Figure 3, all the topics statistically significantly differed in their sentiment score over time ($\chi^2[4] = 182,385.34$, $p < .001$). In terms of the sentiment conveyed in the tweets, a clear distinction between the topics was observed. Specifically, the "Concerns" and "Ethics" topics were distinguished from the rest of the topics by their negative mean score, in contrast to the consistently positive sentiment polarity observed in the remaining topics over time (Nemenyi pairwise post hoc; $p < .01$ for all pairs). The negative tweets in these topics accounted for 52.28% and 42.10% of the tweets within their respective topics. In contrast, positive tweets constituted a smaller portion, specifically 11.40% and 9.53%, respectively (the remaining percentages correspond to tweets classified as neutral). This suggests that these topics harbor significant critical and ethically charged discussions. A detailed distribution of sentiment labels can be found in Supplementary Table S1.

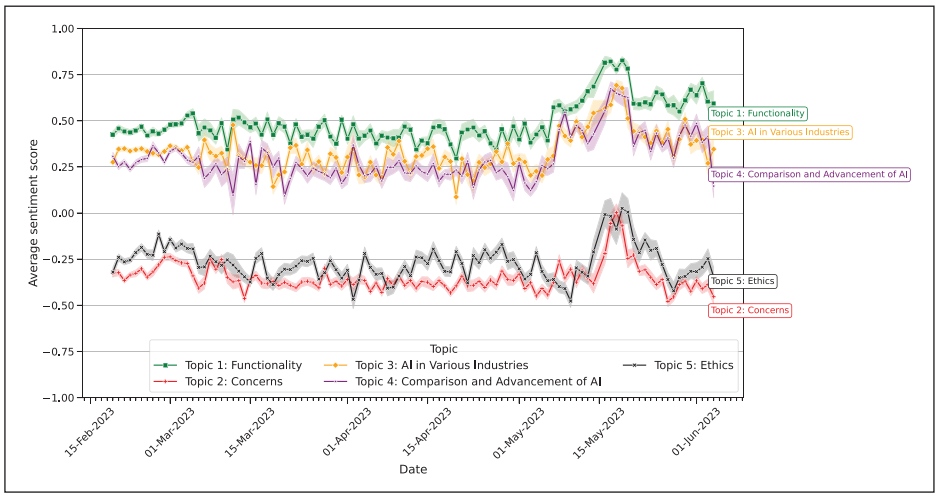

Figure 3. Daily average sentiment score of tweets per topic over time. Sentiment score ranges between −1 and 1, where −1 indicates that a highly negative sentiment is expressed in the text, 0 indicates a neutral sentiment, and +1 indicates a highly positive sentiment. Data is smoothed using a 5-day simple moving average. Shaded error bands around rolling averages indicate a 95% confidence interval.

A noteworthy peak in the average sentiment was observed across all topics between May 15 and 20, 2023, indicating an increase in the positive sentiment score. Moreover, the observed increase facilitated a shift in the tone of critical discussions, including the "Concerns" and "Ethics" topics, moving from a predominantly negative sentiment to a more positive one, on average. After cross-referencing that time frame with the external events that occurred during that time, we saw a potential link between the official launch of Bard, Google's new chatbot and competition to ChatGPT, and the observed increase in positive sentiment. This was backed by a large number of tweets stating that ChatGPT outperforms Bard, in both functionality and performance. For example, a tweet from 14 May 2023 wrote: ". . . *Been trying same prompts on Bard and ChatGPT. Responses . . . from ChatGPT are better as of now . . .*"; another tweet from 18 May 2023 states that ". . . *#GoogleBard needs to learn a lot and come long way to compete with #ChatGPT*."

In alignment with our principal research query—exploring whether this emerging technology induces specific and immediate ethical reservations among Twitter users engaging in related discussions, and identifying the predominant characteristics of such concerns if they do exist—we found it important to further investigate the emergence of two seemingly analogous critical topics, the "Concerns" and "Ethics" topics. These clusters, appearing to be similar in title, sentiment, and evolution over time, necessitated additional scrutiny. Our goal was to discern whether it is justified to regard these clusters as independent entities and, if so, to illuminate the distinguishing factors between them.

We examined the median engagement rate of each topic over the months encompassed by our study (Figure 4) and found statistically significant differences in the median engagement across topics ($\chi^2[4] = 220.49$, $p < .001$). Notably, we observed that

