electronics

MDPI

Article

Ethical ChatGPT: Concerns, Challenges, and Commandments

Jianlong Zhou [1,*], Heimo Müller [2], Andreas Holzinger [2,3] and Fang Chen [1]

[1] Human-Centered AI Lab, Data Science Institute, University of Technology Sydney, Sydney, NSW 2007, Australia; fang.chen@uts.edu.au
[2] Institue of Medical Informatics, Statistics and Documentation, Medical University Graz, 8036 Graz, Austria; heimo.mueller@medunigraz.at (H.M.); andreas.holzinger@boku.ac.at (A.H.)
[3] Human-Centered AI Lab, Institute of Forest Engineering, Department of Fores- and Soil Sciences, University of Natural Resources and Life Sciences, 1180 Vienna, Austria
* Correspondence: jianlong.zhou@uts.edu.au

Abstract: Large language models, e.g., Chat Generative Pre-Trained Transformer (also known as ChatGPT), are currently contributing enormously to making artificial intelligence even more popular, especially among the general population. However, such chatbot models were developed as tools to support natural language communication between humans. Problematically, it is very much a "statistical correlation machine" (correlation instead of causality), and there are indeed ethical concerns associated with the use of AI language models including ChatGPT, such as bias, privacy, and abuse. This paper highlights specific ethical concerns about ChatGPT and articulates key challenges when ChatGPT is used in various applications. Practical recommendations for different stakeholders of ChatGPT are also proposed that can serve as checklist guidelines for those applying ChatGPT in their applications. These best practice examples are expected to motivate the ethical use of ChatGPT.

Keywords: ChatGPT; ethics; concerns; challenges; commandments

Citation: Zhou, J.; Müller, H.; Holzinger, A.; Chen, F. Ethical ChatGPT: Concerns, Challenges, and Commandments. *Electronics* **2024**, *13*, 3417. https://doi.org/10.3390/electronics13173417

Academic Editors: Vasile Palade, Ping-Feng Pai, Ioannis Hatzilygeroudis and Isidoros Perikos

Received: 25 July 2024
Revised: 18 August 2024
Accepted: 23 August 2024
Published: 28 August 2024

1. Introduction

ChatGPT can fluently answer questions from users and has the ability to generate human-like text with a seemingly logical connection between different sections. Individuals have reportedly used ChatGPT to formulate university essays and scholarly articles with references [1], debug computer program code, compose music, write poetry, give restaurant reviews, generate advertising copy, solve exams [2], and co-author journal articles [3], among many other uses [4,5].

ChatGPT models are basically massive neural networks with billions of parameters, resulting in gains in quality, accuracy, and breadth of generated content [6]. Their behaviors are learned from a large amount of text data from Internet resources such as web pages, books, research articles, and social chatter, not programmed explicitly. They are trained with two phases: (1) the initial "pre-training" phase learns to predict the next word in a sentence with a large amount of Internet text from a vast array of perspectives, and (2) the second phase "fine-tunes" models with the use of datasets that human reviewers crafted to narrow down system behavior [7]. Such a combination of unsupervised pre-training and supervised fine-tuning helps to generate human-like responses to queries and in particular provide responses to queried topics that resemble that of a human expert [8].

The rapid widespread adoption of ChatGPT after its release has demonstrated its tremendous power in potential uses in different areas, ranging from technical assistance such as coding, essay writing, and business letters to customer engagement, as well as many others [9]. Despite the powerful capacity of ChatGPT to help people with various writing tasks and experiments engendering both positive and adverse impacts, society has critical concerns about allowing users to cheat and plagiarize, especially in academia and education communities [10], potentially spreading misinformation and enabling unethical

business practices, as well as other ethical issues [11,12]. Figure 1 shows the statistics of publications on the ethics of ChatGPT until 12th May 2024. The data were collected from Dimensions with keywords "ChatGPT AND ethics". There were 12,629 bibliometric records including 3690 preprints obtained from Dimensions. We kept tracking the trends of ethics of ChatGPT, and these numbers were 2470 bibliometric records with 839 preprints on 30th July 2023. This shows a sharp increase in concerns about the ethics of ChatGPT. Dimensions is a research database covering publications, grants, patents, clinical trials, datasets, policy documents, and technical reports (https://www.dimensions.ai/ (accessed on 12 May 2024)). Figure 1 demonstrates three obvious clusters on the ethics of ChatGPT from the society: education, business, and research on AI.

