ChatGPT and the Illusion of Explanatory Depth

Yomn Elsayed (yallam02@gmail.com)
Department of Psychology, Education and Child Studies,
Erasmus University Rotterdam, Post Box 1738 3000 DR Rotterdam, The Netherlands

Steven Verheyen (verheyen@essb.eur.nl)
Department of Psychology, Education and Child Studies,
Erasmus University Rotterdam, Post Box 1738 3000 DR Rotterdam, The Netherlands

Abstract

The recent surge in the use of AI-powered chatbots such as ChatGPT has led to new challenges in academia. These chatbots can enable student plagiarism and the submission of misleading content, undermining educational objectives. With plagiarism detectors unreliable in the face of this issue, educational institutions have been struggling to update their policies apace. This study assesses the effectiveness of sending warning messages - a common strategy used to discourage unethical use of ChatGPT - and investigates the use of the illusion of explanatory depth (IOED) paradigm as an alternative intervention. An international sample of students was asked to rate their understanding of, likelihood to use, and moral stance toward ChatGPT-generated text in assignments both before and after either reading a cautionary university message or explaining how ChatGPT works. Results showed that the explanation task did lead to the expected reduction in ratings of understanding, but despite this, neither moral acceptability nor likelihood to use decreased along with it. Similarly, reading the cautionary message neither resulted in a change in likelihood to use nor in moral acceptability, although it unexpectedly increased ratings of understanding. The results suggest that tackling students' understanding of ChatGPT is insufficient when it comes to deterring its unethical use, and that future interventions might want to have students reflect on moral issues surrounding the use of AI-powered chatbots.

Keywords: illusion of explanatory depth; mechanistic explanations; teleological explanations; warnings; ChatGPT; large language models; chatbots

Introduction

Evelyn Thompson was one of many master's students whose performance declined as the semester went by. Her professor noticed that after executing the first two assignments perfectly, she received a 22/30 on the third assignment and only a 19/40 on the fourth, with the quality of her work decreasing as requirements grew more sophisticated. Though slightly concerning, her professor found nothing overly alarming with her submissions, and with a Turnitin similarity score of only 9%, she successfully passed the course. Except she did not, since Evelyn did not exist. Unbeknownst to the professor, he had assigned a passing grade at a master's level course to ChatGPT (Stutz et al., 2023).

ChatGPT was launched by OpenAI on November 30, 2022. Trained on massive text-based datasets, this AI-powered chatbot can perform a myriad of functions, ranging from simple translation to writing academic articles. Only days after it launched, it had already amassed one million users, a number that was multiplied by 100 globally just two months later (Cotton et al., 2024; Lo, 2023). Interestingly, Google Trends showed that from the top 5 search queries accompanying 'ChatGPT', two included the word 'plagiarism' (Cotton et al., 2024). The problem of plagiarism occurs alongside ChatGPT use in two ways: it is embedded in the output itself, with many expressing concerns that the model is trained on copyrighted work (Xiao, 2022), and is also invoked when students plagiarize by copying text directly from its output (Bašić et al., 2023).

The ability to articulate thoughts through language is arguably the primary ability that separates man from animal. Educational curriculums nurture this skill, but shortcuts enabling assignment completion without real practice threaten the system's purpose. Studies have shown that ChatGPT can pass licensing exams in the fields of medicine and law, can generate high-quality papers (Bašić et al., 2023; Choi et al., 2022; Kung et al., 2023; Rudolph et al., 2023), and can even display critical thinking when prompted (Susnjak, 2022). Others found that ChatGPT's essays are often shallow, characterized by generic claims which are substantiated by non-existent sources (King & ChatGPT, 2023; Rudolph et al., 2023). Moreover, researchers have warned that having ChatGPT to rely on may compromise problem-solving and critical thinking skills by discouraging self-led explorations and impairing truth discernment abilities (Kasneci et al., 2023; Pavlik, 2023). Both its disconcerting strengths and weaknesses threaten student learning, making it crucial for universities to seek solutions for the inappropriate use of ChatGPT and the likes.

Academic dishonesty is by no means a recent problem, and plagiarism detectors have been developed to combat one way it can manifest in digitally submitted assignments. Accessible AI chatbots like ChatGPT and Claude+ make the challenge of detecting academic dishonesty a whole new ballgame, however, with ChatGPT typically avoiding plagiarism detection (Dehouche, 2021; Khalil & Er, 2023; Rudolph et al., 2023; Ventayen, 2023) and AI detectors failing to provide material evidence for AI authorship (Tlili et al., 2023; Sadasivan et al., 2023).

