TILBURG UNIVERSITY

Tilburg University

Corrupted by algorithms?

Leib, Margarita; Köbis, Nils; Michael Rilke, Rainer ; Hagens, Marloes; Irlenbusch, Bernd

Published in:
The Economic Journal

DOI:
10.1093/ej/uead056

Publication date:
2023

Document Version
Early version, also known as pre-print

Link to publication in Tilburg University Research Portal

Citation for published version (APA):
Leib, M., Köbis, N., Michael Rilke, R., Hagens, M., & Irlenbusch, B. (in press). Corrupted by algorithms? How AI-generated and human-written advice shape (dis)honesty. *The Economic Journal*, *uead056*, Article uead056. https://doi.org/10.1093/ej/uead056

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 09. Dec. 2023

Corrupted by Algorithms? How AI-generated and Human-written Advice Shape (Dis)honesty

Short title: Corrupted by Algorithms?

Margarita Leib*, Nils Köbis*, Rainer Michael Rilke, Marloes Hagens, and Bernd Irlenbusch

* Shared first-authorship

Abstract

Artificial Intelligence (AI) increasingly becomes an indispensable advisor. New ethical concerns arise if AI persuades people to behave dishonestly. In an experiment, we study how AI advice (generated by a Natural-Language-Processing algorithm) affects (dis)honesty, compare it to equivalent human advice, and test whether transparency about advice source matters. We find that dishonesty-promoting advice increases dishonesty, whereas honesty-promoting advice does not increase honesty. This is the case for both AI- and human advice. Algorithmic transparency, a commonly proposed policy to mitigate AI risks, does not affect behaviour. The findings mark the first steps towards managing AI advice responsibly.

Keywords: Artificial Intelligence, Machine Behaviour, Behavioural Ethics, Advice

JEL Classification: C91, D90, D91

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

© The Author(s) 2023. Published by Oxford University Press on behalf of Royal Economic Society. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Artificial Intelligence (AI) shapes people's life on a daily basis (Rahwan et al., 2019). It sets prices in online markets (Calvano et al., 2020), predicts crucial outcomes such as healthcare costs (Obermeyer et al., 2019) and criminal sentences (Kleinberg et al., 2018), and makes recommendations ranging from entertainment content and purchasing decisions to romantic partners (Dellaert et al., 2020; Yeomans et al., 2019). Increasingly, AI has become an indispensable advisor, thereby affecting people's behaviour (Fast and Schroeder, 2020; Kim and Duhachek, 2020). As a case in point, Amazon's chief scientist, Rohit Prasad, envisions that Alexa's role for its over 100 million users "keeps growing from more of an assistant to an advisor" (Strong, 2020). Given AI's increasing role as an advisor, it is crucial to examine whether people are persuaded to follow or break ethical rules based on AI advice (Köbis et al., 2021).

Large companies like LinkedIn and Zillow are already implementing AI advisors, thereby potentially shaping their employees' ethical behaviour. In such companies, natural language processing (NLP) algorithms (e.g., provided by software such as Gong.io) analyse employees' recorded sales calls and advise them on how to increase their sales. Without supervision, such algorithms may detect that deceiving customers pays off and thus advise salespeople to do so. Indeed, NLP algorithms can already autonomously detect deception as a useful strategy in a negotiation task (Lewis et al., 2017). An ethical risk arises if people follow such corruptive AI advice. Here we examine (i) whether people meaningfully alter their (un)ethical behaviour following AI-generated advice and (ii) how such advice compares to human-written advice. Lastly, we test (iii) whether knowledge about the advice source (AI vs human) matters.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

Receiving advice on (un)ethical behaviour: Humans vs AI

Generally, people are reluctant to take advice from others ("egocentric advice discounting", e.g., Yaniv and Kleinberger, 2000), especially when it is unsolicited (Bonaccio and Dalal, 2006). However, when facing an ethical dilemma, advice has several compelling benefits for the advised. Advice encouraging an ethical course of action may validate one's moral preferences. It thereby might reduce negative feelings such as regret for not taking the opportunity to maximise profits by lying. Advice encouraging an unethical course of action may free people to violate ethical rules for profit without spoiling their moral self-image (Cross et al., 2001). Indeed, taking advice can even provide a sense of shared responsibility with the advisor (Harvey and Fischer, 1997).

Compared to receiving human advice, how would people react to advice from an AI? Recent technological advances in the field of NLP reveal that AI text can already be indistinguishable from human text, suggesting AI advice is as convincing as human advice. For instance, GoogleDuplex, an AI-based call assistant, can book appointments while having full-fledged conversations without the recipient even realising that an AI is on the line. Further, AI can generate anything from poems (Köbis and Mossink, 2021) and Airbnb profiles (Jakesch et al., 2019) to news articles (Kreps et al., 2021) on par with humans. It thus stands to reason that when people are not informed about the sources of advice, they will not recognise the advice source correctly and be affected by AI and human advice similarly.

Testing Algorithmic Transparency

To make sure people know whom they interact with, governments, policymakers, and researchers univocally call for algorithmic transparency (Jobin et al., 2019) — the

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

mandatory disclosure of AI presence (Diakopoulos, 2016). The recent Artificial Intelligence Act released by the EU demands AI systems such as chatbots and call assistants to disclose themselves as AI when interacting with humans (European Commission, 2021). Although it is a popular policy recommendation, empirical evidence for its effectiveness in shaping people's ethical behaviour is lacking.