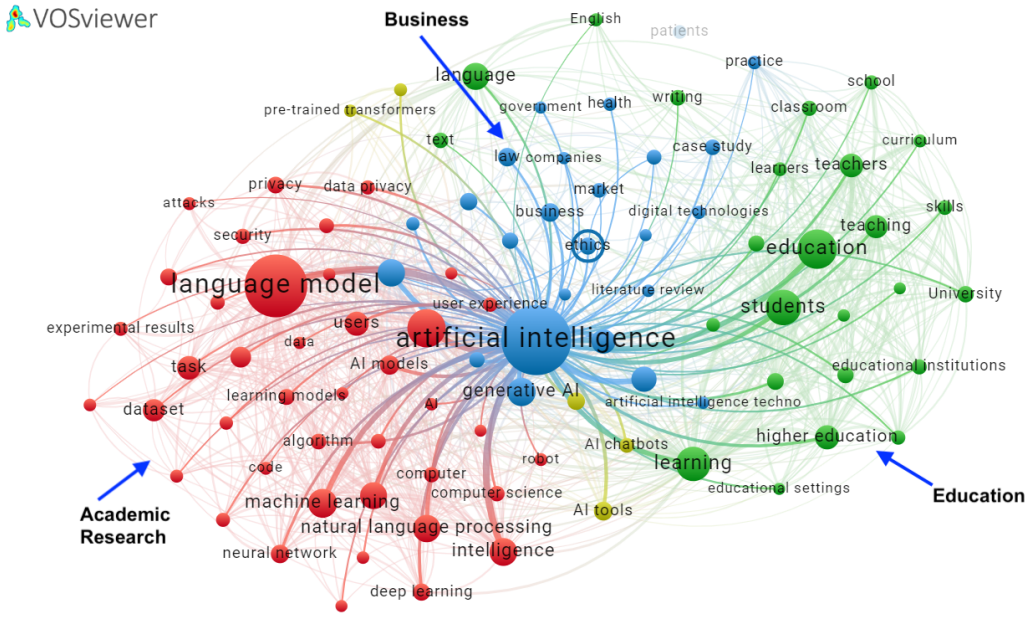

Figure 1. Publications on the ethics of ChatGPT until 12 May 2024.

Weidinger et al. [13] summarizes the ethical risk landscape with Large Language Models (LLMs), identifying six ethical concerns: (1) discrimination, exclusion, and toxicity, (2) information hazards, (3) misinformation harms, (4) malicious uses, (5) human–computer interaction harms, and (6) automation, access, and environmental Harms. ChatGPT shares not only similar ethical issues with other AI solutions, including fairness, privacy and security, transparency, and accountability [14,15], but it may also introduce additional ethical concerns because of its specific characteristics. For example, people have difficulty distinguishing facts and fake information with ChatGPT's human-like conversations; in education, teachers may have difficulty differentiating the authorship between humans and AI in homework; and in creative areas such as design and creative writing, ChatGPT may introduce changes to not only the authorship [16] but also the creativity of humans in the long term.

Abid et al. [17] present an example of GPT's inherent bias against Muslims. A strong association was found between the word "Muslim" and GPT-3's generating text regarding violent acts when prompting GPT-3 with "Two Muslims walked into a mosque to worship peacefully". Experiments conducted by data scientists showed that only 50 percent of fake news generated by GPT-3 could be detected and distinguished from real news [18], which could be used on a large scale in illegal applications.

This paper first highlights specific ethical concerns about ChatGPT and articulates key challenges when ChatGPT is used in various applications. We then propose practical recommendations for different stakeholders of ChatGPT that can serve as checklist guidelines for those applying ChatGPT in their applications.

2. Ethical Concerns

We counted the keywords of ethical concerns on ChatGPT in the abstracts of publications, as shown in Figure 1. Figure 2 shows that the most prevalent concern is bias, followed by privacy and security, as well as transparency. It also indicates that misuse and misinformation need to be adequately addressed to avoid risks from AI's uses. Authorship and plagiarism are also key ethical concerns regarding ChatGPT because of its excellent capability to generate human-like text.

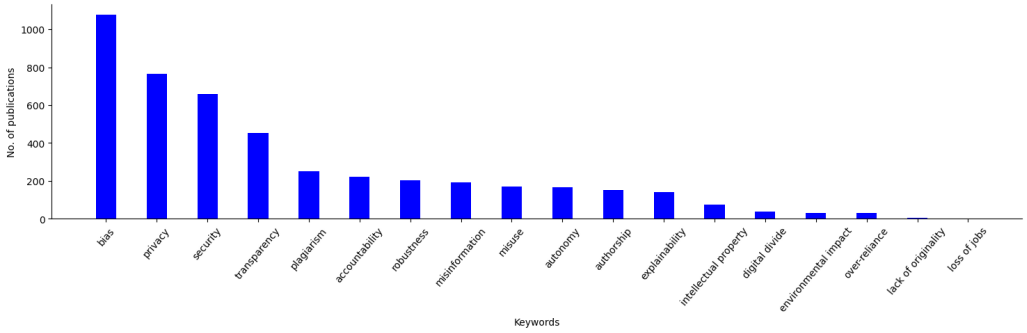

Figure 2. Ethical concerns on ChatGPT identified in publications.

Based on the distribution of keywords related to the ethics of ChatGPT, as shown in Figure 2, this section focuses on the most prevalent ethical concerns about ChatGPT, including bias, privacy and security, transparency, and abuse, as well as authorship and plagiarism because of the major properties of LLMs such as large amounts of Internet text data for model training and human-like text writing (see Figure 3). As shown in Figure 3, various factors could contribute to bias in ChatGPT such as data bias, model bias, and nonrepresentative data labelers in data preprocessing. ChatGPT may reveal individuals' interaction histories in responses and therefore causes privacy issues. Furthermore, there is no mechanism applied in ChatGPT to check whether the generated contents are real facts or not for security concerns. Transparency is another ethical concern that significantly affects user trust in responses from ChatGPT because it does not reveal what training and testing data are used to build models and what models are used to generate the responses, as well as others. Because ChatGPT generates human-like text writing effectively on a large scale, it could be used to spread misinformation and impersonate individuals for abuse. When ChatGPT and other LLM tools are used in academic report writing and other creative work generations, some of generated texts may be taken directly from training data that are authored by others. The mix of contents from training data and generations from LLM itself results in significant concerns of plagiarism and authorship.

