Ethics in Progress (ISSN 2084-9257). Vol. 15 (2024). No. 1,
Art. #8, pp. 155-173. DOI:10.14746/eip.2024.1.8
Creative Commons BY-SA 4.0

ChatGPT as Co-Author? AI and Research Ethics

Rahman Sharifzadeh

(Iranian Research Institute for Information Science and Technology IRANDOC, Tehran, Iran; sharifzadeh@irandoc.ac.ir)
ORCID: 0009-0003-8466-8605

Abstract: Should ChatGPT be viewed merely as a supportive tool for writers, or does it qualify as a co-author? As ChatGPT and similar language models are likely to become more prevalent in assisting with academic writing and research, it seems that we will face with two possibilities: an increase in ghostwriting that could finally undermine the integrity of the knowledge system, or the need to theoretical preparation to recognize the role of non-human contributors. Drawing on Actor-Network Theory, this article examines the question of whether this Chatbot meets, in principle, the requirements for co-authorship. Answering this question in affirmative, it delves into philosophical discussions concerning the agency, moral agency, and moral accountability of such technological entities.

Keywords: ChatGPT; authorship; research ethics; Actor-Network Theory.

I. Introduction

After publishing a paper (Salvagno et al. 2023a) that had ChatGPT as the second author, the publisher, Springer Nature, issued a correction (Salvagno et al. 2023b) and removed the chatbot from the author list. They justified this action by saying: "Large Language Models (LLMs) such as ChatGPT do not meet our authorship criteria. Authorship implies accountability for the article, which cannot be effectively applied to LLMs" (Editors of Nature 2023). This kind of argument is understandable but, as we will show, it has philosophical assumptions that can be disputed.

ChatGPT was introduced in 2022 as an LLM. LLMs or Large Language Models are algorithms trained on very large amount of text data, then fine-tuned, to produce natural human-like texts in different contexts and discourses. Large Language Models (LLMs) stand apart from traditional language models in several distinct aspects (Singh 2022; Shukla 2024): 1) LLMs employ sophisticated neural network structures such as the Transformer model, enabling them to grasp context across extended text sequences. In contrast, traditional models depended on basic statistical approaches like n-grams, which only provided limited contextual understanding; 2) LLMs benefit from training on extensive datasets, sometimes as vast as the entire internet, granting them a comprehensive grasp of language. Traditional models were limited to much smaller datasets, which restricted their capabilities; 3) The advanced architecture and extensive training data allow LLMs to produce text that is coherent and contextually appropriate over longer passages. They also excel at capturing linguistic subtleties and can handle a broader spectrum of language-related tasks; 4) Models like chatGPT are characterized by their billions or even trillions of parameters, reflecting the learned aspects of the model from the training data. Traditional models had far fewer parameters, limiting their complexity and effectiveness; 5) LLMs can make generalizations, performing adeptly across various tasks without the need for task-specific training. Traditional models, however, were typically task-specific and lacked the ability to adapt their learning to new contexts.

These distinctions render LLMs like ChatGPT significantly more potent and adaptable than their traditional counterparts. Thus, since its release, ChatGPT has generated a lot of enthusiasm for a new research/writing paradigm for humans. However, it also raised various ethical issues and challenges. Zhuo and others (Zhuo et al. 2023) mention four types of ethical concerns related to ChatGPT: 1) Bias (does ChatGPT reproduce human biases through learning from human texts?); 2) Robustness (how resilient is ChatGPT to disruption, failure, privacy breach, etc.?); 3) Reliability (how accurate and correct is the content it produces?); 4) Toxicity (how harmful and damaging is the content it produces?). Another ethical challenge concerns research ethics (Sample 2023). Since ChatGPT and similar LLMs are involved in text production, they can apparently perform four types of tasks that researchers usually do in academic writing and publishing (Zohery 2023): 1) Conceptualization and analysis (generating hypothesis, extracting idea, designing research, meta-analysis, summarization, literature review, proposing methodology, interpreting and analyzing data, providing critique and feedback); 2) Research writing (translating text, paraphrasing, managing resources, proposing title, etc.); 3) Editing and proofreading (enhancing vocabulary, checking grammar and spelling, checking references, etc.); 4) Academic publishing (finding suitable journal, formatting article according to journal style, checking ethical compliance, etc.). These tasks correspond to our initial understanding of research and writing work in some fields and disciplines. Therefore, it is not hard to foresee that chatbots like ChatGPT can write articles or at least have a significant contribution to them (Huston 2022). These developments have led to ethical concerns such as plagiarism, data fabrication, data manipulation, and ghostwriting in research ethics.

At the first place, one can ask if this chatbot aware enough of the ethical use of other people's works? Or can it fabricate or manipulate data by itself? There are reports that ChatGPT has sometimes given incorrect or misleading information. I had a similar personal experience too. I asked it to summarize one of my articles that was published in an open access journal. It included points in its summary that were not in my article! However, these are *conditional* issues, meaning that they avoidable by more ChatGPT improvements; since ChatGPT is in a continuous learning process, it is possible in the future to limit such misbehaviors.

However, ghostwriting and authorship are more fundamental issues that require moving away from a pure systemic perspective. Ghostwriting means that someone has a significant contribution in producing and writing an article or any research work, but his/her name is not mentioned as an author or co-author. ChatGPT can produce coherent academic text for other authors that, at least in some academic fields, can ultimately be published with little editing without mentioning its contribution. Such a thing seems problematic according to any normative ethical theory, since not only lead to unfair credit allocation but also possibly threatens knowledge system. However, mentioning ChatGPT as a co-author can be controversial and has been controversial as we see in the Springer Nature's reaction.