Academic institutions have responded by sending messages warning students of being caught using AI tools, requiring the submission of integrity statements with

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

assignments, and banning ChatGPT altogether (Gehen, 2023). Others have taken a more nuanced approach, permitting ChatGPT as a search aid or language assistant, and judging on a case-by-case basis whether learning goals can be reached and whether student ownership of submissions will be maintained depending on how ChatGPT is used. Meanwhile, they prohibit pasting ChatGPT's text as if it is one's own and warn that detectors will likely identify AI-generated text. But do warnings like these suffice to prevent students from using ChatGPT in a dishonest manner?

The Illusion of Understanding

Due to the recency of this challenge and the shortage of empirical evidence on methods meant to address it (Lo, 2023), it is crucial to critically assess interventions designed to tackle the problem as well as explore new ones, especially ones that motivate students to avoid malpractice. Here we consider the possibility that students might be overconfident regarding the abilities of ChatGPT and the possibility to use it unethically without being caught. Overconfidence has been consistently found to promote risky behaviors. Overconfidence leads to risk underestimation (Busenitz & Barney, 1997) and increases the preference for riskier choices fueled by a belief in "beating the odds" (Camerer & Lovallo, 1999). One study even found that it causes CEOs to optimistically interpret company news, assume infallibility, and resist change (Schumacher et al., 2020). More directly related to the current topic, Bucciol et al. (2024) found that overconfident students are more likely to cheat. This has interesting potential implications: if overconfidence can cause CEOs to dismiss negative feedback or notifications of risk, it is likely that it can similarly cause students to underestimate the risks associated with unethical AI use.

A paradigm that shows promise for puncturing overconfidence is that of the illusion of explanatory depth (IOED), wherein after attempting to generate explanations of how something works in a detailed step-by-step manner, people tend to realize that their true level of understanding of it is less than what they had assumed it to be, which can lead them to make more informed decisions regarding its use (Fisher et al., 2014; Muntwiler & Eppler, 2023; Rozenblit & Keil, 2002; Vitriol et al., 2018).

Rozenblit and Keil (2002) found that the IOED phenomenon does not present the same across different objects. Muntwiler and Eppler (2023) argue that digital technologies are prone to an IOED, however, because they adhere to the factors outlined by Rozenblit and Keil (2002) as predictive of the effect: confusing internal representation and environmental support, multi-level complexity without clear endpoints, and rare reproduction of explanations. Using technology to solve a problem successfully may facilitate the formation of a trusted mental representation of the technology that is although teleologically useful not technically mechanistic, essentially confusing 'environmental support' with 'internal representation' (Rozenblit & Keil, 2002). Because technology usually features interconnected parts at different layers of abstraction (e.g., networks, combinations of hardware and software, …), the "problem of unbounded causal complexity" (Keil, 2007, as cited in Muntwiler & Eppler, 2023, p.3) makes a faulty assessment of understanding and explanatory ability likely, especially for laypeople who lack experience in assessing their understanding in this area. Since AI chatbots fall under digital technologies to which these features apply, an illusion of understanding them is likely to take place.

In fact, Chromik et al. (2021) demonstrated the occurrence of the IOED effect with explainable AI (XAI), which involves methods that make AI solutions understandable to users. Since Collaris et al. (2018) noted that users did not scrutinize XAI solutions' validity even when prompted to do so, an IOED was expected given laypeople's tendency for uncritical acceptance. Participants were put in a scenario in which they acted as a lender on a crowdlending platform. They were presented with 16 loan requests displayed alongside a prediction of risk generated by AI. They then performed multiple tasks, including writing a detailed explanation of their understanding of the model's prediction behavior. Results demonstrated the expected IOED effect, with most participants reporting decreased understanding after the procedure (Chromik et al., 2021). These explainable AI findings likely extend to AI chatbots used by students, since both involve laypeople using AI practically. It is perhaps even more likely to occur with regular AI than XAI; if an illusion of understanding occurs with AI that is designed to be understood, it seems even more likely to occur with AI not intended to be understood.

