P&D PHILOSOPHY & DIGITALITY
AN OPEN ACCESS, PEER-REVIEWED JOURNAL

Why a careless use of AI tools may contribute to an epistemological crisis

Anna Strasser

Abstract: With the rise of generative AI technologies, AI tools such as ChatGPT play a prevalent role in our world. This poses new opportunities and risks. Particularly in the field of education, the use of AI tools raises a number of new questions, for example, the challenge of proving human authorship, or whether we can continue to trust electronically distributed texts. This then puts into question whether the widespread and careless use of AI tools is contributing to an epistemological crisis. Based on an analysis of the inherent unreliability of generative AI and the increasing indistinguishability between human-written and machine-generated texts, this paper outlines the impact of generative AI on education and provides suggestions for the responsible use of AI tools. Investigating their use as legitimate tools, it also addresses the question as to whether the use of such tools could be counterproductive because it could possibly result in a deskilling effect.

Keywords: responsible use of AI tools; unreliability of machine-generated texts; indistinguishability; deskilling; epistemological crisis

1 Since ChatGPT's release in 2022, an emotionally charged debate has been occupying scientific communities and the general public, discussing the use of LLMs in education, science, and elsewhere.This debate is characterized by widespread enthusiasm about the fact that AI tools can enable people to solve tasks they would not be able to solve without the tools or for which they would have to invest much more time-consuming work. In contrast, issues concerning the reliability and impracticability of distinguishing between human-written and machine-generated text are also intensely discussed. This concerns, for example, the well-known problem of hallucinations and the fact that we are in many areas increasingly exposed to machine-generated texts that can easily be confused with human-written texts.

2 Within the educational domain, attitudes towards generative AI tools range from radical rejec-

Contact: Anna Strasser <annakatharinastrasser@gmail.com> ORCiD: 0000-0002-3547-9790
https://doi.org/10.18716/pd.v2i1.11662

tions, assuming that any involvement of them constitutes deception/fraud, to calls that they should definitely be integrated into future curricula because they can serve as a legitimate aid (Frye 2022; Hutson 2022; 2023; Lee et al. 2024; Tlili et al. 2023).There is no doubt that this new technology offers interesting and useful applications. The use of AI tools can be a source of inspiration and can help to sift through an infinite space of data and its patterns, and it can help to save time. However, the advantages of a new technology should not prevent us from analyzing its disadvantages. A risk evaluation of the consequences of this new technology is essential, especially when it concerns people at the beginning of their educational journey. The positive use cases are not the focus of this paper. Instead, I will focus on an analysis of critical use cases of AI tools and present considerations on how the responsible use of these tools may be conceivable despite their dangers.

3 One issue that is extensively discussed in the educational domain concerns the consequences that come with the increasing indistinguishability between machine-generated texts and human-authored texts. Neither humans nor detection software can prove beyond doubt whether submitted texts were actually created by machines or by humans. This has the potential to erode established relationships of trust. It is already a major concern for teachers that they are not able to distinguish their students' self-written essays from machine-generated ones (Huang 2023; Marche 2022; Peritz 2022; Sparrow 2022). Even though there is a consensus that texts not written by the students themselves, i.e., machine-generated texts, should not count as an examination performance, the handling of machine-generated texts poses a major problem because we cannot tell them apart with certainty. Therefore, some take the position that this will lead to the demise of conventional educational assessment methods because ChatGPT and other LLM-based AI tools are seen as a serious threat to the credibility of short-form essays as an assessment method (Herman 2022; Yeadon et al. 2022).

4 A further topic concerns the question of which AI aids are classified as legitimate and how their use is to be labeled. The use of spelling software seems uncritical here, but text-generating AI tools raise the question of whether they may undermine the learning of writing skills as well as the ability to understand argumentation structures, which might eventually even contribute to the deskilling of critical thinking. One may wonder to what extent advanced learners who regularly use AI tools are likely to unlearn certain cognitive skills needed to summarize complex texts, develop a line of argument, or write an essay. If AI tools are used in the earliest stages of education, attention should be paid to the extent to which the use of AI tools may contribute to certain cognitive abilities not being acquired at all. An additional factor motivating a critical attitude towards the use of an AI text generator to create academic texts or school essays relates to considerations concerning the extent to which LLMs can contribute to a new form of plagiarism in which the intellectual property of others is misused because LLMs are trained on the work of others (Dehouche 2021; Schwitzgebel, Schwitzgebel, and Strasser 2023).

5 Although it can be helpful in certain learning situations to use an AI tool to create a summary of a difficult text, it is important to realize that AI tools are not as reliable as the majority of users

might think. I will argue that, especially when AI tools are used for knowledge acquisition, it is of immense importance that users know how to verify the results of the AI tool. In contrast to the impressive performance in many tasks, there is still a general problem with reliability with all tools based on generative AI. LLMs are not trained to consider the truth of their linguistic output. They deliver tenable and untenable hypotheses. Outputs can be deeply mistaken or simply wrong. Here, the increasing indistinguishability can have critical additional effects as it becomes more and more difficult to recognize whether, for example, sources found on the internet consist of machine-generated text and whether they have been written or at least checked by humans with expertise. This poses major problems for both teachers and learners, as it is no longer obvious which websites found on the internet can be trusted. Of course, even before the rise of generative AI technology, it was a critical question as to which texts on the internet could be trusted. Already back then, there was a lot of misleading information on the Internet. However, it is foreseeable that the proportion of unreliable machine-generated text, on a quantitative level alone, will massively exacerbate this situation.