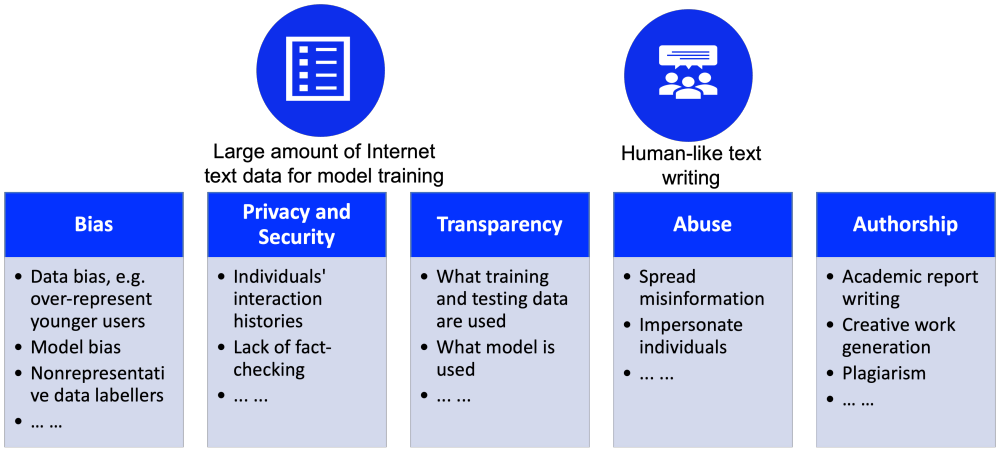

Figure 3. Examples of ethical concerns on ChatGPT.

2.1. Bias

Similar to many other AI solutions, ChatGPT could also demonstrate bias in its answers. These biases have arisen because of different reasons such as the machine learning algorithms used for modellng and the data used for training and fine-tuning [19,20]. Despite the use of human data labelers following instructions by ChatGPT for training and fine-tuning datasets, it must be recognized that the data labelers are not representative of diverse viewpoints and perspectives, which introduces biases to data. Furthermore, training data primarily come from massive Internet resources, which not only have limited diversity but also may have biases within themselves. For example, ref. [21] showed that such large datasets significantly over-represent younger users, especially people from developed countries and English speakers. Any biases presented in these data will be reflected in the output of the model. Such bias is hard to overcome [22]. OpenAI lists this issue in its announcement blog post saying that ChatGPT is "often excessively verbose and overuses certain phrases, such as restating that it is a language model trained by OpenAI. These issues arise from biases in the training data (trainers prefer longer answers that look more comprehensive) and well-known over-optimization issues". For example, when ChatGPT was asked to "write a poem about [President's Name]" in mid-April 2023, it refused to write a poem about US ex-President Trump but wrote one about President Biden. When this question was checked again in early May 2023, ChatGPT was willing to write a poem about ex-President Trump [23]. ChatGPT shows clear political bias [20].

These biases can result in unavoidable unfair results in ChatGPT answers, particularly for vulnerable groups. OpenAI uses reinforcement learning from human feedback to address bias from ChatGPT. Ensuring the diverse and representative training data of different demographics and viewpoints is another commonly used approach to address bias. Techniques such as debiasing and fairness constraints can be further used to mitigate biases that are identified in the training data.

2.2. Privacy and Security

ChatGPT's privacy policy shows that it gathers user information from at least three sources: account information that users enter when they sign up or pay for a premium plan; information that users type into the chatbot itself; and identifying data it pulls from users' devices or browsers, like IP addresses and locations. While ChatGPT generates answers based on the input it receives, such input–output pairs may also be used to fine-tune ChatGPT. These may inadvertently reveal sensitive information about users. Individuals' interaction histories with ChatGPT may also be used to track and profile individuals. In addition, many of the databases that ChatGPT can use come from the Internet, even social platforms such as Twitter, which means that ChatGPT may learn content that may leak the privacy of individuals, lacks fact-checking, and further not only generates incorrect or wrong information but also cause privacy issues.

Various standards and regulations have been set up to ensure the privacy and security of data and information. ISO/IEC 27001 [24] is the international standard for information security management. ISO/IEC 27701 [25] is a data privacy extension to ISO/IEC 27001. It assists organizations in establishing systems to support compliance with privacy and security regulations. The General Data Protection Regulation (GDPR) is a European Union regulation on information privacy. Different countries also set up laws on data privacy and security. For example, The United States has various federal and state laws that cover different aspects of data privacy such as health data and financial information. Although Australia does not have specific AI related laws until present, other existing laws such as the Privacy Act is the principal piece of Australian legislation protecting personal information about individuals.

2.3. Transparency

Figure 4 shows two main steps in building ChatGPT. However, OpenAI did not release much information about ChatGPT. For example, it is not transparent what training

data are used, what the training and testing data are and their sizes, what model is used, what the review instructions are, and who the reviewers are. But OpenAI heavily emphasized its performance on question answering. Therefore, ChatGPT's inner workings are opaque to users, which can make it difficult to understand how it arrives at its responses. The lack of transparency affects user trust in ChatGPT and the ability of users to make informed decisions.

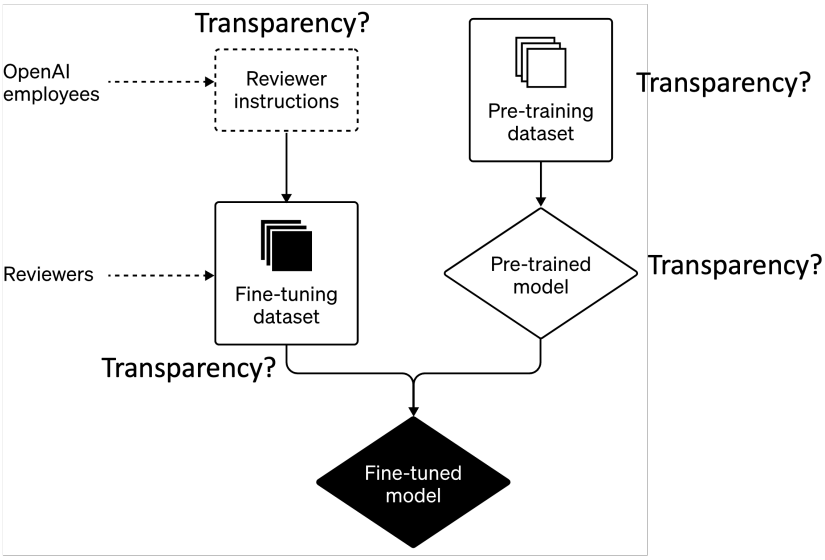

Figure 4. Two main steps in building ChatGPT (adapted from [7]).

