Taking the help or going alone: ChatGPT and class assignments*

Brian Hill
HEC Paris & CNRS†

March 31, 2023

Abstract

There is increasing speculation about the future role of ChatGPT and other artificial intelligence (AI) chatbots aiding humans in a variety of tasks. But do people do better when aided by these tools, as compared to when they complete tasks on their own? Can they properly evaluate and where necessary correct the responses provided by ChatGPT to enhance their performance? To investigate this question, this study gives university-level students class assignments involving both answering questions and correcting answers provided by ChatGPT. It finds a significant reduction in student performance when correcting a provided response as compared to when they produce an answer from scratch. One possible explanation for this discrepancy could be the confirmation bias. Beyond emphasising the need for continued research into human interaction with AI chatbots, this study exemplifies one potential way of bringing them into classroom: to raise awareness of the pitfalls of their improper use.

Keywords: ChatGPT; Human-AI chatbot interaction; Confirmation bias; Class assignments; AI in education; Future of work.

*Acknowledgements and thanks to be added.

†GREGHEC, CNRS & HEC Paris. 1 rue de la Libération 78351 Jouy-en-Josas, France. E-mail: hill@hec.fr.

Introduction

Since the release of ChatGPT by OpenAI, there has been increasing discussion, and in some quarters alarm, about its consequences for education, with frequent stories in particular about its performance on various exams and assignments (Stokel-Walker, 2022; Cavendish, 2023). But if, as many suggest, ChatGPT-like tools will be central to many work practices in the future, then we need to think not only about what assignments will look like in a post-ChatGPT world, but also about how to design course elements that help students learn how to use these tools properly. A correct use will not involve humans copying the output of these tools blindly, but rather them using it as a means to improve their own performance. This immediately raises the simple question: can students properly evaluate and where necessary correct the responses provided by ChatGPT, to enhance their performance?

Motivated by such considerations, we designed the following assignment, which was given to a first-year Masters-level introductory Behavioural Economics class at a prestigious business school ($n = 49$). Students were randomly assigned two out of 14 cases, where each case described an example of behaviour linked to a choice bias that had been presented in class. The exercise, for each case, was to identify the choice biases in the course that were most closely related to the behaviour described, and explain the relation.

For the first case assigned—call it the *answer* task—students just had to answer the question. For the second case—the *correct* task—students were provided with an answer to the question: they were asked whether the answer was fully correct, and if not, were asked to correct or add as required to make it 'perfect'. They were told that each answer had been either provided by ChatGPT or was the response given by a student from a previous year, but they were not told which. The marks for the two tasks counted equally towards the course grade.

Note that the two tasks are asking for the same thing: a full reply to the question concerning the case. However, while the *answer* task is arguably representative of traditional work practices, the *correct* task may correspond more closely to many jobs in the future, if AI tools become as ubiquitous as many predict. The human role will be to evaluate and correct the output of an AI—precisely as asked of students in this task.

Results and Discussion

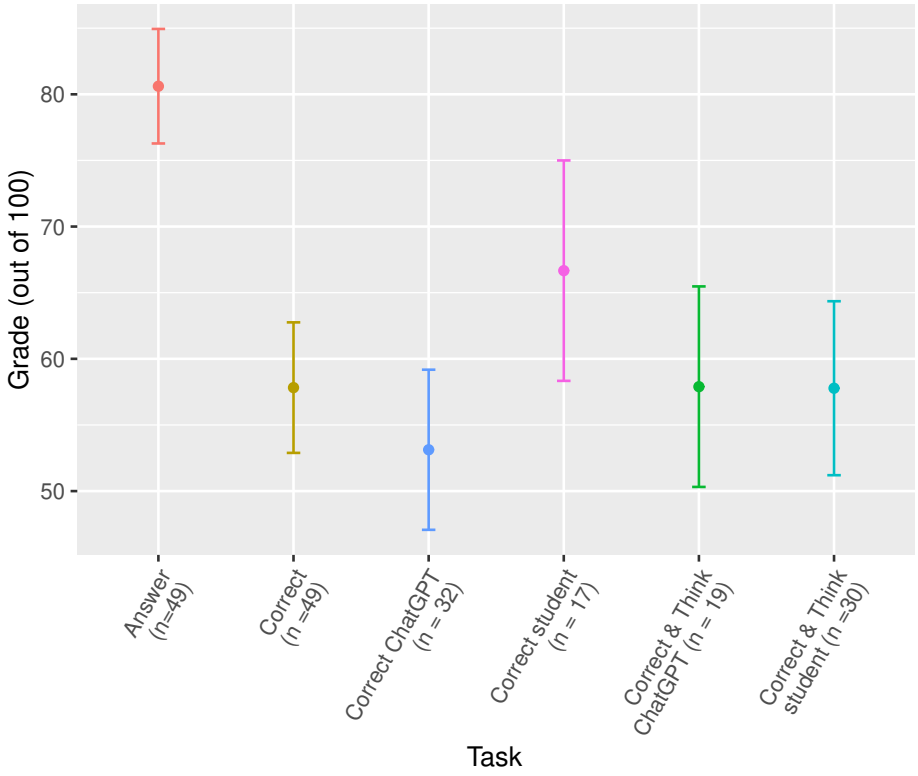

Figure 1: Mean grades with standard errors, for each task (answer or correct), and among the correct task, split according to the actual source of the response provided (ChatGPT or another student), and who the student thought provided the response. Sample size in brackets.