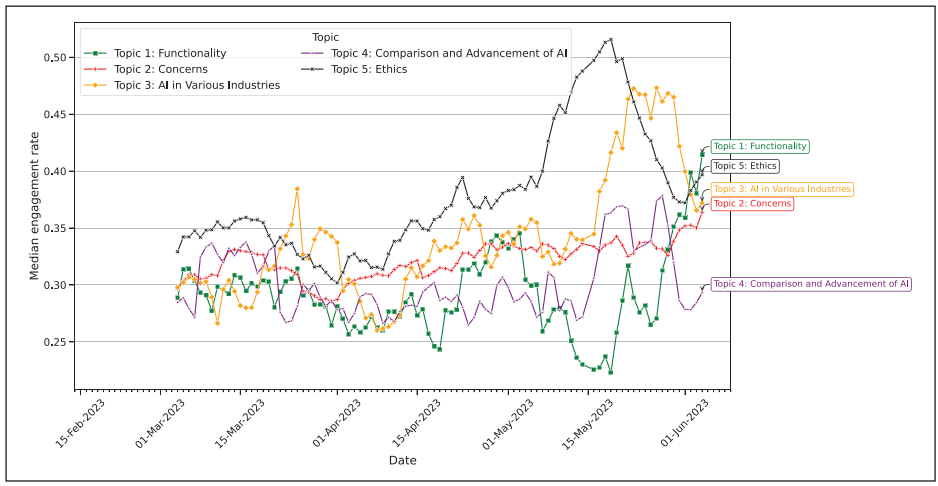

Figure 4. Daily median engagement rate of tweets with negative sentiment across discourse topics. Data is smoothed using a 14-day simple moving average.

the median engagement rate for the "Ethics" topic exceeded that of the "Concerns" topic (Nemenyi pairwise post hoc; $p < .001$), and particularly in late May 2023, where it was 1.5 times higher than that of the "Concerns" topic. Nevertheless, both topics exhibited similar and concurrent engagement trends of their tweets over time from February 2023 to April 2023 (Wilcoxon signed-rank test; $Z = 2052.0$, $p = .736$), that is, increase or decrease in the median engagement of tweets in one topic was also observed at the same time in the other topic. These observations could be interpreted as follows: Discussions under the "Concerns" topic may be plentiful, but each is typically more concise and attracts lesser engagement. Conversely, the "Ethics" topic hosts fewer discussions, indicating a lower number of original tweets sparking conversation (deduced from Supplementary Table S2). However, the individual tweets within the "Ethics" topic receive substantially more interaction compared with those categorized under the "Concerns" topic.

Within-topic main discourse facets

In light of our findings related to the two critical topics—"Concerns" and "Ethics"—we proceeded to further investigate the distinctive discourse characteristics inherent within each topic. By examining the phrases extracted from tweets in each topic using an LDA model, we can more precisely determine the specific nature of the discussions within each topic. Figure 5 shows an aggregated list of selected dominant phrases, scored in descending order across all topics (the full list can be found in Supplementary Table S2).

Using the list of phrases, we can find which phrases have a higher affinity to each topic, each reflecting a different facet of the discourse within the topic. Since LDA modeling learns the distribution of the main "topics" from the phrases in the corpus, all phrases are related to "all" topics, but with a different weight, and hence some phrases

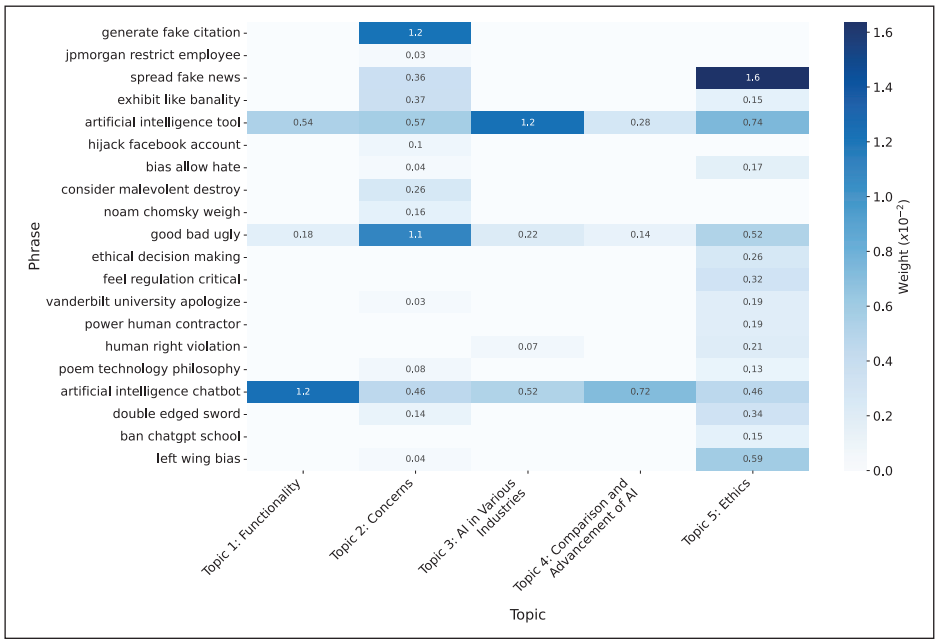

Figure 5. Selected phrases and their weight (i.e., score) across topics based on an LDA model induced from the tweets in each topic. Higher weight (darker color) indicates a higher affinity to the topic (i.e., a word is more informative within the topic discourse and more useful in describing the topic). Although the weight of words in some topics is very low (brighter color, zero-weight cells are not annotated), they can still occur in the tweets belonging to the topics but are less representative of the discourse.

received a very low weight (close to zero) with respect to some of the topics. Consequently, the weight assigned to each phrase conveys which discussion themes are more characteristic of the topic, providing us with a drill-down view of each topic. For example, we can deduce that phrases such as "artificial intelligence tool" and "artificial intelligence chatbot" are not unique to any topic and do not contribute to the analysis since they received relatively high weights across all topics. Conversely, "spread fake news" and "generate fake citations" received a very high weight only with respect to a single topic, which enhances their importance to understanding the underlying discussions in that topic. Furthermore, from the weighting of phrases in Figure 5, we can learn about the differences in the focal discussion within the topics "Concerns" and "Ethics." While some dominant phrases in the "Concerns" topic express the concerns of ChatGPT's users, others were assigned a higher importance within the "Ethics" topic. The "Concerns" topic was characterized by concerns of a more technical nature regarding generative AI, raising the concern of fake content generation ("generate fake citations"), banality ("exhibit like banality"), and the use of offensive language ("good bad ugly"). For example, tweets such as *"Asked ChatGPT for a reference of some info it gave me. It made one up! . . . sounds legitimate, but doesn't exist!!! . . . call that disturbing!"* illustrate those concerns.