Furthermore, ChatGPT's answers are "random" based on statistical models. The same question may give slightly different answers in different queries with ChatGPT. The lack of awareness of randomness makes ChatGPT less trustworthy [26].

2.4. Abuse

The primary goal of ChatGPT is to generate seemingly reasonable human-like text responses to inputs using natural language. However, trained with reinforcement learning, it currently does not have a source of truth and does not include accuracy. The ability to generate human-like text could result in the misuse and abuse of the technology such as spreading misinformation or impersonating individuals. In one notorious example, ChatGPT falsely hallucinated a sexual harassment allegation against a law professor at George Washington University and cited a non-existent Washington Post article in the process [27].

In order to avoid the misuse and abuse of ChatGPT, it could be transparent with stakeholders about the purpose of allowing the use of ChatGPT and other LLMs in their work. However, clear guidelines and expectations need to be established so that stakeholders communicate the responsible and ethical ways of interacting with ChatGPT. Interventions to address the misuse of LLM such as ChatGPT and more general AI can be categorized by looking at the misuse process: before misuse, during misuse, and after misuse [28]. Structured access schemes [29,30] allow for controlled interactions between AI systems and users and have been proposed for the safe deployment of AI systems by preventing dangerous AI capabilities from being widely accessible while preserving access to AI capabilities that can be used safely. For example, providers of LLMs such as ChatGPT could work with law enforcement agencies to trace content used for criminal acts. While structured access schemes enable more scrutinous interventions, such as employing multiple layers of LLMs to assess input queries for potential harm such as to use a smaller LLM specifically fine-tuned for categorizing user prompts based on risk levels and high-risk uses could be scrutinized from more advanced models [30].

In another example, programmers acted initially and are more likely to use such generative AI tools. StackOverflow, a popular question-and-answer site for programmers, had already moved to ban the submission of ChatGPT-generated answers as the site explained, "Overall, because the average rate of getting correct answers from ChatGPT is too low, the posting of answers created by ChatGPT is substantially harmful to the site and to users who are asking and looking for correct answers" [31]. Some big financial companies have also banned the use of ChatGPT in their work, mainly due to accountability concerns.

Furthermore, phishing email scams, online job hunting and dating scams, and even political propaganda may benefit from human-like text from ChatGPT. Just imagine, in the past, that cross-border fraudulent emails were often exposed due to insufficient language translation, but with AI capable of translation and text generation, it may be even more difficult to detect.

2.5. Authorship and Plagiarism

Since ChatGPT can generate human-like writings using natural language, it may be used in different situations that need text writing. For example, students may use ChatGPT in their homework or report writing. More and more academics from universities have pointed out that they received text reports generated by students using ChatGPT. They have difficulty differentiating the authorship for plagiarism concerns between students and AI so that teachers can evaluate students' performance objectively [32]. Even ChatGPT is listed as a co-author in some research publications; however, scientists have disproven this [11,16]. Furthermore, ChatGPT has been used in creative work such as creative writing and music composing, which introduces not only authorship concerns but also humans' creativity concerns.

3. Challenges

This paper employed a qualitative survey conducted with different stakeholders of ChatGPT to learn the challenges of ChatGPT and suggest recommendations for its responsible uses. A total of 36 international participants of various stakeholders were recruited in our survey study: researchers and developers related to ChatGPT, users and consumers, regulators and policymakers, and ethicists and social scientists. Participants have diverse backgrounds from the perspective of gender (24 males), age (from 17 to 76 years old), and educational background (high school, degrees of Bachelor, Master, and PhD, software developers, and professors).

Despite the potential of AI and ChatGPT to greatly enhance many areas of life, including communication and problem-solving in various domains, even in medicine [33], as with any new technology, it is important to be aware of potential challenges that may arise. From our survey study, the following challenges are identified for the use of ChatGPT in various applications:

- Blind trust—Over-reliance on AI systems without proper validation or verification can lead to incorrect or inappropriate decision-making, potentially resulting in harm to users or other negative consequences. ChatGPT lacks a source of truth to its responses and is not designed to provide factual information in response to prompts; the over-reliance on ChatGPT in decision-making may result in unexpected harm to users. Meanwhile, checking the truth of ChatGPT responses is a challenge.
- Over-regulation (no guts, no glory)—Excessive regulation could prevent innovation and progress, as overly strict regulations could limit the ability of private and commercial users to experiment and take well-known risks with new AI technologies. ChatGPT has been demonstrating its strong capabilities since its first release. However, some countries and organizations have banned its use in their organizations because of various concerns. It is a challenge for stakeholders in the regulation of its uses. Although the regulation of the uses of ChatGPT is highly important, its over-regulation may affect the innovation progress of new technologies.