6 In this essay, I shall elaborate on three questions. First, I will examine cases in which the inherent unreliability of tools based on generative AI may make the use of such AI tools for text production and knowledge acquisition highly questionable. Thereby, I will elaborate on the necessity to develop verification skills when using machine-generated text. In a second step, I shall address the effects of the increasing indistinguishability between human-created and machine-generated text. Besides pointing to the impossibility of proofing the use of AI tools, I shall argue that the increasing indistinguishability together with the inherent unreliability of LLMs output will contribute to a situation in which even using the internet and other electronically distributed texts to gain knowledge becomes rather difficult and may thereby contribute to an epistemological crisis. Finally, I will examine the extent to which the careless use of AI tools can lead to acquired abilities being unlearned or even certain abilities no longer being learned, which leads to a critical dependency on these tools.

Unreliability of LLM's outputs

7 The hype around LLMs has been fueled by headlines claiming that these language models are capable of delivering mind-blowing performance and are able to outperform humans in several domains. There is no question that generative AI technology can produce results that far surpass what a learner, a non-expert in a particular domain, can produce. One might think here, for example, of translation (Hatcher and Yu 2018) or the possibility of creating code using an LLM. Even experts are rightly impressed by the performance of generative AI tools. For example, without AlphaFold, the successes in protein structure prediction were not conceivable (Jumper et al. 2021). Without belittling the success story of AI tools, one should be aware that the successes are rooted in the involvement of experts who were able to verify the results of their AI

tools.

8 Furthermore, it should also be noted that not only the amazing output but also the speed of the produced result plays a role in generating an enthusiastic attitude towards such tools. For experts, creating a similar result would often take significantly more time. This especially gains relevance considering that a human lifetime might not be sufficient to search a space of possibilities manually; this was, for example, the case when exploring protein folding without AlphaFold. All of this has led to discussions in many areas of life about whether AI tools should be considered as support for humans since they can relieve humans of time-consuming work and make results possible, which could not have been reached without such tools. But it has also led to fears that AI could replace humans and take away people's jobs. This was, for example, illustrated by the strike of the Writers Guild of America; screenwriters were concerned about the prospect that generative AI technologies could replace them ("2023 Writers Guild of America Strike" 2024; Barnes and Koblin 2023).

9 Despite the critical attitudes toward the societal impacts of AI technologies, it seems as if the reports about the impressive achievements of LLMs are never-ending. There is ongoing news about them passing tests that are established to examine human intelligence. Such reports contribute to the impression that AI tools are getting better and better while ignoring the inherent problems regarding the reliability of this technology that is based on deep neural networks. For example, OpenAI reported in its preprint on GPT-4 that this language model not only performed very well on the *Uniform Bar Exam*, the *Graduate Record Exam*, and several high school *Advanced Placement Tests* but also on several benchmarks that purport to assess language comprehension, reasoning, programming skills, and other abilities (OpenAI et al. 2024). Given that the reports about LLMs getting better and better are so widespread, results of studies that show how easily benchmark test accuracy can be compromised (Mitchell 2023), or papers that collect documented erroneous results from LLMs (Marcus and Davis 2020; 2023), seem to get lost in public attention.

10 For a variety of reasons, one should be cautious in interpreting the enthusiastic reports of the performance as evidence of GPT-4's human-level intelligence. In contrast, fostering the awareness of the inherent problems with reliability is of importance. Especially in educational settings, learners are vulnerable because they often do not have the expertise to verify the results of their AI tools and often over estimate the reliability of AI tools. Instead of being overly impressed by the latest news on benchmark results, one should consider that benchmarks can unfavorably test something that is already included in the training data and thereby give the tested model the opportunity to use shortcuts to solve the test tasks. Such shortcuts will not prove successful for comparable tasks not included in the training data. Furthermore, it is not always clear whether the formulated tasks really necessarily require the use of the supposedly tested cognitive abilities.

11 Melanie Mitchell argued for a skeptical attitude by elaborating on three problems associated with benchmark tests, namely, the problem of data contamination, the problem of robustness,

and the problem of flawed benchmarks (Mitchell 2023). Applying standardized tests to humans, in most cases, one can be relatively sure that they have not yet seen the specific test questions. Consequently, one can exclude that their performance could be based on pure memorization of the answers to certain questions. However, this is far from certain with LLMs, which are trained on a huge amount of data. There, it is more likely that they have already 'seen' specific test questions during their training phase. This is called data contamination, and the effects of having 'seen' test tasks in the training phase are described as overfitting. By stress-testing LLMs, it becomes visible that LLMs exhibit distinct accuracy within a set of comparable test questions. Studies could show that LLMs perform worse on rare tasks than on common ones (McCoy et al. 2024) and support thereby the claim that LLMs demonstrate a strong sensitivity to input and output probability as well as to task frequency. Probing the "memorization" hypothesis via "counterfactual tasks," Zhaofeng Wu and colleagues found consistent and significant degradation of model performance under counterfactual conditions. They concluded that the better performance on standard task variants was due to overfitting (Wu et al. 2024).

12 Contrary to the impression created by headlines emphasizing the successes of generative AI, every output of LLMs can suffer from inherent limitations regarding reliability (Alshemali and Kalita 2020; Bosio et al. 2019; Kurenkov 2021). LLMs are not trained to consider the truth of their linguistic output. Critical voices in the debate about LLMs described, for example, ChatGPT as a "bullshit generator" (Hicks, Humphries, and Slater 2024). One may go so far as to conjecture that with the further development of such language models, a new class of weapons is emerging that can have devastating effects on the war for truth (Guardian Editorial 2023). This is what one could frame as a building block contributing to an epistemological crisis.

13 One notable example of the unreliability of LLM-generated text is a report in *The Guardian* about Amazon selling mushroom-picking guides that were apparently written using ChatGPT or another generative AI (Milmo 2023). This story clearly shows that trusting texts written by generative AI can even have fatal consequences. In this case, poisonous mushrooms were described as edible, and eating poisonous mushrooms is potentially fatal. Therefore, overreliance on LLMs can have disruptive consequences (Hopster 2021).