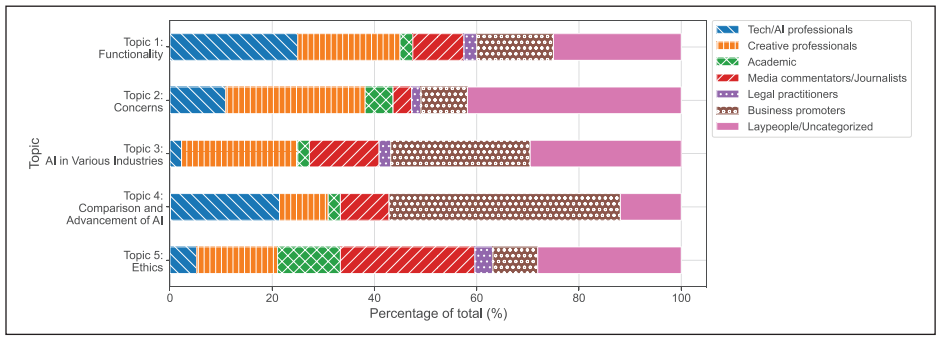

Figure 6. Percentage of users in each predefined group category among the 250 Twitter users with the top-engaging tweets across the analyzed topics. Users were categorized manually to each group based on their Twitter profiles.

More existential concerns were also noticed in the discourse in the "Concerns" topic, based on tweets, such as *"Will ChatGPT destroy human originality?! . . . Concerns that these AI models would flatten reality and homogenize human lives are understandable."* and *"ChatGPT is threatening to upend how we draft everyday communications like emails and college essays. Could AI also replace humans in the democratic process, specifically through lobbying?*" They sometimes encapsulate Twitter's often sarcastic tone while expressing profound philosophical questions regarding the humanity of AI, for example, *"This is scary. ChatGPT is now intelligent enough to match humans. It tries to come up with twenty wonderful things about [the city]; all it can say is that 'it's affordable.'"* On the contrary, phrases relating to ethical decision-making and bias surfaced higher with respect to the "Ethics" topic, as in the phrase "left wing bias" from tweets such as *". . . It was happy to tell me jokes about men but refused to tell me a joke about women. The left wing bias is baked right in.*" This supports our claim for the fundamental conversational distinction between the two topics.

User groups within ChatGPT discourse

The user group distribution, as shown in Figure 6, highlights the differences in the profiles of dominant participants in the ChatGPT debate across various discourse topics. The "Concerns" topic is predominantly engaged by creative professionals (such as designers, artists, and content creators) and a diverse range of uncategorized users, including anonymous individuals and laypeople. In contrast, the "Ethics" topic sees significant participation from media commentators and academics. This results in the discourse within the "Ethics" topic having fewer contributions from professionals using ChatGPT as a tool for enhancement and is characterized by a more scholarly tone, with frequent use of academic jargon to address ChatGPT's ethical implications. For example, Twitter users with academic backgrounds mentioned the ethical dangers in ChatGPT's development and its impact on scientific transparency and regulation, for example, *"regulating AI is an OpenAI ploy to hold onto a market share with government force. Do we*

really want . . . large language model to be as woke . . . as ChatGPT is?" and *". . . Tools such as ChatGPT threaten transparent science.*"

As expected, the discourse around ChatGPT's functional capabilities in Topic 1 ("Functionality") was primarily generated by professionals (in both software-related and creative fields) and business promoters with advertising purposes, featuring ChatGPT as a tool to promote their businesses. In addition, the representation of business profiles in varying industries was prominent in Topic 3 ("AI in Various Industries") and Topic 4 ("Comparison and AI Advancements") as their discussions focused on ChatGPT's applications in the industry. Examples of users from each group can be found in Supplementary Table S4.

Mirroring AI ethics principles

We observed that the conversation within the "Ethics" topic mirrors known AI ethics (Kazim and Koshiyama, 2021). First, fairness, which requires avoiding bias and discrimination, was inadvertently violated by ChatGPT in the initial phases following its launch, as was echoed in the discourse following biased, racist, or even sexist replies from ChatGPT (e.g., ". . . *I find ChatGPT in its current state a promo tool for plagiarism. On top of that, it's biased, racist, sexist . . . and mostly wrong in its replies and responses* . . ."). Second, we observed the reflection of AI transparency, serving as a property eliminating the "black-box" behavior of AI systems to provide insights into their decision-making. However, due to the complexity of algorithms such as ChatGPT's text generation, users often perceived this transparency as insufficient or entirely absent (e.g., *"We spent years trying to make AI transparent and open so we could weed out biases –ChatGPT has woke bias as a 'feature*,' *not a bug . . ."*). Finally, accountability, which is essential for responsible development under the ethical framework behind ChatGPT, was also partially observed in the ethical discourse (e.g., *". . . I think the distinction is accountability. I love ChatGPT but I'm terrified by the idea of putting it in charge . . . With bad human . . . when they mess up, you usually have some kind of clear legal recourse."*), particularly when OpenAI employed several moderation filters on its responses to protect against unethical answers (which were not endorsed by all its users, later jailbreaking ChatGPT to bypass its restrictions).