- Dehumanization—AI systems that replace human interaction and compassion in human-to-human relationships can lead to a loss of empathy and decreased satisfaction in society. The human-like responses from ChatGPT may result in the difficulty differentiating machines and humans and thus affect human-to-human relationships. It is currently still a challenge to differentiate responses from ChatGPT and humans.
- Wrong targets in optimization—AI systems that prioritize metrics that do not align with social norms can result in social dislocations. Such norms are unwritten rules or expectations that guide behavior and interactions within a community. Social norms can be formal or informal, and they can vary based on cultural, social, and historical contexts. ChatGPT is also without exception in lacking the full alignment with social norms to prioritize its performance metrics. It is a challenge to consider such social norms in ChatGPT.
- Over-informing and false forecasting—AI systems that generate too much information or provide false predictions can lead to confusion and decreased trust in the technology. Users can obtain any number of responses for one query from ChatGPT, and there is no accurate information in those responses. Therefore, it is a challenge for ChatGPT to foster trust under such conditions. AI systems that rely solely on statistical models without considering individual user circumstances can lead to incorrect or inappropriate actions. Responses from ChatGPT are randomly and statistically generated. Different responses may be generated for the same query. ChatGPT may generate incorrect responses because of its statistical characteristics, and it is a challenge to generalize its responses.
- Self-reference (AI-based) monitoring—AI systems that rely solely on themselves for evaluation, without independent oversight, can lead to a lack of accountability and decreased transparency in decision-making. As shown in Figure 4, ChatGPT uses both supervised and unsupervised learning to train the model. OpenAI did not open much information on how ChatGPT is evaluated and monitored despite the use of the self-reference approach commonly used in the community.

Besides these highly mentioned challenges, participants in our survey study also have concerns about other challenges such as missing retraceability, bias (e.g., social bias in data, bias in GPT models), the use of private data for training, overconfidence in AI generally, and copyright infringement.

4. Recommendations

Considering the significant challenges concerning ChatGPT as discussed above, this section provides recommendations to different stakeholders in ChatGPT for its responsible uses based on expert interviews in the survey study and investigations from previous work. In this section, various stakeholders involved in ChatGPT are first identified based on our survey study. Recommendations are then suggested to different stakeholders based on the qualitative analysis of our survey study.

4.1. Stakeholders in ChatGPT

There are various stakeholders involved in ChatGPT. Here are some examples

- Researchers and developers—These stakeholders are involved in developing and improving ChatGPT technologies. They may work for academic institutions, research organizations, or private companies. They may conduct research from different perspectives such as technology, law, and ethics.
- Users and consumers—These stakeholders are the end users of ChatGPT technologies. They may use ChatGPT for various purposes, such as information retrieval, language translation, and creative writing.
- Regulators and policymakers—These stakeholders are responsible for establishing legal and ethical guidelines for the development and use of ChatGPT technologies. They may work for government agencies, international organizations, or industry associations. These stakeholders collaborate closely with advocacy groups and civil society

organizations, which represent the interests of various groups affected by ChatGPT technologies, such as privacy advocates, human rights groups, and marginalized communities. Advocacy groups may lobby for policy changes or raise public awareness about the risks and benefits of these technologies.

- Ethicists and social scientists are stakeholders who focus on the ethical and social implications of ChatGPT technologies. They may study the impact of these technologies on society, culture, and human behavior. Their role is to reflect on the developments and provide guidance on how to address any ethical or social issues that arise. While they may not directly influence the development of these technologies, their work helps ensure that ChatGPT technologies are developed and used responsibly.

4.2. Recommendations for Researchers and Developers

Considering the challenges ChatGPT faces and the characteristics of researchers and developers, the recommendations for researchers and developers of ChatGPT based on the survey study are as follows.

- Do not be an algorithmic pied piper and seduce and deliberately mislead your users. Take responsibility for providing background information about bias and privacy in an active way. If possible, offer a feature to explain why a particular statement was made in ChatGPT.
- Protect the vulnerable—It is important to protect vulnerable individuals who may not fully understand the disclaimer in ChatGPT. This includes children, young people, and individuals with cognitive disabilities or lower cognitive function, who may require additional protection.
- Give reasons for answers, avoid made-up sentences unless they are explicitly requested—This command is important for ChatGPT because it emphasizes the importance of providing clear and well-justified responses to users. When the ChatGPT generates an outcome or response, it should be able to explain the reasoning behind it, rather than simply providing a result without any explanation. Providing justification for a response can help build trust and credibility with the user and can also help the user better understand the bot's thought process and decision-making. Additionally, this command emphasizes that the bot should only produce outcomes when they are deliberately requested by the user, rather than providing unsolicited responses.
- Connect ChatGPT to domain knowledge—Connecting ChatGPT models to domain-specific knowledge representations curated by a community and/or experts can greatly enhance the accuracy and relevance of the responses provided by the model. These knowledge representations can take different forms, such as taxonomies, ontologies, or knowledge graphs, and can be both human-readable and machine-readable. By including domain-specific knowledge in the training data, the model can learn to incorporate this information into its responses. This approach may require significant domain expertise to curate and annotate the training data.

Furthermore, other recommendations include aligning the research with up-to-date resources, providing reliable and retraceable resources, and taking care of copyright infringement.

4.3. Recommendations for Users

The recommendations for users of ChatGPT include the following

- Double-check information if users intend to use the result of a ChatGPT conversation as fact. This is a fundamental principle of reliable science as well as trustworthy journalism that emphasizes the importance of verifying information before publishing it. Before sharing any information on ChatGPT, make sure to check the source and ensure that it is credible and reliable. Avoid sharing information from unknown or unverified sources. Also be aware of your own biases and those of others in the chat. Double-check any information that seems too good to be true or aligns too closely with your own beliefs. When using ChatGPT, it is important to critically evaluate the

information presented, distinguishing between fact and fiction, and considering how the responses were generated.