14 Another critical point is the fact that LLMs tend to hallucinate. One may criticize the choice of the term, which was popularized by Google AI researchers (Agarwal et al. 2018), and rather describe this kind of erroneous output as delivering untenable hypotheses. The term 'hallucination' is used to refer to mistakes in generated texts that are semantically or syntactically plausible but are, in fact, incorrect or nonsensical. For example, it is well-known that LLMs frequently hallucinate book and article references. Whether such errors can be rectified in future LLMs is the subject of controversial debates; OpenAI is full of hope, while Yann LeCun raises general objections (Agrawal, Mackey, and Kalai 2023; Smith 2023). To my knowledge, there is no example to date that proves that an LLM-based system could be completely weaned off the tendency to hallucinate, which leads me to claim that hallucinations are a feature and not a bug arising from the architecture of LLMs. In this context, it must, of course, be mentioned that hybrid sys-

tems offer additional verification procedures that can filter out hallucinations.

15 The list of documented erroneous outputs from LLMs is long (Marcus and Davis 2020; 2023), and none of the benchmarks can demonstrate robust 100% accuracy. Moreover, there is a growing body of research showing that LLMs cannot stand up to stress tests. Therefore, it should be obvious that LLM's outputs should better be checked by humans with expert knowledge in the domain in question. All this is not to belittle the impressive performance of LLMs in many areas. The point here is to emphasize that careful use of AI tools requires users to have expertise that enables them to verify the reliability of the results. Of course, translation software can be used to get an idea of what a text written in an unknown language is about. However, when it comes to vital decisions, for example, when signing a contract with far-reaching consequences, you are still well advised to seek the expertise of a person who is proficient in that language.

16 Transferring these considerations to the education sector, I argue that it is particularly important that learners, who do not yet have the necessary expertise to verify the output of the machines, become aware that they cannot blindly trust what the machine tells them. Precisely because one can often experience that much of what the machines produce is extremely valuable, awareness should be raised that the results are not reliably truthful. This is especially important because the way in which LLMs present their results often sounds very convincing. The undisputed ability of LLMs to produce grammatically correct and superficially sophisticated-sounding texts contributes to the fact that their results are often too quickly categorized as trustworthy.

17 However, the inherent unreliability of LLMs clearly indicates that it is a risky business to take over texts from LLMs without being able to verify them. Made-up references are a vivid example of unreliable outputs; they are a reason for failing examinations, regardless of whether the teacher can prove the use of AI tools. Moreover, it must also be stated that LLMs are not a good, or at least not a reliable, guide when it comes to clarifying knowledge questions. They may be a source of inspiration; they may be helpful to get an initial overview of a topic. However, without the expertise to verify the statements of LLMs, it is a dangerous strategy to rely on LLMs blindly. All of this indicates that integrating awareness of fallibility as well as learning the ability to verify results is an important building block for the responsible use of AI tools, especially in the education sector. In addition to teaching students to evaluate the results of AI tools critically, the educational goal should also be to make it natural for students to verify the results of an LLM. In my view, developing skills that allow a critical examination of AI tools' outputs should be incorporated into future curricula. For example, one approach could be to create assignments where students correct the results of LLMs.

18 However, teaching verification skills at the very beginning of a learning process might be quite challenging because successfully implementing verification processes presupposes a certain level of expertise, which beginners still lack. In the same way that calculators are not used to learn basic arithmetic, it would probably make sense to agree on a phased introduction of AI tools in schools. Furthermore, it should also be noted that debugging can prove to be particularly difficult because machines also produce unhuman-like errors and are not easily recognized by hu-

mans. It remains to be seen what results the new research will provide to this topic.

19 In addition to using AI tools to create certain text formats, another field of application for AI tools is knowledge acquisition. Here, in particular, learners who are not experts in a specific domain are especially vulnerable since they do not yet have the necessary expertise to verify the output. Even if the answers of an AI tool represent more knowledge than a learner has, it is not obvious to the learner whether the output may also contain incorrect assertions. I will not go into the debate at this point, which addresses whether the average output of AI tools might be considered good enough. The question here is not whether the use of AI tools can be useful despite their limitations in reliability if it enables individuals to do something that would otherwise be completely beyond them. Instead, the question is to what extent AI tools, no matter how good they are, can undermine a learning process and may pave further ways to misinformation.

20 In order to convey a critical attitude towards the use of AI tools, it is also important that knowledge about the way AI tools work is included in the curriculum. Together with research results from the field of explanatory AI, one could imagine that the creation or fine-tuning of AI tools could also be a useful extension of future curricula.

21 In conclusion, raising awareness of reliability limitations and fostering the development of verification skills is an indispensable step in promoting the responsible use of AI tools. However, this presupposes that you are aware of when you are dealing with machine-generated text. This leads us to the next critical issue: the increasing indistinguishability between human-created and machine-generated text.

Increasing indistinguishability between human-created and machine-generated text

22 Nowadays, it is not always clear whether one is even aware that a particular text is a machine-generated text. In order to make use of the ability to evaluate the results of an AI tool critically, it would be helpful if all machine-generated text were labeled as such. Although efforts are being made to prescribe labeling for machine-generated text (*Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence and Amending Regulations* 2024), it cannot be assumed that this idea is followed everywhere, especially on the World Wide Web. This poses a real problem since it is neither possible for humans nor for detection software to distinguish machine-generated text from human-created text with certainty.

23 Ten years ago, no one gave much thought to their discrimination abilities regarding machine-generated and human-authored text. Back then, the differences were so obvious, and it did not look like this was going to change any time soon. However, with the advent of more and more sophisticated LLMs, this is becoming a serious problem. In the following sections, I will refer to

various studies indicating that neither humans nor detection software are currently able to distinguish with certainty between machine-generated and human-authored text. LLMs have a remarkable capacity to generate texts that resemble human linguistic content. Several studies indicate that humans are not able to recognize machine-generated text with certainty (Brown et al. 2020; Clark et al. 2021; Dugan et al. 2020; Gao et al. 2022; Porter and Machery 2024; Schwitzgebel, Schwitzgebel, and Strasser 2023).