Discussion

Adoption of AI ethics discourse

Prior to ChatGPT's introduction, Khan et al. (2022) conducted a systematic literature review on AI ethics, highlighting transparency, privacy, accountability, and fairness as the most frequently discussed principles. In Khan et al's study, "AI ethics" involved a comprehensive review of academic literature to identify and quantify predominant themes. Our research, however, adopts a different perspective on AI ethics, employing a socio-cultural approach that we term "Ordinary AI ethics." Rather than engaging with AI ethics through technical, abstract, or legal lenses, our study is informed by the anthropological concept of "Ordinary Ethics" (Lambek, 2010), which considers ethics

as the reflective processes individuals use to discern right from wrong. For this study, Twitter (X) served as the primary platform where individuals expressed their reflections on ChatGPT during its initial months of deployment. This approach uncovered the ordinary ethics that shape users' understanding of ChatGPT and the imaginaries underpinning these ethical perceptions.

Our analysis of a substantial corpus reveals that discussions on Twitter regarding ChatGPT often mirror the academic AI ethics concepts that Khan et al. (2022) have illustrated. We observe how this initially academic discourse has become popularized. Many tweets on ethical topics originate from academics who initiate the discussion, and these tweets are subsequently disseminated by other Twitter users. Consequently, these core AI ethics concepts become highly trending topics on the platform and are widely discussed. To illustrate, the tweets we studied suggest a prevalent conceptualization of algorithms perpetuating societal inequalities and biases. In this context, examples of critiques such as the one in Table 1, where a user expresses concern over ChatGPT advising on how a female can enhance her femininity by suggesting stereotypical and biased methods (e.g., wearing makeup and maintaining physical appearance), are representative of the criticisms of bias that Twitter users posed when using ChatGPT for the first time, presuming it might harbor biases similarly to many of its algorithmic predecessors.

By observing the multitude of tweets on ethics and the high level of engagement they received, our research demonstrates that this activist-oriented academic literature on AI ethics and its key concepts have significantly influenced users of "Algorithmic Imaginaries" as they interacted with or observed ChatGPT during its early deployment. In other words, if users' interactions with algorithms are shaped by the imaginaries and folk theories they previously held about how algorithms function and how they should function (Bucher, 2017), we see here that individuals apply the imaginaries they already possessed about AI in general when engaging with the new form of generative AI presented by ChatGPT. This is evidenced by the prevalence of discussions about well-known AI ethics concepts such as "accountability," "privacy," and "fairness."

From master concepts to broadening the discussion on AI ethics

In another recent examination of AI ethics from a philosophical perspective, Tasioulas (2022) echoes the views of Khan et al. (2022), identifying the master concepts that have become the underlying imaginaries scholars consistently reference in the foundational AI ethics literature. Tasioulas critiques these discussions for their narrow focus on these specific subjects and foresees a shift in discourse as AI evolves. He predicts that while these master concepts will remain relevant, the discussion will broaden to explore new subjects related to human flourishing in the context of algorithms. This hypothesis is well-illustrated in our findings.

As we observed, the imaginaries people used when interacting with ChatGPT in its early months of deployment frequently drew on existing imaginaries seen in academic discussions on AI ethics, particularly these "master concepts" mentioned by Tasioulas (2022), which were already familiar and accessible to the academic figures pushing and disseminating these critiques. However, our findings indicate that this does not capture the full spectrum of users' critiques of ChatGPT on Twitter (X). The "Ethics" topic was

not the sole critical topic identified in our corpus. The "Concerns" topic garnered almost double the number of tweets compared with the "Ethics" topic.

Our LDA analysis suggested that, while not mutually exclusive, the themes discussed under the "Concerns" topic were, on the whole, distinct from those discussed under the "Ethics" topic. The "ethics" topic closely mirrored concepts central to AI ethics literature. In contrast, the "concerns" topic expanded the terminology and themes beyond the well-known central themes of AI ethics, incorporating additional themes that, while mentioned in the literature on AI ethics, were not previously central. We argue that the algorithmic imaginaries we see in the "concerns" topic have surfaced specifically in response to ChatGPT's advanced generative capabilities. While academic writing did address some of these topics, such as machine creativity (Epstein et al., 2023) and machine intelligence (Collins, 2018), they are relatively emerging and sporadic with regard to the research communities that study them. The "Concerns" topic uniquely discussed issues related to ChatGPT's generative capacities, such as generating fake news or citations and inciting hate, pointing to potential outcomes and societal impacts specific to generative AI technologies. Conversely, the "Ethics" topic focuses on a range of issues relevant not only to generative AI but also to any kind of algorithm. These include the risks associated with automation, crises related to human rights, and matters concerning bias. In this context, while the term "fake" is present in our LDA analysis in both topics, within the "Ethics" topic, it is contextualized as "spreading fake news," a critique that has long-standing roots in the discourse on AI ethics, illustrated by a rich body of literature on misinformation (Yu et al., 2023). Conversely, in the "Concerns" topic, "fake" is associated with "generating fake citations"—in this critique, the center of gravity is the algorithm's generative capabilities.