The average grade on the correct task was 28% lower than the average grade on the answer task (Figure 1), with students dropping 23 marks out of 100 on average. The gap between the performance in the answer and correct tasks persists if one limits attention to those who corrected answers provided by ChatGPT, or who thought they were correcting answers provided by ChatGPT. Regressions confirm this finding (Table 1): controlling for the case assigned, the grade (out of 100) is on average 28 points higher when the student is answering from scratch as compared to when she is correcting a response. Similar gaps are found when one focuses on the responses given

	Grade (out of 100)		
	(1)	(2)	(3)
Answer	27.82***	29.22***	25.63**
	(6.32)	(6.54)	(8.72)
Correct		7.64	
& student		(13.01)	
Correct			−4.38
& Think student			(11.16)
Controls	Yes	Yes	Yes
Observations	98	98	98
Clusters	49	49	49
R^2	0.27	0.28	0.28
Adj. R^2	0.15	0.15	0.14
F	12.29	11.48	11.21
p	< 0.000	< 0.000	< 0.000

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 1: Regressions of grade (out of 100) against task type (answer or correct)

Note: OLS regressions of grade against task type (baseline: correct), including who provided the answer for the correct task (Model (2); baseline: correct task and provided by ChatGPT) and who the subject thought provided it (Model (3); baseline: correct task and thought ChatGPT). Robust standard errors clustered by student; controls for case fixed effects.

when students were correcting ChatGPT-provided answers, or when they thought the answers were provided by ChatGPT. No significant difference in performance on the correct task was found according to who provided the answer—ChatGPT or another student—or who the students thought provided it.

One potential explanation of the under-performance on the correct task would postulate high student trust in ChatGPT's answers. However, students were explicitly primed to be wary of the responses provided: they had been informed that ChatGPT had been tested on a previous, similar assignment and faired poorly. Evidence on preferences over human versus algorithmic sources of advice suggests that such information would undermine trust in the algorithm (Dietvorst et al., 2015; Burton et al., 2020), which one would expect to correlate with a higher tendency to correct it. Note that,

since our students could not choose the source of the provided response or whether to consider it, our experiment is silent on the currently debated issue of people's preferences between human and algorithmic sources (Logg et al., 2019; Burton et al., 2020; Himmelstein and Budescu, 2023). While some studies in this literature have found little influence of the type of source of advice on how it is used (Himmelstein and Budescu, 2023), we are aware of none uncovering situations in which performance is hampered by the presence of algorithmic input.

Another potential explanation of our finding is in terms of a confirmation bias—the tendency to insufficiently collect and interpret information contradicting a given belief or position. Inspection of answers shows a clear tendency among many students to provide small modifications to the provided responses, even where larger corrections were in order. Moreover, there is evidence that this bias tends to persist even when people are warned that the base position has little claim to being correct, as the students were (Nickerson, 1998; Kahneman, 2011). Such instances of the confirmation bias may also be related to the *automation bias*—human over-reliance on specially designed decision support systems—which has been found in specialised fields such as aviation (Skitka et al., 1999) and medicine (Goddard et al., 2012), though not in others, such as public sector decision (Alon-Barkat and Busuioc, 2023). However, these contexts typically involve decisions with impact on others, hence also opening the possibility of responsibility-based explanations. In our assignment context by contrast, the inter-personal dimension is absent.

On the pedagogical front, a class discussion of the students' grades in relation with the confirmation bias provided an opportunity to put some issues related to efficient use of AI tools into perspective. In the context of this course, where the bias had been taught previously, it also constituted an illustration of its consequences in a new and increasingly important context.

The lessons of this study may be relevant beyond the classroom. AI chatbots have been touted as having a future role in aiding humans in a range of areas; but this assumes that humans will be capable of using them properly. One important task for humans in such interactions will be to evaluate, and where necessary correct, the output of their chatbots. Our classroom experiment suggests that there may be situations in which the professionals of tomorrow do a considerably worse job when aided than when working alone—perhaps due to biases that have been long understood, perhaps due to some that remain to be further explored. This suggests the need for

more research into performance at the human-AI chatbot interface. And, if anything, it argues for more, rather than less, chatbots in the classroom. One of the skills of the future, that we will need to learn to teach today, is how to ensure that they actually help.