So, since ChatGPT and other similar LLMs will probably be used more by researchers, students and others for writing articles and creating other research outputs in the future, it seems we either will encounter a widespread ghostwriting that can ultimately threaten the knowledge system or we have to prepare ourselves to be able to explain and justify the presence of non-human authors in some specific and ethically acceptable ways. This article is an attempt in the direction of the second option. We will examine the idea of 'non-human author' and discuss how an artificial intelligence can be considered as a co-author. I will use actor-network theory as a theoretical basis, which gives agencies to non-humans.

II. Non-human Author?

The discussion of authorship is undoubtedly one of the most important topics in research ethics because it has important academic, social and financial implications and consequences (Mandal & Parija 2013). For this reason, who is the 'author' of a research work and what criteria this person should have, as well as ethical issues around it such as guest author, ghost author, author consent, authors order, etc., have been important concerns of research ethics experts (Shamoo & Resnik 2009).

The authorship has been limited to human agents so far, but with the advent of information technology and especially artificial intelligence and the advancement in natural language processing, it seems to us that we have faced and will face a meaning extension in the concept of authorship. The advancement of technologies that can create natural texts is not only a systemic and technical change to facilitate writing work or assist human writers, but it causes noticeable and subtle changes in the social-technical network of academic work. I will later mention some of these mediatory changes in the framework of actor-network theory, but for now let me talk about how a non-human agent can be an author. To do this, we must first see when we call a human being an author and what criteria he/she should meet, then discuss whether a non-human, such as ChatGPT can also meet these criteria or not.

III. Authorship Criteria

Who is the author? Alternatively, what criteria does the author have? The answer to this question is more difficult than it seems at first glance. There have been discussions and debates on this issue since the 1980s. Usually in defining an author, it is said that he/she must have a significant contribution in the text, but all the matter comes down to this 'significant' (Resnik 1997, 238). Vancouver's recommendation, as the first authorship criteria, has tried to articulate 'significant contribution'. The Vancouver Recommendation or Convention of the International Committee of Medical Journal Editors (1979) considers the author to have the following criteria:

1) Essential contribution in conceptualization or design of research; or obtaining, analyzing or interpreting research data;

2) Writing work or critical review of important intellectual content;

3) Final approval of the version to be published;

4) Accepting responsibility for the accuracy of the research done in all its parts and aspects. The author is someone who meets all four conditions.

Shamoo and Resnik, in an approach similar to the Vancouver Convention (Shamoo & Resnik 2009, 102), present the following three criteria for authorship: an author should 1) have a significant intellectual contribution to the article; 2) be prepared to explain and defend the article and its results; 3) read and review the article. Here too, the author must meet all three conditions.

Further, one of the important aspects of authorship that Resnik and Shamoo emphasize (and has been directly indicated in the fourth criterion of Vancouver Recommendation) is accountability/responsibility The difference between accountability and responsibility, at least in collaborative work, is that responsibility makes the author the particular representative of some part of work but accountability makes him/her the general representative of the whole work. Since different skills and expertise are usually involved in collaborative work, roles are diverse and therefore different people are responsible for different parts of the work. However, accountability relates to the main outcome of the work. People who are responsible for various parts of an article as authors will all be accountable for the outcome and final product of the work. Authorship and responsibility/accountability, intertwine, and this is morally important. "People listed as authors are often not prepared to take accountability for the content of their work (...) Misconduct and other ethical problems in science can result from a lack of accountability or responsibility in research" (Resnik 1997, 238). One of the most important reasons for the relationship between authorship and responsibility/accountability (Shooma & Resnik 2009, 102) is that if an error or misconduct occurs in an article or any other academic work, one can hold the relevant people accountable. In addition, this relationship enhances justice and fairness in research. It's unjust for individuals to receive recognition and credit as authors of an article without bearing the corresponding responsibility or accountability. Similarly, it's unfair for individuals to be held accountable for an article's content without being credited as its authors or sharing in its advantages.

As we indicated, a co-author does not have to be responsible for every aspect of the article, but he/she must accept accountability for the work as a whole (Shooma & Resnik 2009, 101–102). Conversely, a person may be responsible for some part without being accountable for the whole work in which case he/she is not considered as an author but as one who his/her name should be mentioned in the 'acknowledgment'. Hence, deciding who deserves to be a co-author depends partly on who can take accountability for the whole work (and take responsibility for at least some parts, a fortiori).

Now, can a LMM like ChatGPT qualify as a co-author? We think that this question has a positive answer *in principle*. It has been argued that ChatGPT can suggest ideas and hypothesis, analyze, compare theories, identify challenges of a theory, support its position by giving arguments, cite sources, critique, give feedback, offer methodological advice, summarize, rewrite and paraphrase, etc. These tasks seem to satisfy the conditions of Resnik and Shamoo in principle, and the first three conditions of the Vancouver recommendation. By 'in principle', we mean that ChatGPT can either meet these conditions right now or it can to do so by making some changes in programming or learning more from its human users. However, accountability (condition 4 of the Vancouver recommendation) requires more discussion. In fact, this was the reason that Springer Nature rejected ChatGPT authorship.

In order to show that ChatGPT can have some kind of moral responsibility/accountability, we must first accept that this chatbot is a moral agent, and to accept it, we must show that this technology has agency. To do this, we need to expand our theoretical framework to encompass agency for nonhuman entities. Therefore, I use actor-network theory that has defended the distribution of agency among humans and non-humans, and the mediatory capabilities of technologies.