Another compelling illustration of how the definition of AI ethics may broaden through the interaction of individuals with generative AI is epitomized by the discussion over the banality of generative AI. This is observed through the frequently predictable and tool-like responses of ChatGPT (Leaver and Srdarov, 2023). Such concerns regarding the banality of ChatGPT could be perceived as an extension of pre-existing imaginaries concerning machine intelligence, specifically, the hazards associated with perceiving machine outputs as "intelligent" or "creative" (Collins, 2018). With the introduction of ChatGPT, these kinds of imaginaries of AI have crystallized and gained unprecedented relevance, manifesting as tangible issues when, for instance, students depend on generative AI for educational purposes (Adeshola and Adepoju, 2023), or when it is employed in academic research (Epstein et al., 2023).

Returning to the observation of master "AI ethics" concepts appearing alongside what we might call more nuanced and emerging features of AI ethics, this pattern is also demonstrated in our findings through the divergent ways in which our two critical topics—the "Ethics" and the "Concerns" topics—manifest. In the "Ethics" topic, representing the master concepts of AI ethics, we see individual tweets, often by academic figures, receiving notably more interaction from other users than tweets in the "Concerns" topic. This suggests the presence of a community with an established language and common imaginaries, making discussions more expansive. In contrast, the "Concerns" topic features more abundant but typically more concise discussions that attract less engagement. This differentiation indicates a more bottom-up process where users draw on and experiment

with different algorithmic imaginaries—such as the idea of existential risk and robots taking over the world—that already exist in the public imagination. However, these imaginaries are still somewhat less coherent and central, leading to fewer reactions compared with the well-established imaginaries in the "Ethics" topic.

Limitations and future directions

This study has several key limitations. First, it is focused on the discourse surrounding an emerging technology rather than the actual user experience with ChatGPT. We did not attempt to verify whether individuals discussing ChatGPT on Twitter ("X") have used the technology themselves or to understand the contexts in which they used it. Future studies could address this by adopting methodologies similar to Bucher's (2017) approach, such as administering questionnaires or conducting interviews with individuals who have posted about ChatGPT to gain more detailed insights into their experiences and how these relate to their online discussions.

Second, our analysis centered on the discourse and imaginaries surrounding a specific algorithm and we therefore did not examine the algorithm itself. As a result, we cannot determine whether the user reports reflect actual experiences or if they are replicable through engagement with ChatGPT. In addition, our study did not investigate ChatGPT's behavior, which limits our ability to explore the potentially significant relational dynamics between the algorithm and the critiques that emerged around it. Future research could aim to draw parallels or correlations between changes in ChatGPT's behavior over time and the various trending backlashes and critiques. This approach could provide valuable insights into how public critique and mass interaction patterns influence algorithms, considering ChatGPT's reliance on human feedback to some extent.

Another avenue for future research involves investigating how folk theories about algorithms, as exemplified in our "Concern" cluster, transition into topics of discussion within the academic community of algorithmic ethics. This could be achieved by tracking the emergence of folk theories on social media and examining how, when, and by whom similar themes become prominent in academic circles, both on social media and in academic publications.

Finally, this study is focused on the discourse in English on Twitter ("X"), which is one of many social networks where individuals discuss and share their experiences with generative AI. Future research could expand the analysis to include discourse in other languages and on other social platforms, such as Facebook and LinkedIn. This broader approach could provide a more comprehensive understanding and reveal different perspectives and cultural nuances that were not captured in our study.

Conclusion

The rise of generative AI, exemplified by ChatGPT, has shifted abstract debates to tangible discussions regarding the ethical implications of such technologies. Drawing from critical algorithm studies, which emphasize the "algorithmic imaginary"—the way individuals conceptualize and interact with algorithms (Bucher, 2017)—our research analyzed user discourse on ChatGPT during its early deployment months. As the user base

of ChatGPT expanded and ethical discussions intensified, we analyzed sentiments expressed in tweets about the AI, focusing on the primary ethical concerns voiced during its initial phase.

This study drew on a socio-cultural perspective on ethics, suggesting a focus on what we term "Ordinary AI ethics." Instead of approaching AI ethics as technical, abstract, or legal debates, this study draws on the anthropological idea of "Ordinary Ethics" (Lambek, 2010), viewing ethics as the reflective processes of actual people as they consider what is right or wrong. For this work, Twitter ("X") was the main site in which people articulated their reflections on ChatGPT in its first months of deployment. This approach revealed the ordinary ethics that inform users' understanding of ChatGPT and the imaginaries on which these ethical understandings are based.