Effects of Puncturing the IOED

The question remains, even if an illusion of explanatory depth is demonstrated with AI chatbots, would this influence students' use of it? Given that an IOED can lead to faulty decisions based on a false sense of confidence (Fernbach et al., 2013), it is reasonable to infer that addressing this may influence behaviors based on it. The connection between reducing understanding and reducing adoption has been alluded to before. For instance, Fernbach et al. (2013) showed that consumers were less likely to adopt and spend money on innovations they viewed as overly complicated (see also Muntwiler & Epler, 2023). Similarly, Krosnick and Petty (1995) demonstrated that people were less committed to their stance toward a policy when they were less certain about their understanding of it. Additionally, the illusion of understanding can also make one overlook limitations of the object in question (Fernbach et al., 2013). The illusion of understanding how ChatGPT works may cause students to overlook the many limitations of using ChatGPT-generated text in assignments, which implies that puncturing it may make students factor in its limitations more than they would otherwise. A key potential background mechanism for the effects above is that puncturing the IOED has been extensively shown to induce intellectual humility (Sloman & Vives, 2022), which generally makes people more likely to

accept unfavorable information, which in this case might be that ChatGPT is not appropriate to use as a writing shortcut.[1]

However, studies investigating the effects of puncturing the illusion of explanatory depth have also shown that the picture is quite nuanced, with multiple factors exerting an influence. One such factor is the effect of morals. This is clearly demonstrated by Sloman and Vives (2022) who examined the effect of puncturing the IOED of policies on attitude extremity. They accounted for an additional factor: whether the policies alluded to a consequentialist frame or a protected value one. Sloman and Vives found that the IOED paradigm reduced attitude extremity only for consequentialist policies, suggesting that when the policy alluded to sacred moral values, the explanation was of little relevance. People's attitudes were then not a product of knowledge synthesis, but a manifestation of something more deeply settled, so puncturing the illusion of understanding tackled a factor of mere peripheral importance. Similarly, Voelkel et al. (2018) found that the IOED paradigm only reduced prejudice and dissimilarity for political moderates, and when interpreting this result, they invoked political moderates' lower tendency to moralize issues as a possible explanation. This implies that the effectiveness of the paradigm in reducing cheating through ChatGPT may depend on students' moral convictions.

Considering the above, this study aims primarily to explore a possible way to deal with the problem of immoral use of ChatGPT and similar software by students in assignments, and secondarily to gain insight into variables that facilitate or inhibit the readiness to violate academic integrity by using ChatGPT, such as moral convictions. To achieve this, participants from a high school and university population were asked to rate their level of understanding of how ChatGPT works, their willingness to use the text it generates directly in school assignments, and how morally right or wrong they find this to be. The experimental group was then tasked to explain how ChatGPT works and how the information it generates is verified, while the control group was presented with a university message meant to discourage students to use AI by warning about its moral ramifications and the possibility of detection and punishment. Following this, participants rated the same three questions again.

It was hypothesized that having students explain how ChatGPT works would result in a significant decrease in self-rated understanding, and that this in turn would result in students becoming less likely to use ChatGPT-generated text directly in assignments. It was also predicted that having students read a message outlining reasons not to use ChatGPT would result in a reduction of willingness to use, but not understanding. Furthermore, we explored the possibility that the extent to which likelihood of use changes alongside understanding would depend on students' moral ratings. Investigating this can provide valuable insights that may guide universities as they navigate the challenges posed by unregulated student access to large language models.

[1] Intellectual humility can also be argued to be a stance that universities would want to instill in students.

Method

Participants

We aimed to recruit at least 128 participants, as 64 was found to be the number of participants needed per condition in an independent samples design to observe an IOED effect with a power of .80 at alpha =.05 (Gaviria & Corredor, 2021; Johnson et al., 2016). Data collection ran from the 27th of April 2023 to the 24th of May 2023. Participants were recruited both online and on university campuses and participated on a voluntary basis. For students to be eligible to participate, they were required to be enrolled in a high school or university program and know what ChatGPT is. Students who majored in computer science were excluded to ensure that all participants were laypeople in AI mechanisms. Students who self-assessed their English reading proficiency as 'beginner' (CEFR level A1 or A2) were excluded as well to ensure everyone understood the instructions. Additionally, those who provided no explanations or nonsense explanations on the IOED explanation task (e.g., "I know how it works" or "It does everything"), had missing data on the three questions presented before and after the intervention, failed an attention check, or did not confirm reading the cautionary message or the scale training were excluded from the analyses as well.

The final sample after exclusions comprised 170 participants: $N = 80$ in the explanation condition (61 females) and $N = 90$ in the message condition (57 females). Most participants were of Arab descent (80%) and university students (88%). The sample age ranged between 16 and 33 years ($M = 20.9$, $SD = 2.49$). Participants' self-assessed mean writing proficiency was high, at 5.60 out of a maximum of 6 ($SD = .97$), which corresponds to the levels C1 and C2 on the CEFR ('advanced'). Writing proficiency did not differ by condition (explanation condition: $M = 5.69$, $SD = 1.00$; message condition: $M = 5.52$, $SD = 1.06$) as demonstrated by an independent samples t-test ($t(168) = 1.04$, $p = .70$, Cohen's $d_s = .16$). This and all to be reported tests are two-tailed and employ an alpha level of .05.