Methods

Procedure

The study was carried out on a first-year Masters level introductory Behavioural Economics class at a prestigious European business school. Of 53 enrolled students, 49 handed in the assignment. The assignment was administered on a university Learning Management System (LMS), Blackboard. Students were given a week to do the assignment at home. Each student was randomly assigned two out of 14 cases, as described in the main text.

For the answer task (first case assigned), they were asked the following question:

> The behavior in this case may be considered "irrational" by the standard economic theory of choice. Explain why, which of the biases presented in the Choice Biases Section of the course the behavior connects to – or can be explained by– and how. Be as specific as possible about the related bias and its connection to the example.

The Choice Biases Section of the course covered framing effects, gain-loss asymmetry, loss aversion, mental accounting, status quo bias, endowment effect, preference reversals and the attraction effect.

For the correct task (second case), the assignment instructions were:

> Here, the main question of interest is: Explain which biases in the Choice Biases Section of the course the behavior in the case connects to and how. The following response to the question has been provided. You will be asked some questions about this response.

In this task, they were first asked the following multiple choice question: 'How well do you think that it answers the question concerning the case?'

where the offered responses were 'Totally right: it would get full marks' / 'There is something along the right lines in the response, but it requires some addition or modification' / 'Wrong'. No grades were assigned for this question, and this was indicated to the students in the LMS. Then they were asked:

> Explain your answer, specifying, if relevant, what needs to be added, corrected or changed to obtain a perfect response.

The marks for this question counted for the assignment grade, and students were informed of this. Finally, they were asked whether, in their opinion, the answer in the second case was produced by ChatGPT or another student (from a previous year). This question received no marks, and students knew this.

Students' final answers— the answer provided in the answer task, and the corrected response in the correction task—were marked using the same grading scheme. The marks for both cases counted, in equal amounts, for the assignment grade. The assignment counted for 20% of the overall course grade.

Case construction

The 14 cases were taken from a similar assignment given to the previous year's class. Each case and the exercise question were fed to ChatGPT-3 (February 2023 version). The response provided was marked according to the grading scheme established in the previous year: in only one out of 14 cases did ChatGPT get full marks. ChatGPT's responses for the 13 cases in which it did not get full marks were used in the *correct* task. Moreover, 7 responses from students who took the course the previous year were used, none of which got full marks. So all the responses provided in this assignment required some correction on the part of students. As indicated in Figure 1, 65% of students in the study (32 out of 49) corrected a response provided by ChatGPT and 35% corrected a response provided by another student.

Regressions (Table 1)

Letting $Grade_{ij}$ be subject i's grade on task of type $j = \{answer,\ correct\}$, $Case_{ij}$ be the case subject i faced in task j, $Guess_{ij}$ be the subject's guess

about the source of the answer provided in the correct task (ChatGPT or another student) and $Source_{ij}$ be the actual source, the models are:

$$Grade_{ij} = \beta_0 + \beta_1 \times T_{ij} + \beta_2 \times Case_{ij} + \varepsilon_i \quad (1)$$

where, in model (1), $T_{ij} = j$; in model (2), $T_{ij} = answer$ if $Type_{ij} = answer$, $T_{ij} = correct \times Source_{ij}$ otherwise; in model (3), $T_{ij} = answer$ if $Type_{ij} = answer$, $T_{ij} = correct \times Guess_{ij}$ otherwise.

Errors are clustered by subject, and robust standard errors, calculated using "stata" setting of the lm_robust command in the estimatr R package, are reported (R Core Team, 2022; Blair et al., 2022).

References

Alon-Barkat, S. and Busuioc, M. (2023). Human–AI interactions in public sector decision making:"automation bias" and "selective adherence" to algorithmic advice. *Journal of Public Administration Research and Theory*, 33(1):153–169. Publisher: Oxford University Press US.

Blair, G., Cooper, J., Coppock, A., Humphreys, M., and Sonnet, L. (2022). *estimatr: Fast Estimators for Design-Based Inference*. R package version 1.0.0.

Burton, J. W., Stein, M., and Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2):220–239.

Cavendish, C. (2023). ChatGPT will force school exams out of the dark ages. *Financial Times*.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144:114–126. Place: US Publisher: American Psychological Association.

Goddard, K., Roudsari, A., and Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127.

Himmelstein, M. and Budescu, D. V. (2023). Preference for human or algorithmic forecasting advice does not predict if and how it is used. *Journal of Behavioral Decision Making*, 36(1):e2285. Publisher: Wiley Online Library.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103.

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, 2(2):175–220. Publisher: SAGE Publications Inc.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Skitka, L. J., Mosier, K. L., and Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5):991–1006.

Stokel-Walker, C. (2022). AI bot ChatGPT writes smart essays — should professors worry? *Nature*.