24 In one of those studies (Clark et al. 2021b), participants were confronted with samples of a human-made short text out of three domains (stories, news articles, and recipes) and with comparable samples of machine-generated text. With respect to GPT-3 generated texts, participants' accuracy in distinguishing was not significantly different from chance. In addition, they also investigated human discrimination skills in relation to machine-generated outputs using early models such as GPT-2. This comparison indicated a significant decrease in discrimination skills in the later models.

25 Already in 2020, with the publication of the paper 'Language Models are Few-Shot Learners' that introduced GPT-3, Tom Brown and colleagues reported similar results with respect to news articles; they found moderately good discrimination rates for smaller language models and near-chance performance with the largest version of GPT-3 (Brown et al. 2020b).

26 Investigating a different type of text—scientific abstracts—and involving experts as participants, Catherine Gao and colleagues used ChatGPT to generate scientific abstracts. In their study, they asked scientists to distinguish these abstracts from human-written scientific abstracts. Although the scientists were well above the 50% chance rate in distinguishing machine-generated from human-created abstracts, they still had a false negative rate of 32% (classification of machine-generated texts as human-written) and a 14% false-positive rate (classification of human-written texts as machine-generated) (Gao et al. 2022b).

27 In a study I conducted together with Eric and David Schwitzgebel, we examined the performance of a digital replica of the philosopher Daniel Dennett (*DigiDan*) and showed that even experts of Dennett's work were not able to distinguish with certainty between text snippets created by the human from text snippets generated by the machine (Schwitzgebel, Schwitzgebel, and Strasser 2023). This indistinguishability will surely increase with further advances in generative AI.

28 One might hope that if people do not achieve this ability to discriminate, they could perhaps use so-called detection software. But here, too, it is clear that a 100% distinction is out of reach, at least at the present time. Even though various companies providing plagiarism checker software have expanded their offers to include machine-generated text detection software, one has to emphasize that such software cannot prove beyond doubt whether a text has been written by a humanor by an AI tool. At the current state of research, no detection software could distinguish with 100% certainty between machine-generated and human-authored text. All detectors for LLM-generated text commit two types of errors: false-negative (machine-generated text falsely judged to be written by humans) and false-positive errors (human-generated text falsely judged

to be machine-generated). According to a study published in 2023 that covered 12 publicly available tools as well as two commercial systems (*Turnitin* and *PlagiarismCheck*), none of the tested detection software was accurate or reliable; all scored below 80% accuracy, and only five over 70% (Weber-Wulff et al. 2023). The findings of this study are consistent with a series of previously published studies (Anderson et al. 2023; Demers 2023; Elkhatat, Elsaid, and Almeer 2023; Gao et al. 2022a; Gewirtz 2023; Krishna et al. 2023; Pegoraro et al. 2023; van Oijen 2023; Wang et al. 2023). In view of those rather disappointing results, other sources claiming up to 98% accuracy look suspiciously like advertising and cannot refer to experimental studies (Compilatio 2024; Crossplag 2024; Winston AI 2024; Zero GPT 2024).

29 Especially, false positives can be very harmful to humans. Just imagine what it means for students when they are accused of not having written their essays themselves, even though they did (Davalos and Yin 2024). As long as we cannot exclude that such detectors falsely accuse humans of cheating, they should be used with caution and with the knowledge that their judgment could be false (Strasser 2024). It may be that in specific cases, for example, when one can compare a student's previous performance with their presented work, one may assume with some certainty that aids were used. However, one cannot prove this by using detection software. One possible reaction would be to invite students to an additional oral examination in suspicious cases, in which they can then at least show that they understand what the text they submitted is communicating.

30 It is certainly not negligible that increasing indistinguishability plays a role in evaluating exam performance. In addition to the already common demand for a declaration of independence, explanations about the use of AI tools are being developed here. Furthermore, education may also raise awareness that frequent fraud attempts can affect one's own learning process.

31 However, the consequences of indistinguishability also have an impact on other areas like knowledge acquisition. Even before the flood of LLM-produced texts, it was already a specific skill that had to be learned to find relevant and trustworthy content on the Internet. Here, the increasing indistinguishability between machine-generated and human-created texts makes it, in particular, more difficult for students to assess the trustworthiness of texts found on the World Wide Web. Precisely because LLMs are so good at mimicking human linguistic performance, we should consider that with the help of LLMs, it becomes easier than ever to create an infinite amount of text for fake websites, which in turn will lead to a decline in the level of trustworthiness of texts found on the Internet.

32 In various countries, laws are being developed that require the labeling of machine-generated text, for example,as the EU-AI Act implements it (*Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence and Amending Regulations* 2024). However, verifying compliance with these laws is difficult to impossible, and the different legislation in different countries also contributes to the fact that there will be no general legislation for the whole Internet.

33 The combination of the inherent unreliability and the increasing indistinguishability has the

potential to contribute to an epistemological crisis. If one is unable to distinguish machine-generated text from human-made text, one can no longer rely on established strategies for determining the reliability of found text sources. Especially because we cannot rule out the possibility that what were previously considered trustworthy sources are also interspersed with unverified machine-generated text. It is easy to imagine that machine-generated texts often make the appearance of coming from a reputable source. AI tools have the potential to sound like they would present the results of scientific research, even though many of the references do not refer to any published work. Moreover, even with respect to peer-reviewed papers, one cannot exclude that parts of the texts may entail unchecked outputs from AI tools.