Procedure

The study was administered using Qualtrics (Qualtrics, Provo, UT). Participants first received a short introduction to the study, informing them it aimed to investigate perceptions of AI use among students. It also included a disclaimer meant to inhibit demand characteristics and encourage honest responding by stressing the anonymity of the responses. Participants were then presented with the training material from Rozenblit and Keil (2002), which explained how to use a 7-point scale for rating understanding. The training included an example of a level 1, level 4, and level 7 explanation for the workings of a bow. Participants were then presented with the following three questions that all required a response on a 7-point rating scale:

1. Please rate your understanding of how ChatGPT works. (1 = lowest level of understanding; 7 = highest level of understanding)

2. Please rate how likely you are to submit ready-made text from ChatGPT as part of a school assignment. (1 = extremely unlikely; 7 = extremely likely)[2]

3. To what extent do you find using ChatGPT-generated text in school assignments morally acceptable or morally wrong? (1 = morally wrong in most or all cases; 7 = morally right in most or all cases)

Participants were then either presented with an explanation prompt (experimental condition) or a message warning against the use of text generated from ChatGPT and other AI platforms, that was based on an actual message sent to master students at Erasmus University Rotterdam (control condition). The warning message highlighted the importance of students writing assignments themselves and brought to light multiple concerns regarding the use of AI-generated text in assignments, such as the moral issue of fraud, the possibility of text being based on unreliable or even non-existent sources, and the likelihood of it being detected by plagiarism detectors.

The explanation prompt in the experimental condition was derived from Rozenbleit and Keil (2002) and tailored to ChatGPT. It was written to allude to the same concerns regarding source selection and verification addressed in the message presented to participants in the control condition: "As best you can, please describe all the details you know about ChatGPT's generated responses, going from the first step to the last, and providing the causal connection between the steps. That is, your explanation should state precisely how each step causes the next step in one continuous chain from start to finish. In other words, try to tell as complete a story as you can, with no gaps. Consider how information presented by ChatGPT is selected, verified, and worded."

After this, participants were presented with the same three questions (understanding, likelihood to use, moral acceptability) a second time. The survey ended with several individual differences measures (academic integrity culture, deliberation, intuition, moralized rationality, importance of rationality), but these are not reported here because they were found to be unreliable (all Cronbach's alphas < .72). Participants reported their English reading and writing proficiency on the CEFR self-assessment scale (Council of Europe, n.d.) as part of the demographic information.

Results

The pre-intervention ratings reflect participants' default standing without intervention. The mean understanding rating was 4.17 (SD = 1.62), while the mean likelihood (M = 2.90, SD = 2.00) and moral acceptability ratings (M = 3.18, SD = 1.59) were markedly lower. Both understanding and moral acceptability were significantly positively correlated with likelihood to use: $r(168) = .40$, $p < .001$ and $r(168) = .50$, $p < .001$, respectively. Understanding and moral acceptability were also significantly positively correlated ($r(168) = 0.32$, $p < .001$).

Mixed ANOVAs were conducted to investigate whether the interventions (writing an explanation or reading a warning message) had an effect on the three measures. A mixed ANOVA investigating the effect of the intervention on understanding with time as the repeated measure and condition as a between-subjects factor demonstrated that there was no main effect of time: understanding ratings did not significantly change after intervention ($F(1,168) = 1.890$, $p = .171$, $\eta_p^2 = .001$). Understanding ratings overall did not differ between interventions either, with no main effect of condition ($F(1,168) = 2.15$, $p = .14$, $\eta_p^2 = .011$). There was, however, a significant time*condition interaction ($F(1,168) = 22.044$, $p < .001$, $\eta_p^2 = .013$). To better understand this interaction, we followed it up with a simple main effects analysis, which indicated that the effect of the intervention on self-rated understanding was in opposite directions for the two conditions (see Figure 1). In the explanation condition, understanding decreased as hypothesized ($F(1,168) = 15.648$, $p < .001$), whereas in the message condition understanding unexpectedly increased after the intervention ($F(1,168) = 6.502$, $p = .012$).