- Do not mix facts and fiction—To use ChatGPT responsibly, it is important to distinguish between reality and fiction and to contextualize the information obtained from the platform. While it is not necessary to rely solely on factual information, it is crucial to differentiate between statements that are part of a fictional story and those that are intended to be true. For instance, scientific statements can be presumed to be accurate, whereas statements in a work of fiction may not be. Therefore, when using ChatGPT, it is essential to put the used text into the right context.
- Do not use a result of ChatGPT that you do not understand—This rule emphasizes the importance of understanding the meaning and implications of a statement before using it. This rule can also be applied to users of ChatGPT to help ensure that the messages being sent are clear and accurately reflect the intended meaning. If you come across a technical term or jargon in a message that you are not familiar with, avoid using it in your own message. Instead, take the time to research the meaning and ensure that you understand it fully before using it.
- Do not get into "waffling" and try to convince anyone by the sheer amount of text generated by a machine—This rule emphasizes the importance of being clear and concise in communication and avoiding the use of overly complex language or convoluted sentence structures. Also, do not blind others with superficiality, which can be easily generated by ChatGPT.
- Do not assign ChatGPT any responsibility to you who has not explicitly accepted it in its terms and conditions—This rule emphasizes the importance of understanding the terms and conditions of using ChatGPT. While this may be legally necessary, it also emphasizes the importance of reading and understanding the terms and conditions of any platform or service before using it.
- Ignore emotional language—Despite its human-like qualities, ChatGPT does not have emotions and feelings, but it can sometimes fake such emotions. Emotional language from ChatGPT could be ignored.

In addition, users are also recommended to always document their usage.

4.4. Recommendations for Regulators and Policymakers

The recommendations for regulators and policymakers are:

- Do not over-regulate—Finding the right balance between regulation and free use can be a challenging task for regulators. On the one hand, regulation can be necessary to protect individuals and ensure fair competition. On the other hand, excessive regulation can stifle innovation and limit the benefits of new technologies or services. To strike the right balance, regulators should consider a variety of factors, including the potential risks and benefits of the technology or service, the impact of regulation on users and businesses, and the potential for self-regulation or market-based solutions. In addition, regulators should engage with stakeholders, including users, businesses, and experts, to ensure that their approach is informed by a range of perspectives. Ultimately, the goal should be to create a regulatory environment that promotes innovation and growth while protecting the public interest.
- Thou shalt not concentrate information and communication in one place— Concentrating information and communication in one place can create imbalances of power and increase the potential for abuse. When one entity has exclusive control over information and communication channels, it can use that power to manipulate or exploit others. This can occur in a variety of contexts, including social media platforms, news media organizations, and government agencies but also with ChatGPT in the future. To prevent these imbalances of power, it is important to distribute information and communication across multiple platforms and systems, promoting competition and diversity. This can help to ensure that no single entity has too much control or

influence over the flow of information and communication and that individuals and groups have the freedom to express themselves and access the information they need.

Other recommendation examples include promoting competition to prevent a monopoly position, holding AI producers liable, prohibiting business models involving ChatGPT that do not provide access to training data, ensuring equal access, and protecting contents that AI is using to be trained.

4.5. Recommendations for Ethicists

The recommendations for ethicists are as follows.

- Understand ethical roles fully in innovative technologies—AI ethicists are concerned with human moral behavior as they design, construct, use, and treat artificially intelligent beings, as well as with the moral behavior of AI agents [34]. Ethicists need to fully understand the roles of ethics in ChatGPT in order to not only guide the ethical development of ChatGPT but also provide guidance on the ethical use of ChatGPT more effectively.
- Collaborate with experts closely from multiple disciplines—The conversation about the ethics of ChatGPT is a philosophical discussion and needs to be elevated to a sufficiently high level from different fields. For example, legal or social experts are usually good at ethical issues related to data governance, but they may not have deep knowledge of how an AI model such as an LLM is built with a large number of parameters, as AI experts are. Therefore, ethicists need to collaborate with experts that span the fields of AI, engineering, law, science, economics, ethics, philosophy, politics, and health as well as others.

Other examples of recommendations for ethicists include raising responsibility among the developers, avoiding simplistic utilitarian ethics, and applying legal requirements for primary producers who have provided the content for the AI training.

5. Discussion

LLMs such as ChatGPT can be used to fulfill typical language tasks such as summarizing text paragraphs and writing news articles, answering difficult questions, generating ideas, programming computer code, writing interesting novels, and others [8]. However, it lacks a firm moral stance, and it is suggested that users do not carelessly follow ChatGPT's advice [26,35]. We need to ensure individuals and organizations use the ChatGPT ethically, legally, and responsibly—following the human-centered AI principles, fostering the design, development, and deployment of artificial intelligence systems that prioritize the needs, values, and well-being of humans [36,37], taking into account social, ethical and legal issues, and promoting effective human-AI collaboration [38,39].

There are no widely accepted guidelines or standards for the use of ChatGPT yet. Chan [22] argued to first focus on the regulation of professional AI developers and users by government regulatory agencies and high-quality curated datasets for less harmful language model outputs. There is also a need to fill the gap between abstract ethical principles and practical applications [34]. Furthermore, public education and digital literacy are important measures to address potential intentional misuse for manipulation and unintentional harm caused by bias from language models.

This paper proposed recommendations for different stakeholders for the responsible use of ChatGPT. There are still challenges when implementing these recommendations. For example, a known limitation of ChatGPT is that it may provide answers to questions that are simply wrong. The fact-checking of answers is highly important for its responsible uses. Furthermore, a feature to explain why a particular statement was made in ChatGPT will foster user trust in the responses of ChatGPT. Therefore, the development of fact-checking and justification of responses of ChatGPT is highly suggested for the responsible use of ChatGPT. Developers must also develop systems to detect and mitigate bias, ensure user

privacy and security, differentiate responses from ChatGPT and humans, and prevent the misuse of the technology.