34 In this paper, I focus on machine-generated text, but it should also be noted that even videos can be created in which a well-known person is said to appear and make statements that this person never made. In this context, I would like to refer to a paper by Daniel Dennett, who has dealt with the possibility of creating so-called counterfeit people and pointed out the disastrous consequences of such counterfeits for our societies. Thinking about the future impact of digital replicas, Daniel Dennett claimed that "counterfeit people are the most dangerous artifacts in human history, capable of destroying not just economies but human freedom itself." (Dennett 2023)[1]

35 In the face of this uncertainty, which becomes all the more relevant, the more texts are accessible primarily in electronic form, the crucial question arises as to whether we are in the epistemological position to know when we are confronted with machine-generated output. And this puts the applicability of verification skills to the test in a completely different dimension. If, over the long term, you can no longer clearly distinguish between texts that are based on sound expert knowledge and texts that are based on machine-generated, unverified content, then this naturally leads to problems when you want to apply your verification skills. In my view, such a development has the potential to contribute to an epistemological crisis. Regardless of which sources you use to verify the output of an AI tool, you will be confronted with the uncertainty that even the sources you use may contain unchecked machine-generated text. It is, therefore, obvious to point out that the widespread careless use of AI tools can contribute to the erosion of established relationships of trust.

36 Laws requiring special labeling of machine-generated texts may counteract this development partly. However, it must be clear that due to the increasing indistinguishability, we are not in a good position to verify and enforce compliance with such laws. In this respect, a special responsibility arises here in the field of education to ensure that they establish protected areas in which trustworthy texts are accessible.

1. At this point, as one of the creators of the digital replica of Daniel Dennett (DigiDan) (Strasser, Crosby, and Schwitzgebel 2023), I would like to point out that we developed this model with his consent but under the condition that it should not be publicly accessible and, furthermore, should be deleted after his death, which we have of course adhered to. Daniel Dennett passed away on April 19, 2024, and is greatly missed. A video recorded with him—the DigiDan installation—gives an impression of this model and Dennett's attitude towards it (Strasser 2023).

Deskilling as a form of cognitive atrophy

37 The above considerations show that the inherent unreliability of LLMs outputs, as well as the increasing indistinguishability, already poses severe challenges for the educational sector. In an educational setting, however, the aim is not only to learn how to use aids critically but also to develop cognitive skills such as the ability to write and to argue. For instance, one goal of our education institutions is to help students further develop the cognitive skills they need for elaborating the main claims of an argument, summarizing text, and writing text. In other words, education should foster critical thinking.

38 At the same time, a huge variety of AI tools are made publicly available, with which some of those effortful and time-consuming tasks can be solved without much effort. As a strict restriction on the use of such tools is not feasible, we have to think about the question of how to integrate such tools into the curricula. As argued above, it is important that students learn when and how to use such tools and get insights about critical limitations. However, early and frequent use of AI tools might present an obstacle to learning (and maintaining) the ability to solve certain tasks themselves. Moreover, this will lead to a critical dependence on AI tools (Tacca and Gilbert 2024).

39 An example illustrating the loss of certain abilities is the widespread use of GPS navigation devices. A majority of people nowadays rely on such devices to go from A to B. Thereby, the users tend to lose skills and abilities that are important for spatial orientation, like the ability to read a physical map (Dahmani and Bohbot 2020). Of course, users have learned the new useful skill of using navigation devices, but the price seems to be that their overall orientation skills are diminished or limited. One might object here that reading maps is no longer a necessary skill these days. But this does not hide the fact that a strong dependency on such devices is developing.

40 Turning to AI tools that can be used for text production, it seems that the situation is somehow different. I argue that due to the limitations of current AI tools in terms of reliability, responsible use of AI tools requires the user to have verification skills. But in order to acquire such skills, it is crucial that the users themselves also acquire the abilities required to perform the tasks that AI tools will later take over. It seems obvious that for learning verification skills, the development of cognitive abilities is essential, even if some of those abilities are no longer needed if one uses AI tools. Examples of such abilities are the ability to summarize long and complex essays, to work out and understand the structure of argumentation as well as to express one's own thoughts in an essay.

41 Considering that responsible use of AI tools requires a certain level of expertise, it is essential that the education sector, in particular, successfully contributes to training future experts who

are able to verify the outputs of future AI tools. Using an example related to the education of programmers, it can be argued that the widespread use of AI tools may have a negative impact. For example, it is feasible for so-called seniors, who have the necessary expertise to verify the code suggested by AI tools, to use AI tools instead of junior employees. The juniors are then no longer needed, and costs can be saved if AI tools are used instead. In that case, the question arises as to how and where future seniors will be trained in the required verification skills.

42 Especially the application of AI tools in highly specialized areas requires human experts who are able to recognize erroneous machine output. For example, there is an ongoing discussion about whether AI tools that can detect malignant tumors are potentially more reliable than experts because these tools can process a much larger amount of training data. That may be true, but I would argue that it would be irresponsible to develop future AI tools with these abilities without drawing on human expertise and, above all, without placing a high value on training future experts as well.

43 Even though deep neural networks outperform humans in pattern recognition, they can nonetheless also recognize features entailed in patterns that are not useful for the task at hand. For example, a random distribution of certain patterns in the training data used can cause an AI tool to incorrectly evaluate certain features as significant features that are not related to the pattern that should be recognized. A vivid example of this was provided by the development of AI tools that should recognize malignant tumors. Brian Christian refers to the example of a neural network that is known to have achieved accuracy comparable to that of dermatologists in diagnosing malignant skin lesions (Christian 2021; VentureBeat 2021). However, a closer examination of the model's methods revealed that the feature this model looked for in an image of a person's skin was the presence of a ruler. Since medical images of cancerous lesions often include a ruler, the model learned to identify the presence of a ruler as a marker of malignancy.