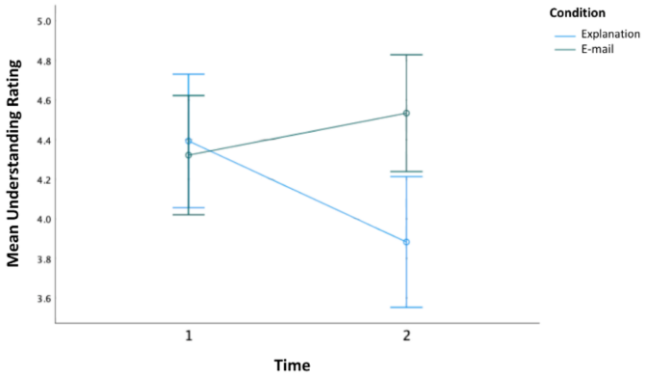

Figure 1: Change in understanding across time per condition. Error bars represent 95% confidence intervals.

Neither the likelihood to use nor the moral acceptability ratings changed after the interventions, with mixed ANOVAs showing no significant main effect of time on the likelihood ($F(1, 168) = .011$, $p = .92$, $\eta_p^2 = .00$) or moral acceptability ratings ($F(1,168) = .598$, $p = .44$, $\eta_p^2 = .00$). Similarly, there was no main effect of condition (intervention) on the likelihood ($F(1,168) = .01$, $p = .94$, $\eta_p^2 = .00$) or moral acceptability ratings ($F(1,168) = .00$, $p = .998$, $\eta_p^2 = .00$), and no significant time*condition interaction for likelihood ($F(1,168) = .94$, $p = .33$, $\eta_p^2 = .00$) or acceptability ($F(1,168) = .28$, $p = .60$, $\eta_p^2 = .00$).

To further understand these results, we also investigated whether changes in understanding, likelihood, and moral acceptability were related to each other. To this end, variables

[2] We opted for this formulation to address a clearly unethical and uncritical use of generative AI: Relying entirely on ChatGPT to solve an assignment and presenting the resulting output as one's own without any scrutiny or editing.

that encompass change from pre to post were created by subtracting post-intervention ratings from pre-intervention ratings. In keeping with IOED literature norms, a positive score on understanding change would thus indicate that the understanding rating had decreased after the task (i.e., a positive change indicates the illusion of explanatory depth), while a negative score would indicate that the understanding rating increased. A multiple linear regression was conducted on the sample as a whole to test whether change in understanding and moral acceptability predicted change in likelihood to use, and results showed that they did ($F(2, 167) = 20.44$, $p < .001$ with $R^2 = .20$). Further inspection revealed that only change in moral acceptability was a significant predictor of change in likelihood with $b = .63$ ($SE = .10$, $t = 6.24$, $p < .001$), while change in understanding was not ($b = -.01$, $SE = .06$, $t = -0.19$, $p = .85$). Since adding condition and its interactions with morality and understanding to the regression equation, yielded a significant condition*morality interaction, we repeated the original regression analysis per condition, and found that the observed pattern only held true for the message condition ($F(2,87) = 32.72$, $p < .001$, $R^2 = .43$). For this condition, change in moral acceptability predicted change in likelihood to use with $b = .82$, $SE = .11$, $t = 7.58$, $p < .001$, while change in understanding did not ($b = -.023$, $SE = .09$, $t = -0.27$, $p = .79$). This pattern could not be reproduced when the regression analysis was conducted for the explanation condition ($F(2,77) = .262$, $p = .77$, $R^2 = .01$), with neither change in understanding ($b = .01$, $SE = .10$, $t = 0.07$, $p = .94$) nor change in moral acceptability ($b = .144$, $SE = .20$, $t = 0.71$, $p = .48$) reliably predicting a change in the likelihood to use ChatGPT.

Discussion

The purpose of this study was twofold. On the one hand, it aimed to investigate whether the illusion of explanatory depth would occur with ChatGPT, and if so, whether this would influence other domains such as the likelihood of using it and its perceived moral acceptability. The study thus evaluated the IOED's potential as a tool for dissuading students to use ChatGPT unethically. On the other hand, the study also aimed to assess the effectiveness of sending cautionary messages to warn against using ChatGPT in ways that breach academic integrity, a practice that many universities have resorted to.

The results for the pre-intervention ratings suggested that using the IOED paradigm as an intervention seemed reasonable: students' self-rated understanding of ChatGPT was positively correlated with their likelihood of using it and their judgment of its moral acceptability. The observation that students who believed they understood ChatGPT better were also more likely to use it and tended to view it as more morally acceptable, allows for the hypothesis that puncturing such a potentially illusory understanding could possibly lead to reducing the likelihood of using the AI chatbot unethically.