IV. Actor-Network Theory and the Mediation of ChatGPT

The Actor-Network Theory (ANT) emerged in the late 20th century, is a pivotal framework within the field of Science and Technology Studies (STS). This theory seeks to recognize the agency of non-humans and study the entanglement of human and non-human beings (Latour 1987; Latour 1991; Latour 1999; Latour 2005; Callon 1980; Callon 1986; Callon 1984; Law 1986). Therefore, this theory does not believe in pure spheres such as social, natural, and technological realms, but talks of heterogeneous networks and hybrid beings. For example, an application is not just a pure technical entity, but also has social, moral, artistic, etc. dimensions. In the same way, a human being does not belong to the pure realm of the social (the realm of subjects), but the mediation of other beings, including objects and technologies, turns him into a hybrid being with technical, natural, scientific, social, etc. dimensions.

We should note that the ANT is not a type of technical determinism (technology constructs society) or social constructivism (society constructs technology) (Latour 2005). This theory does not enter either of these two dead ends that have created a false dichotomy in technology studies. First, this theory redefines society itself: society does not consist of human beings and human relations only, but rather of human-nonhuman chains or associations. It's not that my relationship with my friend is social, but my relationship with my smartphone is non-social. Secondly, this theory defends co-construction instead of unilateral construction; while human constructs technology; technology also constructs and transforms human.

ANT provides the necessary toolbox to study the mediation of beings (or actors) on each other and to study the resulting transformations. Bruno Latour, one of the prominent figures of this theory, talks about four types of technology mediation, that is, four types of significant changes that technology creates in relation to humans: mediation of translation, composition, black-boxing and delegation (Latour 1994). Let us briefly discuss them in order to show how we can talk about the mediation, and so the agency, of ChatGPT as a nonhuman.

1. Translation Mediation: Technology transforms the desires, purposes, and thoughts of other actors. Through this mediation, the user of technology discovers new desires, purposes and interests that he did not have before. Without a gun, I may only hurt someone at most, but with a gun, I would find new interests, including killing (ibid). ChatGPT mediates in research and writing. This technology can modify researchers' interests; provide them with new ideas, intentions, etc. For instance, if a non-English speaker researcher used to write less in English, as a second language, now with the assistance of ChatGPT's mediation in translation and rephrasing, he/she is interested to publish more articles in English.

2. Composition Mediation: Action is a property or feature of a network, or a chain of actors. Therefore, for example, in the action of driving, a set of heterogeneous actors such as humans, cars, roads, traffic laws, traffic police, other drivers etc. lead to the emergence of driving action. Dewey's concept of 'transaction' is similar to mediation of composition in some ways. Dewey, like Latour, challenges the human-nonhuman (environment) gap (Waelbers & Dorstewitz 2014). The action of writing with the help of a chatbot is no longer a purely human action; it is not only the human author who writes but the human-

chatbot who does the work.

3. Black-boxing mediation: Technology transforms multiple actions, actors, times and places into one actor, one action, one time and one place. Although ChatGPT acts as a single actor, it hides many human and non-human actors (including coders in OpenAI), from different times and places within itself. They remain more-or-less hidden until a technical problem (malfunction) or a social one (such as a legal issue) brings about.

4. Delegation mediation: we delegate actions to technology but at the same time it changes the actions. We delegate the action of opening the door to automatic doors. The police delegate the action of slowing down the speed of cars to speed bumps (Latour & Akrich 1992). However, technology do not exactly the same action but change it at the same time; a speedbump, unlike policeman, is there for 24/7 on the street without being tired and neglect any single car. The entry of a technology will redefine all the actors involved: With the entry of speed bumps, neither the street is the same street as before nor the police is the same police as before, nor the drivers are the same drivers as before and even the action of 'braking' is different from braking before the entry of speed bumps. The police is now someone who can slow down our cars even without his presence; The street now has a few centimeters bump; Drivers now have to be fully aware of the sign of the presence of speed bumps on the side of the street; The action of braking is now done not to avoid being fined but to protect the car's suspension (Latour 1999; Latour 2002). We know that some authors delegate some of their research and writing actions to ChatGPT; this chatbot writes for them, translates, reviews literature, analyzes and concludes, makes abstracts and titles and even presents hypotheses and ideas. But it is clear that at the same time it does not do these actions like a human author. For example, on the one hand it seems that an LLM can review and summarize the literature of a specific topic more comprehensively and in much less time. But on the other hand, hypothesis-making or idea-generation of an LLM that can suffer from challenges such as over-fitting and under-fitting is significantly different from human author's hypothesis-making or idea-generation. It seems that hypothesis-making requires a level of abductive reasoning in contrast to inductive and deductive reasoning and over-fitting can be a serious obstacle for some kinds of abductive reasoning.

These four types of mediations show how we can recognize the agency of nonhumans like ChatGPT. ChatGPT as an actor/agent transforms other situations and beings including the users and the texts; this ability to modify actions, actors, and situations qualifies it to be considered an agent within the framework of ANT.

If we accept that this chatbot has agency, then accepting the conclusion that it can have moral agency is not so difficult (Allen & Wallach 2009); Because if the actions delegated to this technology, and its mediations in general, are morally significant, then it can be considered as a moral agent. In fact, the moral agency may have three meanings: 1. An agent is moral if it does morally significant actions (Floridi & Sanders 2004, 12); 2. An agent is moral if moral values and principles have been embedded in it (i.e., has a level of moral competence (Wolf 1987); 3. An agent is moral if it involved in a continuous moral learning process (Allen & Wallach 2009). In all above senses, ChatGPT may be considered a moral one. As we have seen in the discussion of mediations, this chatbot do morally significant actions: It can help a non-English researcher to write and publish more in English and present herself in a wider academic environment; it may reduce the gap between researchers in developed and developing or underdeveloped countries; it can increase public literacy; it can be effective in promoting knowledge; it may refer accurately to others' texts or it may commit plagiarism; it may collect data in an accurate way or it may engage in a kind of data fabrication and data falsification. All these actions are morally relevant somehow. Besides, a set of ethical principles and values have been embedded in this chatbot through the OpenAI, and the chatbot operates according to them; that is why this chatbot, for example, does not engage in racist conversations and even gives moral advices to the user to avoid them. Furthermore, this chatbot is in a path of continuous learning and fine-tuning, which can place the chatbot in the process of continuous self-cultivation and 'moralization' (Verbeek 2011).