Concluding remarks

44 In this paper, I focused on critical issues related to AI tools. I have left out areas where AI tools can have a positive impact. It is not my intention to deny that these technological innovations can have a positive impact; they allow us to process large amounts of data and can enable us to accomplish tasks that would be unthinkable without them. AI tools can certainly be helpful in education, too; examples include the use of tools that can help provide access to complex texts by creating simplified summaries or can help provide a first overview of a knowledge domain. In addition, I think AI tools offer the potential to develop individualized learning software, but that was not the topic of this paper.

45 Therefore, I would like to emphasize that AI tools can only meaningfully support the educational mission if a critical awareness of the inherent reliability problems accompanies their use. I conclude with the following recommendations.

46 Due to the inherent reliability problems of tools based on generative AI technology, it is important to take care of learning verification skills. In order to meet the need for verification skills, it should be a goal in education to teach a critical approach to these tools. In doing so, the connection between the development of the required verification skills and the individual learning of their own problem-solving skills in relation to tasks that can admittedly often be solved by AI tools should be taken into account. This can also counteract a critical kind of dependency on such tools.

47 Given our inability to distinguish machine-generated text with certainty from human-written text that has been authored by experts or at least verified by experts, I advocate that it is in our own interest to develop procedures for labeling text that human experts have verified. In order to use verification skills, we need protected areas where we can find sources that we can trust. This leads to the general advocacy that we should strive for unverified machine-generated texts also to be clearly labeled as such.

48 Even if future AI tools become more reliable, we should consider whether we want to live in a world where humans are completely dependent on all kinds of AI tools because they no longer learn to solve tasks that are solved by tools. Not to mention that training future experts who can evaluate the outputs of AI tools is certainly a prerequisite for the further development of AI tools. To counteract the development of deskilling, we should continue to teach our students the skills they need to solve the tasks that can also be performed by AI tools. And to prevent the use of learned verification skills from becoming impractical, we should ensure that there are still places where reliable sources can be found.

49 I would go so far as to formulate the last recommendation as a general one: machine-generated text should be labeled in principle because, for future scientific progress, it is necessary that we do not lose the ability to rely on sources that already contain verified knowledge. Even if you think that the considerations put forward here represent a technophobic and very dystopian view of the future, you should perhaps think about what measures could be taken to prevent a future in which potentially every text could contain machine-generated errors. To sum up, I argue that careless and widespread use of AI tools has the potential to trigger an epistemological crisis.

References

"2023 Writers Guild of America Strike." 2024. In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=2023_Writers_Guild_of_America_strike&oldid=1258070372.

Agarwal, Ashish, Clara Wong-Fannjiang, David Sussillo, Katherine Hawley, and Orhan Firat. 2018. "Hallucinations in Neural Machine Translation." In *ICLR*. https://research.google/pubs/hallucinations-in-neural-machine-translation/.

Agrawal, Ayush, Lester Mackey, and Adam Tauman Kalai. 2023. "Do Language Models Know When They're Hallucinating References?" arXiv. http://arxiv.org/abs/2305.18248.

Alshemali, Basemah, and Jugal Kalita. 2020. "Improving the Reliability of Deep Neural Networks in NLP: A Review." *Knowledge-Based Systems* 191 (March):105210. https://doi.org/10.1016/j.knosys.2019.105210.

Anderson, Nash, Daniel L Belavy, Stephen M Perle, Sharief Hendricks, Luiz Hespanhol, Evert Verhagen, and Aamir R Memon. 2023. "AI Did Not Write This Manuscript, or Did It? Can We Trick the AI Text Detector into Generated Texts? The Potential Future of ChatGPT and AI in Sports & Exercise Medicine Manuscript Generation." *BMJ Open Sport & Exercise Medicine* 9 (1): e001568. https://doi.org/10.1136/bmjsem-2023-001568.

Barnes, Brooks, and John Koblin. 2023. "On Day 146, Screenwriters Reach Deal With Studios to End Their Strike." *The New York Times*, September 25, 2023, sec. Business. https://www.nytimes.com/2023/09/25/business/media/hollywood-writers-strike-deal.html.

Bosio, Alberto, Paolo Bernardi, Annachiara Ruospo, and Ernesto Sanchez. 2019. "A Reliability Analysis of a Deep Neural Network." In *2019 IEEE Latin American Test Symposium (LATS)*, 1–6. Santiago, Chile: IEEE. https://doi.org/10.1109/LATW.2019.8704548.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020a. "Language Models Are Few-Shot Learners." https://doi.org/10.48550/arXiv.2005.14165.

———. 2020b. "Language Models Are Few-Shot Learners." https://doi.org/10.48550/arXiv.2005.14165.

Christian, Brian. 2021. *The Alignment Problem: Machine Learning and Human Values.* First published as a Norton paperback. New York, NY: W. W. Norton & Company.

Clark, Elizabeth, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021a. "All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text." https://doi.org/10.48550/arXiv.2107.00061.

———. 2021b. "All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text." https://doi.org/10.48550/arXiv.2107.00061.

Compilatio. 2024. "What Is the Best AI Detector?" Compilatio. January 19, 2024. https://blog.compilatio.net/en/blog/best-ai-detectors.

Crossplag. 2024. "AI Content Detector." *Crossplag* (blog). 2024. https://crossplag.com/ai-content-detector/.

Dahmani, Louisa, and Véronique D. Bohbot. 2020. "Habitual Use of GPS Negatively Impacts

Spatial Memory during Self-Guided Navigation." *Scientific Reports* 10 (1): 6310. https://doi.org/10.1038/s41598-020-62877-0.

Davalos, Jackie, and Leon Yin. 2024. "AI Detectors Falsely Accuse Students of Cheating—With Big Consequences." *Bloomberg.Com*, October 18, 2024. https://www.bloomberg.com/news/features/2024-10-18/do-ai-detectors-work-students-face-false-cheating-accusations.