The illusion was in fact punctured by the explanation task. Students who were tasked to explain how ChatGPT works reported a significant decrease in their understanding of how the AI chatbot works. This adds to the growing body of literature on the IOED in phenomena other than the physical devices the effect was initially observed for (Rozenblit & Keil, 2002; Zeveney, 2016) and indicates it also holds for algorithms.

In contrast, participants in the control condition showed a surprising increase in the self-reported understanding of ChatGPT after reading a cautionary message about the chatbot's use, despite the message providing no explanation of how ChatGPT works. We see two possible reasons for this. One possibility is that this represents a demand characteristic: participants might have felt that they were supposed to understand ChatGPT more after being presented with a message about it. The second is the known mishap of assessing understanding based on a teleological approach or an abstract construal rather than approaching the question literally and mechanistically (Alter et al. 2010; Johnson et al., 2016). The study design did attempt to counter this by including a disclaimer meant to inhibit demand characteristics and encourage honest responding, and by providing a mandatory training on how to rate understanding prior to the interventions. It is, however, possible that this training was not entirely effective due to the differences between the workings of the tangible object (a bow) used during training and that of an intangible item like an AI chatbot.[3] If the message condition was approached with an inappropriately abstract view of what understanding entails, the information presented in the message may have enhanced the impression of understanding rather than challenged it.

While it is not possible to directly assess whether participants in the control condition entertained such an abstract view of understanding, we can get a sense of how likely it is to have been an issue by examining the explanations provided by the participants in the explanation condition, seeing that their starting situation was very similar (same population, training, and initial questions). A post-hoc examination of the provided explanations showed that approximately 36% of responses were teleological, focused on explaining how to use ChatGPT and what it can do (e.g., "I ask it questions and it answers me") instead of how it works mechanistically. Although one should exert care with drawing inferences about one condition based on another one, this finding could indicate that encountering any information about ChatGPT - such as the little information about it in the control condition's cautionary message - could have increased participants' impression of understanding it. With the literature suggesting that an inaccurate assessment of one's understanding can make one unqualified to assess shortcomings (e.g., Fernbach et al., 2013; Schumacher et al., 2020), this is a concerning outcome in the context of existing attempts to dissuade students from unethical ChatGPT use by highlighting its limitations.

[3] One might argue that the decrease in understanding in the explanation condition counters this argumentation.

Neither the explanation generation task nor the cautionary message resulted in a change in the likelihood to use ChatGPT in assignments or its perceived moral acceptability, despite the change in understanding. The stagnant moral acceptability ratings are not entirely surprising, as previous studies have shown that moral convictions are resistant to change (Voelkel et al., 2018). The observation that the decrease of understanding in the explanation condition was not accompanied by a decrease in the likelihood to use ChatGPT is disheartening, however, as this speaks directly to the worry voiced by the chatbots' opponents that students will use it uncritically in light of its limitations. One reason why the explanation intervention may not have led to the expected results is also related to this: participants might not have engaged sufficiently with the task. The explanation prompt addressed the questions of how ChatGPT generates text and how its information is verified. This was meant to make students reflect on the gaps in their understanding and on how AI-generated text cannot be trusted blindly. However, many of the explanations did not address the prompt in sufficient depth to expect them to influence students' likelihood to use ChatGPT. If participants mainly focused on teleological features, this may have inhibited the level of reflection necessary to cause a puncturing in understanding that is likely to affect decision making. We may also need to entertain the possibility that students might not care how AI chatbots work, as long as the potential benefits outweigh the disadvantages.

The results regarding the changes that occur alongside each other yield some interesting insights. In the message condition, likelihood to use changed with changes in moral acceptability. When understanding was reduced, however, this did not lead to a reduction of likelihood to use. A reason why change in moral acceptability predicted change in likelihood to use in the message but not the explanation condition could be that the extent to which the explanation prompt and the cautionary message alluded to morality was slightly different. The message explicitly alluded to fraud, while the explanation prompt left this implicit. It only directed students to think about how information is selected and verified. The idea that generated text could be unethically sourced or fraudulent is one that students were meant to realize as a result of reflecting on the prompt instead of being directly provided. Students who did not put sufficient thought and effort in the explanation task, might not have realized this. The impact of morality might thus have been reduced in the explanation condition because moral issues were not communicated explicitly.