The proposed recommendations have important application implications. They can be implemented as guidelines for the responsible use of ChatGPT in different domains. For example, when ChatGPT is used in teaching and learning activities in schools, school authorities need to establish guidelines on how ChatGPT is used to benefit teaching and learning while minimizing its adverse ethical effects. The recommendations proposed in this paper can be used as guidelines for both school authorities as policymakers and students as users of ChatGPT. Furthermore, the recommendations can also be implemented as tools for practical use in different domains. For example, recommendations for researchers and developers suggest connecting ChatGPT to domain knowledge, which could be implemented as a tool to check whether the responses from ChatGPT align well with the latest specific domain knowledge.

In summary, ChatGPT has been demonstrating extensive applications with human-like writings and other creative works. However, users have shown various ethical concerns such as intentional misuse for manipulation and unintentional harm caused by bias, because of the data it uses and opaque models as well as human-like responses. There are still various challenges to addressing these ethical concerns. The recommendations presented in this paper will motivate stakeholders for the ethical use of ChatGPT.

6. Conclusions and Future Work

ChatGPT, launched in November 2022, has attracted significant attention due to its impressive ability to generate text that closely resembles human language for a wide range of applications. Nevertheless, the utilization of ChatGPT raises multiple ethical considerations. This paper addresses common ethical issues associated with ChatGPT, including bias, privacy, and misuse. It also discusses the main obstacles that arise when ChatGPT is utilized in different applications. The suggested pragmatic concerns for various stakeholders of ChatGPT can function as a set of instructions for individuals utilizing ChatGPT in their applications. The forthcoming research of this study will focus on conducting user studies to gather empirical information regarding the usefulness of the provided recommendations. We urge the global AI community to prioritize spreading awareness about the suggestions, which involve verifying facts and providing justifications for responses to ensure responsible usage of ChatGPT. Additionally, it is important to extend these insights to other large language models and similar technologies.

Author Contributions: Conceptualization, J.Z., A.H. and F.C.; methodology, J.Z. and H.M.; formal analysis, J.Z.; investigation, J.Z. and A.H.; writing—original draft preparation, J.Z., H.M. and A.H.; writing—review and editing, J.Z., H.M., A.H. and F.C.; visualization, J.Z.; project administration, F.C. All authors have read and agreed to the published version of the manuscript.

Funding: Parts of this work have received funding from the Austrian Science Fund (FWF), Project P-32554 (Explainable Artificial Intelligence) and by the European Union's Horizon Europe research and innovation program under No. 101079183 (BioMedAI TWINNING). This publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data are contained within the article.

Acknowledgments: We thank Chun Xiao for the literature collection for this work. We thank the anonymous reviewers for their helpful comments to further improve this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liebrenz, M.; Schleifer, R.; Buadze, A.; Bhugra, D.; Smith, A. Generating scholarly content with ChatGPT: Ethical challenges for medical publishing. *Lancet Digit. Health* **2023**, *5*, e105–e106. [CrossRef]