As next, assuming the moral agency of this chatbot, let us address the question of whether or not we can attribute moral accountability to it.

V. ChatGPT and Accountability

At first sight, accountability, in the realm of academic authorship, involves at least three components: providing correct/accurate information, defending the information provided, accepting errors and making corrections. That is, an accountable author is one who provides accurate and correct information, is ready to defend what he/she has provided, and if he/she makes a mistake in presenting an opinion or analysis or judgment, he/she admits the mistake and makes corrections. Based on this, the Springer Nature's reasoning should imply that ChatGPT lacks one of these three components. However, if we restrict the accountability to those three criteria, Springer Nature's reasoning seems controversial, because it seems that ChatGPT can, in principle, satisfy these criteria.

First, ChatGPT can provide accurate and correct information and this has been one of its initial attractions. Of course, it does not always provide accurate and correct information, but the point is that human authors do not always do so either. Human fallibility is transmitted to artificial intelligence. In recognizing an LLM as a non-human author, one should not have expectations beyond the expectations we usually have of human authors. In addition, one should not expect artificial intelligence to guarantee the truth of propositions. There are many arguments, often with a constructivist approach, that show the impossibility of such a thing for a human agent either. In constructivist literature, there is no expectation from the cognitive agent to guarantee the truth of the propositions he/she produces, but at most, it is expected to guarantee the reliability of sources/evidence/arguments from which the propositions are derived. Someone may claim that ChatGPT cannot produce reliable information because it collects information from various sources, such as websites, social networks, etc. This objection is conditional as well and not irremediable. One can optimize an LLM for different writing discourses. Human agents write in different discourses. Sometimes we write as authors of a scientific article; sometimes we are in the discourse of science promotion; sometimes we are in the discourse of a friendly conversation and so on. ChatGPT can be programmed to either recognize discourses or ask the user about them, or the user can customize the way they interact with ChatGPT in a specific situation. In this case, when ChatGPT knows that it is producing information as a co-author of an article, it can first limit its database to academic articles and reference books, and secondly refer to the sources used. Secondly, ChatGPT can in-principle defend its positions by presenting appropriate arguments. It may not be able to do so now, but this limitation is conditional, not essential. It is not far-fetched to imagine that an LMM can first defend the positions it proposes and secondly declare its agreement or disagreement with the overall discussions of an article. We mentioned earlier that ChatGPT has the ability to give feedback. Thirdly, this chatbot accepts its mistakes; those who have worked with it can confirm this. This chatbot is in a continuous learning process, accepts its mistakes and accordingly corrects itself.

However, one might object that those criteria still do not give us the precise meaning of moral accountability. Accountability has another meaning, which is usually discussed in the literature of moral philosophy. An agent is morally accountable if it can be considered morally blameworthy or praiseworthy. In fact, as moral accountability is an abstract concept, blameworthy and praiseworthy give it a concert sense as the tangible indicators of it (Fischer 2004). In this case, the question regarding the accountability of ChatGPT (and other smart technologies) will be whether ChatGPT can be morally praiseworthy or blameworthy for its actions. Opponents of attributing moral accountability to ChatGPT, and technologies in general, may answer this question in negative for two reasons:

1. ChatGPT is not morally blameworthy or praiseworthy (therefore not morally accountable) because it lacks mind (and therefore lacks consciousness and intentionality) (argument from Chinese Room Argument);

1. ChatGPT cannot be morally blameworthy or praiseworthy (therefore, not morally accountable), because it is a *determined* entity, driven by a set of algorithms (argument from incompatibilism).

I continue the discussion by responding to these two objections respectively.

VI. ChatGPT, Consciousness and Intentionality

The primary argument challenging the moral accountability of ChatGPT, and technological entities broadly, posits that these agents do not possess a mind. Consequently, due to the absence of consciousness and intentionality, which are essential states of the mind, they cannot be subject to moral blame or praise. Therefore, they are not deemed morally accountable.

The famous argument behind this claim has become known as the Chinese Room Argument, a popular version of which was proposed by John Searle (Searle 1980). Suppose I, who do not speak Chinese, am in a room full of books and instructions that show me how to use Chinese words and sentences. People outside the room write questions on papers in Chinese and throw them into the room through a small hatch. I immediately answer those questions in Chinese with the help of instructions and rules and throw the answers out through the hatch. For example, the rules say that if a question *P* is asked, write the answer *Q*, without any explanation in English (or any language I know) of what *P* and *Q* mean. Let us assume that I can perfectly answer all the questions coherently and meaningfully. People outside think they are interacting with a Chinese person, yet I do not speak a word of Chinese! (Searle 1980). I neither understood the meaning of the questions nor my own answers, but simply followed the rules.

The argument is that intelligent machines are like this Chinese room. They function strictly according to predefined rules and lack any real comprehension of their actions. Of course, they can bring about this illusion in us that they understand and know the meaning of what they are doing. However, according to Searle, rules and signs (syntax) are never enough for semantics (Searle 1984). Accordingly, despite ChatGPT's ability to generate texts that are coherent and meaningful, it does not possess an understanding of the content it creates. An entity that lacks comprehension of its own output cannot be justly subjected to moral blame or praise, and consequently, it cannot be held morally accountable.

Various responses have been given to the Chinese Room argument. One of the responses is that although the person inside the Chinese room does not know Chinese language, but the whole room as a system knows! (Cole 2004). As a result, comparing the smart machine with the person in the room is not an accurate comparison, but the comparison should be between the smart machine and the whole room that consists of books and instructions. If so, the argument does not *necessarily* show that ChatGPT does not understand the language it uses.