Limitations and Directions for Future Research

In addition to the limitations inherent to collecting data online, it needs to be acknowledged that participants in our study were not compensated. This might have deterred people who avoid "unnecessary" work (and who might also be more inclined to use ChatGPT to cut corners) from participating. The pre-intervention likelihood to use ratings underscore this point with a mean of 2.90 that is lower than the scale's midpoint. This rather low number could also reflect students' hesitancy to admit they use AI-generated text in school assignments (despite the anonymous response format) or could be an indication that many students use ChatGPT differently (e.g., to check spelling, rephrase, suggest literature, or propose counterarguments rather than having it write their entire paper for them, which is what we queried them about). We therefore also need to entertain the possibility that the lack of an intervention effect is due to a floor effect on this dependent variable. A linear regression analysis indicated that students who were initially more likely to use ChatGPT showed a greater reduction in their likelihood to use ChatGPT-text in assignments across interventions ($F(1,168) = 10.44$, $p < .001$, with $b = .02$, $SE = .01$, $t = 3.23$, and $p < .001$). Future studies might therefore want to target students who already use or are considering using AI chatbots. Similarly, given that our sample had a self-reported writing proficiency average of 5.60 out of 6 and that the literature suggests that those with low writing proficiency are more likely to use ChatGPT unethically (Namira et al., 2021), it might be worthwhile to include students with a wider range of writing abilities.

Another limitation is that data collection took place at around the same time universities were beginning to send out notices warning students of ChatGPT use in assignments. Because of this, some students may have already seen similar messages while others had not, which may have affected the pre-task likelihood ratings. Future studies could investigate whether the effects of the IOED paradigm differ between those who previously received a warning and those who did not, broadening the scope by exploring the paradigm's potential as a supplement to current interventions rather than a standalone one, and assess the impact of several interventions over time rather than immediately after one of them takes place.

Conclusion

Our study is the first to demonstrate the IOED with AI chatbots. Having participants explain how ChatGPT works decreased self-rated understanding. However, neither the IOED paradigm nor a cautionary message led to a reduction of students' likelihood to use ChatGPT or their perception of its moral acceptability. Our results thus show that the common idea that puncturing the illusion of understanding will influence behavior should not be directly assumed.

This study approached the discussion on the influence of ChatGPT in education from the angle of prevention and academic integrity, but it could also be approached differently, for instance by teaching students to use ChatGPT in assignments productively. This study has shown that appealing to students' awareness and understanding of chatbots' limitations is insufficient to combat the threats posed by AI applications. As universities continue to take on this challenge, we advise they aim to touch students' core values and ethics rather than merely raise awareness, noting that personal integrity seems to be a precursor of academic integrity.

References

Alter, A. L., Oppenheimer, D. M., & Zemla, J. C. (2010). Missing the trees for the forest: A construal level account of the illusion of explanatory depth. *Journal of Personality and Social Psychology, 99*(3), 436–451.

Bašić, Ž., Banovac, A., Kružić, I., & Jerković, I. (2023). Better by you, better than me, chatgpt3 as writing assistance in students essays. arXiv preprint arXiv:2302.04536.

Busenitz, L. W., & Barney, J. B. (1997). Differences between entrepreneurs and managers in large organizations: Biases and heuristics in strategic decision-making. *Journal of Business Venturing, 12*(1), 9-30.

Bucciol, A., Cicognani, S., & Montinari, N. (2024). It's time to cheat! *Journal of Behavioral and Experimental Economics, 108*, 102156.

Camerer, C., & Lovallo, D. (1999). Overconfidence and excess entry: An experimental approach. *American Economic Review, 89*(1), 306–318.

Choi, J. H., Hickman, K. E., Monahan, A. B., & Schwarcz, D. (2022). ChatGPT goes to law school. *Journal of Legal Education, 71*(3), 387-400.

Chromik, M., Eiband, M., Buchner, F., Krüger, A., & Butz, A. (2021). I think I get your point, AI! The illusion of explanatory depth in explainable AI. *26th International Conference on Intelligent User Interfaces*. https://doi.org/10.1145/3397481.3450644

Collaris, D. A. C., Vink, L. M., & van Wijk, J. J. (2018). Instance-level explanations for fraud detection: A case study. 28-33. Paper presented at 2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018), Stockholm, Sweden. https://arxiv.org/abs/1806.07129

Cotton, D. R., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in education and teaching international, 61*(2), 228-239.

Council of Europe. (n.d.). Self-assessment grids (CEFR). Retrieved May 3, 2023, from https://www.coe.int/en/web/portfolio/self-assessment grid

Dehouche, N. (2021). Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3). *Ethics in Science and Environmental Politics, 21*, 17-23.