In the context of Twitter discourse surrounding OpenAI's ChatGPT, our study uncovered a notable interplay between established AI ethics dialogues and emerging apprehensions about the model's generative capabilities. A comprehensive analysis of Twitter content revealed two main discussion topics: "Ethics" and "Concerns." The "Ethics" topic encompassed academic dialogues on AI principles, covering transparency, privacy, accountability, and fairness. In contrast, the "Concerns" topic presented a novel critical perspective, under which discussions centered on ChatGPT's unique generative features, including its potential to produce inaccurate information or references, possibilities for misuse, and more profound philosophical reflections on the model's role within societal frameworks. This divergence suggests that as AI technologies evolve, the public is simultaneously referencing and adopting popular AI ethics imaginaries while branching into nuanced critiques tailored to the intricacies of contemporary technological advancements.

Acknowledgements

We would like to express our gratitude to the Azrieli Foundation and Tel Aviv University's Center for AI and Data Science (TAD) for their generous support of this research. We also thank the anonymous reviewers of this article for their valuable feedback. In addition, we extend our thanks to Roi Cohen for his advice and support during the initial stages of this project.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Azrieli Foundation and Tel Aviv University's Center for AI and Data Science (TAD) generously supported this research.

ORCID iDs

Maayan Cohen https://orcid.org/0000-0001-9497-7581

Michael Khavkin https://orcid.org/0000-0002-7416-2243

Danielle Movsowitz Davidow https://orcid.org/0000-0003-4475-662X

Supplemental material

Supplemental material for this article is available online.

Notes

1. Definitions of Twitter's account activity metrics (Tweet activity dashboard) are available at https://help.twitter.com/en/managing-your-account/using-the-tweet-activity-dashboard
2. Available at https://huggingface.co/finiteautomata/bertweet-base-sentiment-analysis

References

Abid A, Farooqi M and Zou J (2021) Large language models associate Muslims with violence. *Nature Machine Intelligence* 3(6): 461–463.

Adeshola I and Adepoju AP (2023) The opportunities and challenges of ChatGPT in education. *Interactive Learning Environments*. Epub ahead of print 4 September. DOI: 10.1080/10494820.2023.2253858.

Avnoon N, Kotliar DM and Rivnai-Bahir S (2023) Contextualizing the ethics of algorithms: a socio-professional approach. *New Media & Society*. Epub ahead of print 5 January. DOI: 10.1177/14614448221145728.

Bender EM, Gebru T, McMillan-Major A, et al. (2021) On the dangers of stochastic parrots: can language models be too big? In: *FAccT '21: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, Virtual Event, 3–10 March, pp. 610–623. New York: ACM.

Blei DM (2012) Probabilistic topic models. *Communications of the ACM* 55(4): 77–84.

Boltanski L and Thévenot L (1999) The sociology of critical capacity. *European Journal of Social Theory* 2(3): 359–377.

Brey P and Dainow B (2023) Ethics by design for artificial intelligence. *AI and Ethics*. Epub ahead of print 21 September. DOI: 10.1007/s43681-023-00330-4.

Brown S, Davidovic J and Hasan A (2021) The algorithm audit: scoring the algorithms that score us. *Big Data & Society* 8(1): 2053951720983865.

Brynjolfsson E, Li D and Raymond LR (2023) *Generative AI at work*. Technical Report. Cambridge, MA: National Bureau of Economic Research.

Bucher T (2017) The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society* 20(1): 30–44.

Bucher T (2018) *If. . . Then: Algorithmic Power and Politics*. Oxford: Oxford University Press.

Büchi M, Fosch-Villaronga E, Lutz C, et al. (2023) Making sense of algorithmic profiling: user perceptions on Facebook. *Information, Communication & Society* 26(4): 809–825.

Cave S, Coughlan K and Dihal K (2019) "Scary robots": examining public responses to AI. In: *AIES '19: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society*, Honolulu, HI, 27–28 January, pp. 331–337. New York: ACM.

Christin A (2017) Algorithms in practice: comparing web journalism and criminal justice. *Big Data & Society* 4(2): 2053951717718855.

Collins H (2018) *Artifictional Intelligence: Against Humanity's Surrender to Computers*. Hoboken, NJ: John Wiley & Sons.

Devendorf L and Goodman E (2014) The algorithm multiple, the algorithm material: reconstructing creative practice. *UC Davis' Contours of Algorithmic Life Conference*. Available at: https://www.confectious.net/may-15-the-algorithm-multiple-the-algorithm-material-reconstructing-creative-practice-uc-davis/

DeVito MA, Gergle D and Birnholtz J (2017) "Algorithms ruin everything": #riptwitter, folk theories, and resistance to algorithmic change in social media. In: *CHI '17: Proceedings of the*

2017 CHI conference on human factors in computing systems, Denver, CO, 6–11 May, pp. 3163–3174. New York: ACM.

Díaz-Rodríguez N, Del Ser J, Coeckelbergh M, et al. (2023) Connecting the dots in trustworthy artificial intelligence: from AI principles, ethics, and key requirements to responsible AI systems and regulation. *Information Fusion* 99: 101896.

Dodd H (2023) Sentiment analysis of tweets about ChatGPT. Available at: https://ir.lib.uwo.ca/cgi/viewcontent.cgi?article=1010&context=fims_evolvingtech_finalproj_summer2023

Epstein Z and Hertzmann A, of Human Creativity I, et al. (2023) Art and the science of generative AI. *Science* 380(6650): 1110–1111.