Dehouche, Nassim. 2021. "Plagiarism in the Age of Massive Generative Pre-Trained Transformers (GPT-3)." *Ethics in Science and Environmental Politics* 21 (March):17–23. https://doi.org/10.3354/esep00195.

Demers, Tom. 2023. "16 of the Best AI and ChatGPT Content Detectors Compared." Search Engine Land. April 25, 2023. https://searchengineland.com/ai-chatgpt-content-detectors-395957.

Dennett, Daniel C. 2023. "The Problem With Counterfeit People." *The Atlantic*. https://www.theatlantic.com/technology/archive/2023/05/problem-counterfeit-people/674075/.

Dugan, Liam, Daphne Ippolito, Arun Kirubarajan, and Chris Callison-Burch. 2020. "RoFT: A Tool for Evaluating Human Detection of Machine-Generated Text." arXiv. https://doi.org/10.48550/arXiv.2010.03070.

Elkhatat, Ahmed M., Khaled Elsaid, and Saeed Almeer. 2023. "Evaluating the Efficacy of AI Content Detection Tools in Differentiating between Human and AI-Generated Text." *International Journal for Educational Integrity* 19 (1): 17. https://doi.org/10.1007/s40979-023-00140-5.

Frye, Brian L. 2022. "Should Using an AI Text Generator to Produce Academic Writing Be Plagiarism?" SSRN Scholarly Paper. Rochester, NY. https://papers.ssrn.com/abstract=4292283.

Gao, Catherine A., Frederick M. Howard, Nikolay S. Markov, Emma C. Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T. Pearson. 2022a. "Comparing Scientific Abstracts Generated by ChatGPT to Original Abstracts Using an Artificial Intelligence Output Detector, Plagiarism Detector, and Blinded Human Reviewers." https://doi.org/10.1101/2022.12.23.521610.

———. 2022b. "Comparing Scientific Abstracts Generated by ChatGPT to Original Abstracts Using an Artificial Intelligence Output Detector, Plagiarism Detector, and Blinded Human Reviewers." https://doi.org/10.1101/2022.12.23.521610.

Gewirtz, David. 2023. "Can AI Detectors Save Us from ChatGPT? I Tried 5 Online Tools to

Find Out." ZDNET. October 19, 2023. https://www.zdnet.com/article/can-ai-detectors-save-us-from-chatgpt-i-tried-5-online-tools-to-find-out/.

Guardian Editorial. 2023. "The Guardian View on ChatGPT Search: Exploiting Wishful Thinking." *The Guardian*, February 10, 2023, sec. Opinion. https://www.theguardian.com/commentisfree/2023/feb/10/the-guardian-view-on-chatgpt-search-exploiting-wishful-thinking.

Hatcher, William Grant, and Wei Yu. 2018. "A Survey of Deep Learning: Platforms, Applications and Emerging Research Trends." *IEEE Access* 6:24411–32. https://doi.org/10.1109/ACCESS.2018.2830661.

Herman, Daniel. 2022. "The End of High-School English." *The Atlantic*, December. https://www.theatlantic.com/technology/archive/2022/12/openai-chatgpt-writing-high-school-english-essay/672412/.

Hicks, Michael Townsen, James Humphries, and Joe Slater. 2024. "ChatGPT Is Bullshit." *Ethics and Information Technology* 26 (2): 38. https://doi.org/10.1007/s10676-024-09775-5.

Hopster, Jeroen. 2021. "What Are Socially Disruptive Technologies?" *Technology in Society* 67 (November):101750. https://doi.org/10.1016/j.techsoc.2021.101750.

Huang, Kalley. 2023. "Alarmed by A.I. Chatbots, Universities Start Revamping How They Teach." *The New York Times*, January 16, 2023, sec. Technology. https://www.nytimes.com/2023/01/16/technology/chatgpt-artificial-intelligence-universities.html.

Hutson, Matthew. 2022. "Could AI Help You to Write Your next Paper?" *Nature* 611 (7934): 192–93. https://doi.org/10.1038/d41586-022-03479-w.

———. 2023. "Rules to Keep AI in Check: Nations Carve Different Paths for Tech Regulation." *Nature* 620 (7973): 260–63. https://doi.org/10.1038/d41586-023-02491-y.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. "Highly Accurate Protein Structure Prediction with AlphaFold." *Nature* 596 (August):583–89. https://doi.org/10.1038/s41586-021-03819-2.

Krishna, Kalpesh, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. "Paraphrasing Evades Detectors of AI-Generated Text, but Retrieval Is an Effective Defense." arXiv. https://doi.org/10.48550/arXiv.2303.13408.

Kurenkov, Andrey. 2021. "The Inherent Limitations of GPT-3." 2021. https://lastweekin.ai/p/the-inherent-limitations-of-gpt-3.

Lee, Victor R., Denise Pope, Sarah Miles, and Rosalía C. Zárate. 2024. "Cheating in the Age of Generative AI: A High School Survey Study of Cheating Behaviors before and after the Release of ChatGPT." *Computers and Education: Artificial Intelligence* 7 (December):100253. https://doi.org/10.1016/j.caeai.2024.100253.

Marche, Stephen. 2022. "The College Essay Is Dead." The Atlantic. December 6, 2022. https://www.theatlantic.com/technology/archive/2022/12/chatgpt-ai-writing-college-student-essays/672371/.

Marcus, Gary, and Ernest Davis. 2020. "GPT-3, Bloviator: OpenAI's Language Generator Has No Idea What It's Talking about." *MIT Technology Review*. https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion.

———. 2023. "Large Language Models like ChatGPT Say The Darnedest Things." Substack newsletter. *The Road to AI We Can Trust* (blog). January 10, 2023. https://garymarcus.substack.com/p/large-language-models-like-chatgpt.