Fernbach, P. M., Rogers, T., Fox, C. R., & Sloman, S. A. (2013). Political extremism is supported by an illusion of understanding. *Psychological Science, 24*(6), 939–946.

Fernbach, P. M., Sloman, S. A., Louis, R. St., & Shube, J. N. (2013). Explanation fiends and foes: How mechanistic detail determines understanding and preference. *Journal of Consumer Research, 39*(5), 1115–1131.

Gaviria, C., & Corredor, J. (2021). Illusion of explanatory depth and social desirability of historical knowledge. *Metacognition and Learning, 16*(3), 801–832.

Gehen, G. (2023, January 6). CHATGPT, artificial intelligence, and the future of writing. Psychology Today. Retrieved April 19, 2023, from https://www.psychologytoday.com/intl/blog/darwins-subterranean-world/202301/chatgpt-artificial-intelligence-and-the-future-of-writing

Johnson, D. R., Murphy, M. P., & Messer, R. M. (2016). Reflecting on explanatory ability: A mechanism for detecting gaps in causal knowledge. *Journal of Experimental Psychology: General, 145*(5), 573–588.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274.

Keil, F. C. (2007). Getting to the truthgrounding incomplete knowledge. *Brooklyn Law Review,73*(3), 1035-1052.

Khalil, M., & Er, E. (2023). Will ChatGPT get you caught? Rethinking of plagiarism detection. *Learning and Collaboration Technologies*, 475–487.

King, M. R., & ChatGPT (2023). A conversation on artificial intelligence, chatbots, and plagiarism in higher education. *Cellular and Molecular Bioengineering, 16*(1), 1–2.

Krosnick J. A., & Petty R. E. (1995). Attitude strength: An overview. In Petty R. E., & Krosnick J. A. (Eds.), *Attitude strength: Antecedents and consequences* (pp. 1–24). Erlbaum.

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digital Health, 2*(2), e0000198.

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences, 13*(4), 410.

Muntwiler, C., & Eppler, M. J. (2023). Improving decision-making through visual knowledge calibration. *Management Decision, 61*(8), 2374–2390.

Namira, M., Ping, M., & Suhatmady, B. (2021). Plagiarism awareness and academic writing ability: The relationship with the EFL Students' plagiarism practice. *Educational Studies: Conference Series*, 1(1). https://doi.org/10.30872/escs.v1i1.867

Pavlik, J. V. (2023). Collaborating with ChatGPT: Considering the implications of generative artificial intelligence for journalism and media education. *Journalism & Mass Communication Educator, 78*(1), 84-93.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 26*(5), 521–562.

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning & Teaching, 6*(1), 1-22.

Sadasivan, V. S., Kumar, A., Balasubramaniam, S., Wang, W., & Feizi, S. (2023). Can AI-generated text be reliably detected? arXiv preprint arXiv:2303.11156

Schumacher, C., Keck, S., & Tang, W. (2020). Biased interpretation of performance feedback: The role of CEO overconfidence. *Strategic Management Journal, 41*(6), 1139–1165.

Sloman, S. A., & Vives, M.-L. (2022). Is political extremism supported by an illusion of understanding? *Cognition, 225*, 105146.

Stutz, P., Elixhauser, M., Grubinger-Preiner, J., Linner, V., Reibersdorfer-Adelsberger, E., Traun, C., Wallentin, G., Wöhs, K., & Zuberbühler, T. (2023). Ch(e)atGPT? An anecdotal approach addressing the impact of ChatGPT on teaching and learning GIScience. *GI_Forum, 1*, 140–147.

Susnjak, T. (2022). ChatGPT: The end of online exam integrity? arXiv preprint arXiv:2212.09292

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (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.

Ventayen, R. J. (2023). OpenAI CHATGPT generated results: Similarity index of artificial intelligence-based contents. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4332664

Vitriol, J. A., & Marsh, J. K. (2018). The illusion of explanatory depth and endorsement of conspiracy beliefs. *European Journal of Social Psychology, 48*(7), 955–969.

Voelkel, J. G., Brandt, M. J., & Colombo, M. (2018). I know that I know nothing: Can puncturing the illusion of explanatory depth overcome the relationship between attitudinal dissimilarity and prejudice? *Comprehensive Results in Social Psychology, 3*(1), 56–78.

Xiao, Y. (2023). Decoding authorship: Is there really no place for an algorithmic author under copyright law? *IIC-International Review of Intellectual Property and Competition Law, 54*(1), 5-25.

Zeveney, A. S.(2016). Illusion of understanding in a misunderstood field: The illusion of explanatory depth in mental disorders. Master's Thesis, Lehigh University.