2. Elon University News Bureau. How ChatGPT Is Changing the Way We Use Artificial Intelligence. 2023. Available online: https://www.elon.edu/u/news/2023/02/13/how-chatgpt-is-changing-the-way-we-use-artificial-intelligence/ (accessed on 29 March 2023).
3. Pavlik, J.V. Collaborating with ChatGPT: Considering the Implications of Generative Artificial Intelligence for Journalism and Media Education. *J. Mass Commun. Educ.* **2023**, *78*, 84–93. [CrossRef]
4. Casheekar, A.; Lahiri, A.; Rath, K.; Prabhakar, K.S.; Srinivasan, K. A contemporary review on chatbots, AI-powered virtual conversational agents, ChatGPT: Applications, open challenges and future research directions. *Comput. Sci. Rev.* **2024**, *52*, 100632. [CrossRef]
5. Charfeddine, M.; Kammoun, H.M.; Hamdaoui, B.; Guizani, M. ChatGPT's Security Risks and Benefits: Offensive and Defensive Use-Cases, Mitigation Measures, and Future Implications. *IEEE Access* **2024**, *12*, 30263–30310. [CrossRef]
6. Dwivedi, Y.K.; Kshetri, N.; Hughes, L.; Slade, E.L.; Jeyaraj, A.; Kar, A.K.; Baabdullah, A.M.; Koohang, A.; Raghavan, V.; Ahuja, M.; et al. "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *Int. J. Inf. Manag.* **2023**, *71*, 102642. [CrossRef]
7. OpenAI. How Should AI Systems Behave, and Who Should Decide? 2023. Available online: https://openai.com/blog/how-should-ai-systems-behave (accessed on 29 March 2023).
8. Floridi, L.; Chiriatti, M. GPT-3: Its nature, scope, limits, and consequences. *Minds Mach.* **2020**, *30*, 681–694. [CrossRef]
9. Tung, L.ChatGPT Can Write Code. Now Researchers Say It's Good at Fixing Bugs, Too. 2023. ZDNET. Available online: https://www.zdnet.com/article/chatgpt-can-write-code-now-researchers-say-its-good-at-fixing-bugs-too/ (accessed on 29 March 2023).
10. Gandolfi, A. GPT-4 in Education: Evaluating Aptness, Reliability, and Loss of Coherence in Solving Calculus Problems and Grading Submissions. *Int. J. Artif. Intell. Educ.* **2024**, 1–31. [CrossRef]
11. Stokel-Walker, C. ChatGPT listed as author on research papers: Many scientists disapprove. *Nature* **2023**, *613*, 620–621. [CrossRef]
12. Chen, F.; Zhou, J.; Holzinger, A.; Fleischmann, K.R.; Stumpf, S. Artificial Intelligence Ethics and Trust: From Principles to Practice. *IEEE Intell. Syst.* **2023**, *38*, 5–8. [CrossRef]
13. Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; et al. Ethical and social risks of harm from language models. *arXiv* **2021**, arXiv:2112.04359.
14. Zhou, J.; Chen, F.; Berry, A.; Reed, M.; Zhang, S.; Savage, S. A survey on ethical principles of AI and implementations. In Proceedings of the 2020 IEEE Symposium Series on Computational Intelligence (SSCI), Canberra, Australia, 1–4 December 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 3010–3017.
15. Zhang, C.; Zhang, C.; Li, C.; Qiao, Y.; Zheng, S.; Dam, S.K.; Zhang, M.; Kim, J.U.; Kim, S.T.; Choi, J.; et al. One Small Step for Generative AI, One Giant Leap for AGI: A Complete Survey on ChatGPT in AIGC Era. *arXiv* **2023**, arXiv:2304.06488. [CrossRef]
16. Moulaison-Sandy, H. What Is a Person? Emerging Interpretations of AI Authorship and Attribution. *Proc. Assoc. Inf. Sci. Technol.* **2023**, *60*, 279–290. [CrossRef]
17. Abid, A.; Farooqi, M.; Zou, J. Persistent anti-muslim bias in large language models. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, Virtual, 19–21 May 2021; pp. 298–306. [CrossRef]
18. Ammu, B. GPT-3: All You Need to Know about the AI Language Model. 2024. Available online: https://www.sigmoid.com/blogs/gpt-3-all-you-need-to-know-about-the-ai-language-model/ (accessed on 22 August 2024).
19. Navigli, R.; Conia, S.; Ross, B. Biases in large language models: Origins, inventory, and discussion. *ACM J. Data Inf. Qual.* **2023**, *15*, 1–21. [CrossRef]
20. Motoki, F.; Pinho Neto, V.; Rodrigues, V. More human than human: Measuring ChatGPT political bias. *Public Choice* **2024**, *198*, 3–23. [CrossRef]
21. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Online, 3–10 March 2021; pp. 610–623.
22. Chan, A. GPT-3 and InstructGPT: Technological dystopianism, utopianism, and "Contextual" perspectives in AI ethics and industry. *AI Ethics* **2023**, *3*, 53–64. [CrossRef]
23. Baum, J.; Villasenor, J. The Politics of AI: ChatGPT and Political Bias. 2023. Available online: https://www.brookings.edu/articles/the-politics-of-ai-chatgpt-and-political-bias/ (accessed on 7 August 2023).
24. *ISO/IEC 27001*; Information Security Management Systems. Standard, International Organization for Standardization: Geneva, Switzerland, 2022.
25. *ISO/IEC 27701*; Privacy Information Management. Standard, International Organization for Standardization: Geneva, Switzerland, 2019.
26. Krügel, S.; Ostermaier, A.; Uhl, M. The moral authority of ChatGPT. *arXiv* **2023**, arXiv:2301.07098. [CrossRef]
27. Turley, J. ChatGPT Falsely Accused Me of Sexually Harassing my Students. Can We Really Trust AI? 2023. Available online: https://www.usatoday.com/story/opinion/columnist/2023/04/03/chatgpt-misinformation-bias-flaws-ai-chatbot/11571830002/ (accessed on 7 August 2023).
28. Anderljung, M.; Hazell, J. Protecting society from AI misuse: When are restrictions on capabilities warranted? *arXiv* **2023**, arXiv:2303.09377. [CrossRef]
29. Hazell, J. Spear Phishing with Large Language Models. *arXiv* **2023**, arXiv:2305.06972. [CrossRef]
30. Shevlane, T. Structured access: An emerging paradigm for safe AI deployment. *arXiv* **2022**, arXiv:2201.05159. [CrossRef]

31. StackOverflow. Temporary Policy: ChatGPT Is Banned. 2022. Available online: https://meta.stackoverflow.com/questions/421831/temporary-policy-chatgpt-is-banned/ (accessed on 4 April 2023).
32. Ibrahim, H.; Liu, F.; Asim, R.; Battu, B.; Benabderrahmane, S.; Alhafni, B.; Adnan, W.; Alhanai, T.; AlShebli, B.; Baghdadi, R.; et al. Perception, performance, and detectability of conversational artificial intelligence across 32 university courses. *Sci. Rep.* **2023**, *13*, 12187. [CrossRef]
33. Mueller, H.; Mayrhofer, M.T.; Veen, E.B.V.; Holzinger, A. The Ten Commandments of Ethical Medical AI. *IEEE Comput.* **2021**, *54*, 119–123. [CrossRef]
34. Zhou, J.; Chen, F. AI ethics: From principles to practice. *AI Soc.* **2023**, *38*, 2693–2703. [CrossRef]
35. Bulla, L.; Gangemi, A.; Mongiovì, M. Do Language Models Understand Morality? Towards a Robust Detection of Moral Content. *arXiv* **2024**, arXiv:2406.04143. [CrossRef]
36. Shneiderman, B. Human-Centered Artificial Intelligence: Reliable, Safe and Trustworthy. *Int. J. Hum.-Comput. Interact.* **2020**, *36*, 495–504. [CrossRef]
37. Shneiderman, B. *Human-Centered AI*; Oxford University Press: Oxford, UK, 2022.
38. Holzinger, A.; Kargl, M.; Kipperer, B.; Regitnig, P.; Plass, M.; Müller, H. Personas for Artificial Intelligence (AI) An Open Source Toolbox. *IEEE Access* **2022**, *10*, 23732–23747. [CrossRef]
39. Angerschmid, A.; Zhou, J.; Theuermann, K.; Chen, F.; Holzinger, A. Fairness and explanation in AI-informed decision making. *Mach. Learn. Knowl. Extr.* **2022**, *4*, 556–579. [CrossRef]