McCoy, R. Thomas, Shunyu Yao, Dan Friedman, Mathew D. Hardy, and Thomas L. Griffiths. 2024. "Embers of Autoregression Show How Large Language Models Are Shaped by the Problem They Are Trained to Solve." *Proceedings of the National Academy of Sciences* 121 (41): e2322420121. https://doi.org/10.1073/pnas.2322420121.

Milmo, Dan. 2023. "Mushroom Pickers Urged to Avoid Foraging Books on Amazon That Appear to Be Written by AI." *The Guardian*, September 1, 2023, sec. Science. https://www.theguardian.com/technology/2023/sep/01/mushroom-pickers-urged-to-avoid-foraging-books-on-amazon-that-appear-to-be-written-by-ai.

Mitchell, Melanie. 2023. "How Do We Know How Smart AI Systems Are?" *Science* 381 (6654). https://doi.org/10.1126/science.adj5957.

Oijen, Vivian van. 2023. "AI-generated text detectors: Do they work? | SURF Communities." March 31, 2023. https://communities.surf.nl/en/ai-in-education/article/ai-generated-text-detectors-do-they-work.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, et al. 2024. "GPT-4 Technical Report." arXiv. https://doi.org/10.48550/arXiv.2303.08774.

Pegoraro, Alessandro, Kavita Kumari, Hossein Fereidooni, and Ahmad-Reza Sadeghi. 2023. "To ChatGPT, or Not to ChatGPT: That Is the Question!" arXiv. https://doi.org/10.48550/arXiv.2304.01487.

Peritz, Aki. 2022. "A.I. Is Making It Easier Than Ever for Students to Cheat." *Slate*, September 6,

2022. https://slate.com/technology/2022/09/ai-students-writing-cheating-sudowrite.html.

Porter, Brian, and Edouard Machery. 2024. "AI-Generated Poetry Is Indistinguishable from Human-Written Poetry and Is Rated More Favorably." *Scientific Reports* 14 (1): 26133. https://doi.org/10.1038/s41598-024-76900-1.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence and Amending Regulations. 2024. http://data.europa.eu/eli/reg/2024/1689/oj/eng.

Schwitzgebel, Eric, David Schwitzgebel, and Anna Strasser. 2023. "Creating a Large Language Model of a Philosopher." *Mind & Language*, July, 1–23. https://doi.org/10.1111/mila.12466.

Smith, Craig S. 2023. "Hallucinations Could Blunt ChatGPT's Success - IEEE Spectrum." 2023. https://spectrum.ieee.org/ai-hallucination.

Sparrow, Jeff. 2022. "'Full-on Robot Writing': The Artificial Intelligence Challenge Facing Universities." *The Guardian*, November 18, 2022, sec. Australia news. https://www.theguardian.com/australia-news/2022/nov/19/full-on-robot-writing-the-artificial-intelligence-challenge-facing-universities.

Strasser, Anna, dir. 2023. *DIGITAL LAB DGPhil 2024: DigiDan Installation.* https://youtu.be/rNEHzQ15tJQ?si=VlsChaeTcJE5VJb9.

———. 2024. "On Pitfalls (and Advantages) of Sophisticated Large Language Models." In *Ethics in Online AI-Based Systems. Risks and Opportunities in Current Technological Trends.*, edited by Santi Caballé, Joan Casas-Roma, and Jordi Conesa. S.l.: Elsevier Academic Press.

Strasser, Anna, Matthew Crosby, and Eric Schwitzgebel. 2023. "How Far Can We Get in Creating a Digital Replica of a Philosopher?" In *Social Robots in Social Institutions*, edited by Raul Hakli, Pekka Mäkelä, and Johanna Seibt, 371–80. Frontiers in Artificial Intelligence and Applications 366. IOS Press. https://doi.org/10.3233/FAIA220637.

Tacca, Alessio, and Frederic Gilbert. 2024. "Just Copy-Paste Me! Assessing the Risks of Epistemic Dependence on Large Language Models." In *Anna's AI Anthology. How to Live with Smart Machines?*, edited by Anna Strasser, 31–52. Xenomoi.

Tlili, Ahmed, Boulus Shehata, Michael Agyemang Adarkwah, Aras Bozkurt, Daniel T. Hickey, Ronghuai Huang, and Brighter Agyemang. 2023. "What If the Devil Is My Guardian Angel: ChatGPT as a Case Study of Using Chatbots in Education." *Smart Learning Environments* 10 (1): 15. https://doi.org/10.1186/s40561-023-00237-x.

VentureBeat. 2021. "When AI Flags the Ruler, Not the Tumor — and Other Arguments for

Abolishing the Black Box (VB Live)." *VentureBeat* (blog). March 25, 2021. https://venturebeat.com/business/when-ai-flags-the-ruler-not-the-tumor-and-other-arguments-for-abolishing-the-black-box-vb-live/.

Wang, Jian, Shangqing Liu, Xiaofei Xie, and Yi Li. 2023. "Evaluating AIGC Detectors on Code Content." arXiv. https://doi.org/10.48550/arXiv.2304.05193.

Weber-Wulff, Debora, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. 2023. "Testing of Detection Tools for AI-Generated Text." *International Journal for Educational Integrity* 19 (1): 1–39. https://doi.org/10.1007/s40979-023-00146-z.

Winston AI. 2024. "The Most Trusted AI Detector | ChatGPT Detection Tool." Winston AI. 2024. https://gowinston.ai/.

Wu, Zhaofeng, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. "Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks." arXiv. https://doi.org/10.48550/arXiv.2307.02477.

Yeadon, Will, Oto-Obong Inyang, Arin Mizouri, Alex Peach, and Craig Testrow. 2022. "The Death of the Short-Form Physics Essay in the Coming AI Revolution." arXiv. https://doi.org/10.48550/arXiv.2212.11661.

Zero GPT. 2024. "AI Detector - Trusted AI Checker for ChatGPT, GPT4 & Bard." 2024. https://www.zerogpt.com/.