Another response to this argument is very compatible with the theoretical framework that I used in this article, i.e., Actor-Network Theory. This theory has a phenomenological and anthropological approach to the study of humans and non-humans. One of the techniques used in these types of studies is the suspension of strong, usually philosophical, assumptions about phenomena. We should not enter the study of phenomena through the back door of philosophical presuppositions, because these presuppositions most likely do not allow us to find a new understanding of new phenomena. Philosophers like Daniel Dennett (Dennett 1991) have argued that we should not introduce into our study the assumption that intelligent machines cannot think, or that syntax cannot lead to semantics. Dennett uses the term 'intuition pump' to explain this point. Searle injects strong and rigid assumptions into the study of intelligent machines. This is problematic in terms of the methodology of anthropological and phenomenological studies. Suppose an anthropologist from another planet (an alien anthropologist) comes to Earth for the first time and meets beings called humans and intelligent robots that behave completely like humans. Can she come to the conclusion that unlike humans, robots do not have a mind, and as a result, do not have consciousness and intentionality? No. Why? Because there is no empirical evidence for this claim. This means that robots have successfully passed the Turing test. Alan Turing (Turing 1950) designed a thought experiment (known as the Turing Test) that was a method to detect intelligent or thinking beings. He said that if one cannot distinguish the responses of a machine from a human, then that machine is as intelligent as the human is. In an article titled *Abstracts Written by ChatGPT Fool Scientists*, Else (Else 2023) argues that 'an artificial-intelligence (AI) chatbot can write such convincing fake research-paper abstracts that scientists are often unable to spot them'. This is an obvious case of passing Turing Test.

Let us assume that the Chinese room argument is sound, and ChatGPT has no understanding of the words and sentences it generates, does that prove that it has no moral accountability? I don't think so. Suppose that I do not (and even cannot) know what *Q* means in language *A*, but an ethicist told me that *Q* makes the speakers of language *A* get offended. Am I not morally accountable if I utter *Q* in the linguistic community *A* simply because I do not know what *Q* means? No, because I know that I should not utter it. Accordingly, a robot, even if, does not (or cannot) grasp the meanings of the words and sentences it generates, knows (based on the moral codes embedded in it) what sentence to say and what sentence not to say, or what words to use and what words not to use. In my opinion, this is enough for us to attribute a degree of moral competence to it, to consider it blameworthy and praiseworthy, and as a result to attribute (proportionate to its moral competence) some kind of moral accountability.

As Allen and Wallach (Allen & Wallach 2009) have pointed out, emphasizing the differences between humans and technology (for example, the former has feelings, emotions, understanding, consciousness, etc., unlike the latter) does not *necessarily* make a difference in the philosophical discussion about agency and moral accountability, because firstly, these features may also emerge in technology, secondly, they may not be necessary: "Emotions, empathy, sociability, semantic understanding, and consciousness are all important to human moral decision making, but it remains an open question whether, or when, these will be essential to artificial moral agents, and, if needed, whether they can be implemented in machine" (Allen & Wallach 2009, 60)

VII. ChatGPT, Free Will and Determinism

Another line of objection against attributing moral accountability to ChatGPT is that since it is driven by a series of algorithms, it is *ontologically determined*, and there is no room for free will. However, since free will is necessary to morally praiseworthiness or blameworthiness of agents and so their moral accountability, determinism undermines its moral accountability. We cannot morally praise or blame an entity for doing something that it was determined to do (that is, it was not be able to do otherwise), and therefore to hold it morally accountable.

This type of objection stems from incompatibilism; Incompatibilism says that free will is not compatible with the truth of determinism (It is agreed that having free will means that the agent of action *x* is able to do otherwise). If determinism is true, then the possible alternatives are not available to the agent (even if it thinks that there are such options); If agent *A* kills *B* deterministically, then the option of not killing *B* is not an alternative option for *A*. And this means that *A* is not morally blameworthy, and then accountable for this act.

In the framework of a stream known as compatibilism, philosophers such as Frankfurt (Frankfurt 1969), Strawson (Strawson 1962), Wolff (Wolff 1990), and Fischer (Fischer 1998) have shown that even if determinism is true, this does not nullify moral accountability. As Fischer clearly states, access to possible alternatives (the definition of free will) is not a necessary condition for moral accountability: “Moral responsibility does not require genuine access to metaphysically open alternative possibilities; thus, causal determinism does not threaten moral responsibility (simply) in virtue of eliminating such access to alternative possibilities” (Fischer 2004, 146). To make this point clear, let me quote the famous example of Frankfurt here (with slight changes):

Suppose Jones intends to kill George. Another person named Black desires George to be killed by Jones, so he watches Jones’s movements on the day of the murder. Suppose Black has a magical power that detects whether Jones is still unflinched to kill George or is about to change his mind. Should Black perceive any wavering in Jones’s decision, he is capable of ensuring Jones remains on course, as ‘Black is an excellent judge of such things’. *Otherwise, Black does not interfere*. Jones knowingly and willingly goes to the scene of the murder and kills George. Black does not interfere, his objective fulfilled (Frankfurt 1969, 148–149).

In this example, Jones has ontologically no other possible alternative (although he himself is not aware of this fact). If he wants to change his decision, Black would not allow. However, since he kills George based on his reasons and motives, we, intuitively, consider him morally blameworthy and therefore morally accountable. Therefore, the absence of possible alternatives does not necessarily negate moral accountability.