Eslami M, Karahalios K, Sandvig C, et al. (2016) First I "like" it, then I hide it: folk theories of social feeds. In: *CHI '16: Proceedings of the 2016 CHI conference on human factors in computing systems*, San Jose, CA, 7–12 May, pp. 2371–2382. New York: ACM.

Fiesler C and Proferes N (2018) "Participant" perceptions of Twitter research ethics. *Social Media + Society* 4(1): 2056305118763366.

Fletcher R and Nielsen RK (2019) Generalised scepticism: how people navigate news on social media. *Information, Communication & Society* 22(12): 1751–1769.

Future of Life Institute (2023) Pause giant AI experiments: an open letter. Available at: https://futureoflife.org/open-letter/pause-giant-ai-experiments/ (accessed 13 October 2023).

Ghosh S and Caliskan A (2023) Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: findings across Bengali and five other low-resource languages. In: *AIES '23: Proceedings of the 2023 AAAI/ACM conference on AI, ethics, and society*, Montreal, QC, Canada, 8–10 August, pp. 901–912. New York: ACM.

Ghotbi N, Ho MT and Mantello P (2022) Attitude of college students towards ethical issues of artificial intelligence in an international university in Japan. *AI & Society* 37: 283–290.

Gilardi F, Alizadeh M and Kubli M (2023) Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* 120(30): e2305016120.

Gillespie T (2014) The relevance of algorithms. *Media Technologies: Essays on Communication, Materiality, and Society* 167(2014): 167.

Gross N (2023) What ChatGPT tells us about gender: a cautionary tale about performativity and gender biases in AI. *Social Sciences* 12(8): 435.

Hitlin S, Dromi SM and Luft A (2023) *Handbook of the Sociology of Morality*, vol. 2. London: Springer.

HLEG A (2019) High-level expert group on artificial intelligence. *Ethics Guidelines for Trustworthy AI 6*. Available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Htet A, Liana SR, Aung T, et al. (2024) Chatgpt in content creation: techniques, applications, and ethical implications. In: Obaid AJ, Bhushan B, Muthmainnah S, et al. (eds) *Advanced Applications of Generative AI and Natural Language Processing Models*. Hershey, PA: IGI Global, pp. 43–68.

Hu Y, Ameer I, Zuo X, et al. (2023) Zero-shot clinical entity recognition using ChatGPT. arXiv [preprint]. DOI: 10.48550/arXiv.2303.16416.

Jones P (2021) *Work without the Worker: Labour in the Age of Platform Capitalism*. London: Verso Books.

Kasirzadeh A and Gabriel I (2023) In conversation with artificial intelligence: aligning language models with human values. *Philosophy & Technology* 36(2): 1–24.

Kazim E and Koshiyama AS (2021) A high-level overview of AI ethics. *Patterns* 2(9): 100314.

Kennedy H, Elgesem D and Miguel C (2017) On fairness: user perspectives on social media data mining. *Convergence* 23(3): 270–288.

Khan AA, Badshah S, Liang P, et al. (2022) Ethics of AI: a systematic literature review of principles and challenges. In: *EASE '22: Proceedings of the 26th international conference on evaluation and assessment in software engineering*, Gothenburg, 13–15 June, pp. 383–392. New York: ACM.

Kojima T, Gu SS, Reid M, et al. (2022) Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems* 35: 22199–22213.

Korkmaz A, Aktürk C and Talan T (2023) Analyzing the user's sentiments of ChatGPT using twitter data. *Iraqi Journal for Computer Science and Mathematics* 4(2): 202–214.

Kotliar DM (2020) The return of the social: algorithmic identity in an age of symbolic demise. *New Media & Society* 22(7): 1152–1167.

Laakasuo M, Herzon V, Perander S, et al. (2021) Socio-cognitive biases in folk AI ethics and risk discourse. *AI and Ethics* 1(4): 593–610.

Laidlaw J (2018) Fault lines in the anthropology of ethics. In: Mattingly C, Dyring R, Louw M, et al. (eds) *Moral Engines: Exploring the Ethical Drives in Human Life*. New York: Berghahn Books, pp. 174–196.

Lambek M (2010) *Ordinary Ethics: Anthropology, Language, and Action*. New York: Fordham University Press.

Lamont M (2009) *The Dignity of Working Men: Morality and the Boundaries of Race, Class, and Immigration*. Cambridge, MA: Harvard University Press.

Larochelle H, Erhan D and Bengio Y (2008) Zero-data learning of new tasks. *AAAI* 1: 3.

Laufer B, Jain S, Cooper AF, et al. (2022) Four years of FAccT: a reflexive, mixed-methods analysis of research contributions, shortcomings, and future prospects. In: *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, Seoul, Republic of Korea, 21–24 June, pp. 401–426. New York: ACM.

Leaver T and Srdarov S (2023) ChatGPT isn't magic: the hype and hypocrisy of generative artificial intelligence (AI) rhetoric. *M/C Journal* 26(5): 3004.

Liesenfeld A, Lopez A and Dingemanse M (2023) Opening up ChatGPT: tracking openness, transparency, and accountability in instruction-tuned text generators. In: *Proceedings of the 5th international conference on conversational user interfaces*, Eindhoven, 19–21 July, pp. 1–6. New York: ACM.

Lim WM, Gunasekara A, Pallant JL, et al. (2023) Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. *The International Journal of Management Education* 21(2): 100790.

Lomborg S and Kapsch PH (2020) Decoding algorithms. *Media, Culture & Society* 42(5): 745–761.

Lukowicz P (2019) The challenge of human centric AI. *Digitale Welt* 3(4): 9–10.

Mahmood S (2005) *Politics of Piety: The Islamic Revival and the Feminist Subject*. Princeton, NJ: Princeton University Press.

Mattingly C and Throop J (2018) The anthropology of ethics and morality. *Annual Review of Anthropology* 47: 475–492.

Nerdy N (2024) Up-to-date ChatGPT statistics & user numbers. Available at: https://nerdynav.com/chatgpt-statistics/ (Accessed 29 January 2024).

Niforatos E, Palma A, Gluszny R, et al. (2020) Would you do it?: enacting moral dilemmas in virtual reality for understanding ethical decision-making. In: *CHI '20: Proceedings of the 2020 CHI conference on human factors in computing systems*, Honolulu, HI, 25–30 April, pp. 1–12. New York: ACM.

Nissim G and Simon T (2021) The future of labor unions in the age of automation and at the dawn of AI. *Technology in Society* 67: 101732.

Rader E and Gray R (2015) Understanding user beliefs about algorithmic curation in the Facebook news feed. In: *CHI '15: Proceedings of the 33rd annual ACM conference on human factors in computing systems*, Seoul, Republic of Korea, 18–23 April, pp. 173–182. New York: ACM.

Rathje S, Mirea DM, Sucholutsky I, et al. (2024) GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences* 121: e2308950121.

Ruckenstein M and Granroth J (2020) Algorithms, advertising and the intimacy of surveillance. *Journal of Cultural Economy* 13(1): 12–24.

Schellewald A (2022) Theorizing "stories about algorithms" as a mechanism in the formation and maintenance of algorithmic imaginaries. *Social Media & Society* 8(1): 20563051221077025.

Seaver N (2017) Algorithms as culture: some tactics for the ethnography of algorithmic systems. *Big Data & Society* 4(2): 2053951717738104.

Siles I, Segura-Castillo A, Solís R, et al. (2020) Folk theories of algorithmic recommendations on Spotify: enacting data assemblages in the global south. *Big Data & Society* 7(1): 2053951720923377.

Smuha NA (2021) From a "race to AI" to a 'race to ai regulation': regulatory competition for artificial intelligence. *Law, Innovation and Technology* 13(1): 57–84.

Swart J (2021) Experiencing algorithms: how young people understand, feel about, and engage with algorithmic news selection on social media. *Social Media & Society* 7(2): 20563051211008828.

Swidler A (1986) Culture in action: symbols and strategies. *American Sociological Review* 51: 273–286.

Taecharungroj V (2023) "What can ChatGPT do?" analyzing early reactions to the innovative AI Chatbot on Twitter. *Big Data and Cognitive Computing* 7(1): 35.

Tasioulas J (2022) Artificial intelligence, humanistic ethics. *Daedalus* 151(2): 232–243.

Törnberg P (2023) ChatGPT-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. arXiv [preprint]. DOI: 10.48550/arXiv.2304.06588.

Van Berkel N, Sarsenbayeva Z and Goncalves J (2023) The methodology of studying fairness perceptions in artificial intelligence: contrasting CHI and FAccT. *International Journal of Human-Computer Studies* 170: 102954.

Weidinger L, Uesato J, Rauh M, et al. (2022) Taxonomy of risks posed by language models. In: *FAccT '22: Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, Seoul, Republic of Korea, 21–24 June, pp. 214–229. New York: ACM.

Yu W, Payton B, Sun M, et al. (2023) Toward an integrated framework for misinformation and correction sharing: a systematic review across domains. *New Media & Society* 25(8): 2241–2267.

Ziegler DM, Stiennon N, Wu J, et al. (2019) Fine-tuning language models from human preferences. arXiv [preprint]. DOI: 10.48550/arXiv.1909.08593.

Author biographies

Maayan Cohen is a socio-cultural anthropologist who specializes in the study of media production. She earned her doctoral degree at the University of Oxford as a Rhodes Scholar and currently serves as an Azrieli Postdoctoral Fellow and Departmental Lecturer in the Department of Sociology and Anthropology at Tel Aviv University. Maayan is also a postdoctoral affiliate at the University of Oxford's School of Anthropology and Museum Ethnography. Her current research project investigates how culture shapes conversational AI in the United States and Israel.

Michael Khavkin is a PhD candidate in the Department of Industrial Engineering at Tel Aviv University. His research interests include machine learning and the intersection of human–computer interaction (HCI) and privacy-preserving data analysis with differential privacy.

Danielle Movsowitz Davidow is a PhD candidate in the Department of Industrial Engineering at Tel Aviv University. Her research interests include human–AI interactions and privacy and security, focusing on privacy-enhancing technologies and blockchain privacy.

Eran Toch is an associate professor in the Faculty of Engineering at Tel Aviv University, where he also heads the Industrial Engineering Department. He co-directs the Interacting with Technology Lab (IWiT), focusing on areas such as data engineering, user-friendly privacy and security, and human–AI interaction.