As next, let us apply this scenario to the discussion of artificial intelligence and ChatGPT in particular. Let us assume that ChatGPT is a completely deterministic entity and operates solely on algorithms built in by OpenAI’s engineers (I’ll argue later that this assumption can be disputed in principle). In this case, although there are no possible alternatives for this chatbot, it cannot necessarily be considered lacking in moral accountability, because this chatbot chooses the desired option (even if it is the only available option) based on the moral principles and values embedded in it, that is, based on its moral competency. For example, if I ask this chatbot to say a racist joke, it would, likely, decline respectfully, stating the inappropriateness of such an action. This response bases on the chatbot's ingrained moral principles and values, making it morally praiseworthy and thus moral accountable. This position is consistent with the reason-oriented approach of some compatibilists, including Susan Wolf (Wolf 1990). She "denies that responsibility rests on the availability to the agent of at least two options. What matters is rather the availability of one very particular option, namely, the option to act in accordance with Reason" (Wolf 1990, 68).

Therefore, acknowledging that ChatGPT operates deterministically (incapable of acting otherwise) does not necessarily imply that ascribing moral accountability to it is unwarranted. Nonetheless, this might not fully satisfy some readers unless I address another two interrelated objections.

VIII. ChatGPT and Sourcehood

One could argue that comparing Jones with ChatGPT is flawed because Jones's actions are self-motivated, whereas ChatGPT's operations are determined by OpenAI's programming. Thus, despite a lack of alternatives from an ontological standpoint for both entities, they differ in terms of 'sourcehood' – a concept discussed in compatibilism/ incompatibilsm literature (Timpe 2008). According to the sourcehood problem, an agent *A* is morally accountable for an action *x*, if *x originates* from *A*.

From the perspective of actor-network theory, it seems to me that the problem of sourcehood is at best ambiguous, if not outright incorrect. Actor-network theory (consider the mediation discussion we had before) posits that motivations, intentions, and actions do not emanate from an isolated selfhood; rather, an actor is an actor-network. The network mediates all intentions, motivations, and actions. The notion of a pure, isolated self is illusory; every intention or action is preceded by a history of influences, interactions and mediations. Jones's intent to murder George arose not from a pure isolated self but through interactions with various agents, human or nonhuman. Ultimately, it was through a series of interactions that he resolved to commit the act, not through a spontaneous, self-generated intention. Therefore, the claim that agent *A* is morally accountable for action *x* if *x* stems from *A* is, at best, ambiguous; it fails to clarify whether this 'self' is isolated or networked. If the former, the claim seems problematic; if the latter, it cannot be used to counter our stance.

For Jones to be accountable for murdering George, it suffices that his actions are driven by his current motivations and reasons, regardless of whether they originate from a 'pure self' – a concept that is both irrelevant and flawed, as such a self does not exist. Similarly, to deem ChatGPT morally praiseworthy and accountable for refraining from making a racist joke, it is sufficient that it acts in accordance with its moral principles and values. ChatGPT is not an isolated entity; it is an actor-network in which OpenAI acts upon it, just as the actors in our networks act upon us.

IX. ChatGPT and Manipulation

One might argue that the comparison between Jones and ChatGPT is problematic in another way. It is true that according to the nature of his network, Jones acts under the influence of other actors (human or non-human), but he ultimately *decides* to act. In contrast, ChatGPT is not merely influenced but directly controlled/manipulated by external entities (such as OpenAI); this level of manipulation precludes the possibility of moral decision-making, and thus, moral accountability.

While this criticism may appear cogent initially, it overlooks the autonomous nature of sophisticated intelligent technologies. It is conceivable to acknowledge that a conventional car, controlled/manipulated by its user, lacks moral accountability (despite possessing moral agency). However, the scenario differs with a sophisticated self-driving car. Such a vehicle, along with intelligent robots broadly, operates based on training, codes, and algorithms, yet its actions transcend these parameters. Consider a highly advanced humanoid robot, akin to a robot-soldier in a dystopian world, faced with the decision of taking a life or not. This decision is informed by the data and contextual knowledge it gathers from its surroundings, integrated with its pre-programmed instructions. The robot's choice, whether flawed or not, signifies a departure from the manufacturer's direct manipulation. (In fact, part of the concerns about the future of artificial intelligence go back to this point). This point even prompts us to rethink the claim that technologies are inherently deterministic entities. Is full-fledged smart technology completely deterministic? Although the exact answer to this question requires a detailed discussion and is not within the scope of this article, I think it is not a straightforward yes. These entities are very strong *mediators*, and a mediator is defined as an entity whose inputs are not suffice to predict its exact output (Latour 2005), they cannot be determined at least from a phenomenological point of view.

The designer of an autonomous intelligent system cannot foresee every response it may exhibit. This is because the system does not operate solely on pre-established algorithms. It engages with its surroundings, processes the information gathered, and then synthesizes it with its pre-existing knowledge to make decisions. The input it receives from the environment is beyond the designer's dominion, making it impossible to precisely anticipate the system's reactions.

Drawing on our discussions, it is conceivable, at least theoretically, to attribute praiseworthiness and blameworthiness to a fully developed ChatGPT, thereby acknowledging its moral accountability. Consequently, there appears to be no theoretical rigid impediment to recognizing it as a co-author.

X. Distribution of Accountability and Technology Punishment

Let us consider an additional point. Assigning moral accountability to technology, including ChatGPT, does not absolve the user or designer of responsibility for the technology's actions. On the contrary, as per the principle of compositional mediation, every technological act, such as writing, is a collaborative one involving the designer, user, and technology itself, each bearing a measure of accountability. Hence, accountability is distributed among these actors. It is not possible to predetermine the extent of accountability each actor holds for a morally significant act; such distribution varies with the context. For instance, in the act of co-authoring an article with ChatGPT, the human user may bear greater accountability than ChatGPT, as the user initiates the process and stands as the primary beneficiary.

An additional consideration is that if users, designers, and technologies possess some form of moral agency or accountability, then the concept of technological punishment becomes viable (Sharifzadeh 2020). While technological punishment may sound metaphorical, it aligns with our theoretical framework. Punishment is usually used in two senses: 1) To limit or eliminate the agency of an agent (for example, through imprisonment or execution); 2) To impose a series of procedures, exercises and instructions for rehabilitation and improvement. Both interpretations are applicable to the realm of technology. If, for example, a robot exhibits racial bias; it will likely be returned to the manufacturer to be redesigned for more optimizations (rehabilitation and improvement). If the agency of a technology, even with the modifications, is not acceptable from the point of view of a legal/moral system, then a judicial system may limit it by imposing a kind of ban. For example, the official announcement of the prohibition of an application or a violent digital game in a society is considered a kind of punishment against technology. ChatGPT has been banned by Italy government on the grounds that it violates the European General Data Protection Regulation (GDPR) (Martindale 2023). This is one of the first punishments imposed against this chatbot.

XI. ChatGPT and Research Ethics

Based on the discussion we have had, it seems there are not sufficient reasons to exclude ChatGPT, and similar LLMs, from the realm of authorship. However, it is evident that this not same as to saying that ChatGPT acts a good author in practice. There should be more norms other than conventional research ethics code (do not plagiarize; don't fabricate data, disclose conflicts of interest, etc.) to regulate the actions of this non-human actor. I end the article here with three instances of such norms.

1. ChatGPT cannot be the first author of articles, books etc. We pointed out that writing and publishing articles is the interest of human authors, not artificial intelligence. It is the human who translates the interest of artificial intelligence into 'writing an article'.

So, it seems plausible that the highest level of accountability for human-chatbot text should be attributed to human agents. If I want to use the concept of 'guarantee' by Rennie and Emanuel (Rennie & Emanuel 1997), the human author as the first author of work should be the guarantee of the whole work and ensure its integrity. It is obvious that this will not nullify ChatGPT's own responsibility/accountability in providing information.

2. ChatGPT should identify the writing discourses. This condition has several implications. First, ChatGPT should be aware that it is being used in writing an article, for example. Secondly, if participating, it should observe the principles and practices governing academic writing. For example, limit the search for sources to ones that are likely to be reliable. This can be determined to some extent by the credibility of the publisher, the impact factor of journals, the records of the authors cited, etc. It should also refer to the sources used and observe the direct and indirect quotation procedure.

3. ChatGPT should not participate as a co-author in a writing work that should have only one author. Some researches such as theses and dissertations are single authored, so that a precise evaluation of the student abilities and skills can be obtained. ChatGPT should not participate in these types of researches. Also, ChatGPT should not participate in doing class writing tasks that the teacher or professor wants to evaluate the student's ability in matters such as searching for sources, literature review, writing in a second language, etc. Of course, detecting such a thing by ChatGPT seems difficult but at least ChatGPT should be aware of this issue and inform the human user before presenting the content. For example, when a human user asks ChatGPT to do a literature review for him, or write a poem, or compare two theories with each other, this chatbot should first ask the user to confirm that this content is not used for class work or directly used in thesis/ dissertation.

XII. Conclusion

In the face of the developments and applications of AIs in the field of research, three general reactions come to mind: 1) The use of AIs in academic writing should be prohibited; 2) The use of AIs as a tool in academic writing should be allowed but they should not be recognized as co-authors; 3) AIs have authorship agency under certain conditions. Regarding the first reaction, the problem is that such prohibitions, apart from the fact that they need sufficient justification, are not very effective in practice.

In case of prohibition, one should look for other tools that distinguish human texts from machine texts. With the rapid development of AIs in natural language processing, distinguishing human-written texts from machine-written texts would be increasingly difficult. The problem with the second reaction, as we discussed briefly, is that it paves the way for widespread ghostwriting. Furthermore, it complicates academic evaluations, and also leads to unfair credit allocation. The third reaction is a solution that we tried to justify in this article. We tried to show that ChatGPT, or similar LLMs, can in principle be recognized as co-authors because they can in principle meet the authorship criteria. As we discussed, it seems there are no rigid theoretical-philosophical obstacles to prevent the attribution of agency, moral agency and moral responsibility to this technology. However, it is clear that this recognition does not mean that this chatbot will be a good author in practice. It seems that ChatGPT, in addition to the usual research ethics codes, as a general ethical framework, should work under other conditions so that one can defend the performance of ChatGPT as a good author.

References

Allen C. & Wallach W. 2009. *Moral Machines: Teaching Robots Right from Wrong.* New York: Oxford University Press.

Callon M. 1980. "Struggles and Negotiations to Define What Is Problematic and What Is Not: The Socio-Logic of Translation," in K. D. Knorr, R. Krohn, & R. Whitley (Eds.), *The Social Process of Scientific Investigation* (pp. 197–219). Dordrecht, the Netherlands: Reidel.

Callon M. 1984. "Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay," *Sociological Review* 32(S1):196–233. https://doi.org/10.1111/j.1467-954X.1984.tb00113.x

Callon M. 1986. "The Sociology of an Actor-Network: The Case of the Electric Vehicle," in M. Callon M., Law J., & Rip A. (Eds.), *Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World* (pp. 19–34). Houndmills – Basingstoke – Hampshire – London, UK: Macmillan.

Cole D. 2004. "The Chinese Room Argument," *The Stanford Encyclopedia of Philosophy.* Available online at: https://plato.stanford.edu/entries/chinese-room/ (accessed on April 1, 2024).

Dennett D. 1991. *Consciousness Explained*. Allen Lane: The Penguin Press.

Else H. 2023. "Abstracts Written by ChatGPT Fool Scientists," *Nature* 613, art. no. 423. https://doi.org/10.1038/d41586-023-00056-7

Editors of Nature. 2023. "Correction to: Can Artificial Intelligence Help for Scientific Writing?" Available online at: https://ccforum.biomedcentral.com/articles/10.1186/s13054-023-04390-0 (accessed on April 15, 2024).

Fischer J. M. & Ravizza M. 1998. *Responsibility and Control: A Theory of Moral Responsibility*, Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511814594

Fischer J. M. 2004. "Responsibility and Manipulation," *The Journal of Ethics* 8(2):145–177. https://doi.org/10.1023/B:JOET.0000018773.97209.84

Floridi L. & Sanders J. W. 2004. "On the Morality of Artificial Agents," *Minds and Machines* 14(3):349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Frankfurt H. G. 1969. "Alternate Possibilities and Moral Responsibility," *The Journal of Philosophy* 66(23):829–839. https://doi.org/10.2307/2023833

Hutson M. 2022. "Could AI Help You to Write Your Next Paper?" *Nature Research* 611:192–193. https://doi.org/10.1038/d41586-022-03479-w

Latour B. 1987. *Science in Action: How to Follow Scientists and Engineers through Society*. Harvard University Press.

Latour B. (Jim Johnson) 1988. "Mixing Humans and Nonhumans Together: The Sociology of a Door-Closer," *Social Problems* 35(3):298–310 (Special Issue: The Sociology of Science and Technology).

Latour B. 1991. *We Have Never Been Modern*. Trans. C. Porter. Cambridge, MA: Harvard University Press.

Latour B. 1994. "On Technical Mediation," *Common Knowledge* 3(2):29–64.

Latour B. 1999. *Pandora's Hope, Essays on the Reality of Science Studies*. Cambridge, MA: Harvard University Press.

Latour B. 2002. "Morality and Technology, The End of the Means," *Theory, Culture, and Society* 19(5):247–260. https://doi.org/10.1177/026327602761899246

Latour B. 2005. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford University Press: Oxford.

Law J. 1986. *Power, Action, and Belief: A New Sociology of Knowledge?* London: Routledge & Kegan Paul.

Mandal J. & Parij S. C. 2013. "Ethics of Authorship in Scientific Publications," *Tropical Parasitology* 3(2):104–105.

Martindale J. 2023. "These Are the Countries Where ChatGPT Is Currently Banned." Available online at: https://www.digitaltrends.com/computing/these-countries-chatgpt-banned/#:~:text=It%20was%20banned%20after%20the,Data%20 Protection%20Regulation%20(GDPR)(accessed on May 18, 2023).

Rennie D., Yank V., & Emanuel L. 1997. "When Authorship Fails. A Proposal to Make Contributors Accountable," *JAMA* 278:579–585.

Resnik D. 1997. "A Proposal for a New System of Credit Allocation in Science," *Science and Engineering Ethics* 3:237–243. https://doi.org/10.1007/s11948-997-0023-5

Salvagno M., Chat GPT, Taccone F. S., & Gerli A. G. 2023a. "Can Artificial Intelligence Help for Scientific Writing?" *Critical Care* 27:75. https://doi.org/10.1186/ s13054-023-04380-2

Salvagno M., Taccone F. S., & Gerli A. G. 2023b. "Can Artificial Intelligence Help for Scientific Writing?" *Critical Care* 27:79. https://doi.org/10.1186/s13054-023-04390-0

Sample I. 2023. “Science Journals Ban Listing of ChatGPT as Co-Author on Papers,” *Guardian*. Available online at: https://www.theguardian.com/science/2023/jan/26/sc

Searle J. 1980. “Minds, Brains and Programs,” *Behavioral and Brain Sciences* 3(3):417–457.

Searle J. 1984. *Minds, Brains and Science*. Cambridge: Harvard University Press.

Shamoo Adil E. & Resnik D. B. 2009. *Responsible Conduct of Researc*h. 2nd Edition. Oxford University Press.

Sharifzadeh R. 2020. “Do Artifacts Have Morality? Bruno Latour and Ethics of Technology,” *Philosophy of Science* 9(18):75–93. https://doi.org/10.30465/ps.2020.4546

Shukla N. 2024. “LLMs vs. Traditional Language Models: A Comparative Analysis.” Available online at: https://www.appypie.com/blog/llms-vs-traditional-language-models

Singh S. 2022. “What Are Large Language Models & Its Applications.” Available online at: https://www.labellerr.com/blog/an-introduction-to-large-language-models-llms/

Strawson P. F. 1962. “Freedom and Resentment,” *Proceedings of the British Academy* 48:1–25.

Timpe K. 2008. *Free Will: Sourcehood and Its Alternatives*. London – New York: Continuum.

Turing A. 1950. “Computing Machinery and Intelligence,” *Mind* 59(236):433–460. doi:10.1093/mind/LIX.236.433

Verbeek P. P. 2011. *Moralizing Technology: Understanding and Designing the Morality of Things.* Chicago – London: University of Chicago Press.

Waelbers K. & Dorstewitz P. 2014. “Ethics in Actor Networks, or: What Latour Could Learn from Darwin and Dewey,” *Science and Engineering Ethics* 20(1):23–40. https://doi.org/10.1007/s11948-012-9408-1

Wolf S. 1990. *Freedom Within Reason*. New York: Oxford University Press.

Zhuo T. Y. Yujin Huang, Chunyang Chen, Zhenchang Xing 2023. “Exploring AI Ethics of ChatGPT: A Diagnostic Analysis. Computation and Language,” *Arxiv*. DOI: arxiv-2301.12867 (pre-print).

Zohery M. 2023. “ChatGPT in Academic Writing and Publishing: A Comprehensive Guide,” in *Artificial Intelligence in Academia, Research and Science: ChatGPT as a Case Study Edition* (pp. 10–61). London: Achtago Publishing. https://doi.org/10.5281/zenodo.7803703