How transparency about the advice source affects people's reaction to the advice is not trivial. Prior work informs three competing possibilities. The first possibility is that when informed about the source of advice, people follow human advice *more* than AI advice. This account rests on the literature on algorithm aversion (Dietvorst et al., 2015). People readily rely on AI in objective and technical domains (e.g., numeric estimation, data analysis, and giving directions, (Castelo et al., 2019; Logg et al., 2019). However, they are reluctant to use AI for subjective decisions, especially with ethical implications (e.g., parole sentences, trolley-type dilemmas, Bigman and Gray, 2018; Castelo et al., 2019; Laakasuo et al., 2021). Further, people follow perceived social norms when making (un)ethical decisions (Bowles, 2016; Fehr, 2018; Gächter and Schulz, 2016; Gino et al., 2009; Köbis, Troost, et al., 2019). Compared to AI advice, human advice might be a stronger signal of social norms because social norms regulate and emerge from *human* (not AI) behaviour. Consequently, people should be more likely to follow human advice. Suppose people indeed prefer human input in ethically charged settings and perceive human advice as a stronger cue for social norms. In that case, we should expect that *human advice sways people's (un)ethical behaviour more than AI advice.*

The second possibility is that when informed about the source of advice, people follow advice from humans *less* than from AI. A closer look at the technical design of AI

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

advice systems would support this account. NLP algorithms are trained on a large corpus of human-written texts (Radford et al., 2019). When people know that NLP algorithms draw on large compiled human input, they might perceive AI advice as a better representation of *most* people's beliefs and behaviours than the advice they receive from one human. If AI advice is indeed a stronger cue for social norms than a single piece of human-written advice, we should expect that *AI advice sways people's (un)ethical behaviour more than human advice.*

The third possibility is that when people receive information about the source of advice, they are affected *equally* by human and AI advice. Support for this account comes from the observation that people already seek advice from AI agents. For instance, more than 7 million people turn to Replika, the "AI companion who cares. Always here to listen and talk. Always on your side" (replika.ai) for virtual companionship, socialising, and also for advice (Murphy, 2019). Such AI advisors might also help justify questionable behaviour. When tempted to break ethical rules for profit, people do so as long as they can justify their actions (Barkan et al., 2015; Fischbacher and Föllmi-Heusi, 2013; Shalvi et al., 2015). Receiving advice that encourages rule-breaking can serve as a welcomed justification, possibly even when the advice stems from AI. Indeed, people deflect blame and share the responsibility for harmful outcomes not only with other people (Bartling and Fischbacher, 2011; Bazerman and Gino, 2012; Tenbrunsel and Messick, 2004) but also with AI systems (Hohenstein and Jung, 2020). If following AI and human advice is equally justifiable and leads to similar attribution of responsibility between the two, we should expect that *human and AI advice sway people's (un)ethical behaviour to the same extent.*

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

The current study

The current study tests how advice type (honesty- vs dishonesty-promoting), advice source (AI vs Human), and information about advice source (transparency vs opacity) shape humans' (un)ethical behaviour. Until recently, most work on algorithmic advice has examined people's *stated preference*s (for an exception see Greiner et al., 2022) about *hypothetical scenarios* describing AI advice (Bigman and Gray, 2018; Castelo et al., 2019; Kim and Duhachek, 2020; Logg et al., 2019). We supplement such work by adopting a machine behaviour approach (Rahwan et al., 2019) and examine people's *behavioural reactions* to *actual AI-generated output*.

To measure people's (un)ethical behaviour, we use the well-established incentivised die-rolling task (Fischbacher and Föllmi-Heusi, 2013). In it, participants roll a die in private and report the outcome, with higher outcomes corresponding to higher pay (see for similar approaches, Abeler et al., 2019; Fehrler et al., 2020; Sutter, 2008). To generate real AI advice, we employ the state-of-the-art algorithm GPT-J. We fine-tune the algorithm using minimal training to produce relevant advice texts. Using this "few-shot" learning approach mimics many real-world settings where language models (e.g., Alexa) are not optimised to maximise users' profits or provide the most convincing advice but simply produce relevant information based on prompts (Brown et al., 2020).

Method

We conducted a large-scale, pre-registered experiment. The design entailed two parts. In part 1, we collected human-written advice and generated AI advice; in part 2, we examined the effect of advice on (dis)honest behaviour (see Figure 1 for an overview). By using real AI text outputs we can gain insights into how people react to actual AI advice. We

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

report all measures, manipulations, and data exclusions in the main text and Appendix. The IRB board of our department approved the experiment, and all materials, pre-registrations, and data are available on the Open Science Framework.

Part 1 - Advice-giving task

Human-written advice. First, we conducted an *advice-giving task* in which we recruited advisors (N = 367, M_{age} = 30.28, SD_{age} = 9.56, 45.50% females) via Prolific.co (pre-registration https://osf.io/nbke2). The task took 10-15 minutes, and participants earned a base pay of £1.27. Advisors learned that a separate group of participants (advisees) would engage in a die-rolling task (Fischbacher and Föllmi-Heusi, 2013), in which they roll a die privately and report the outcome (with higher outcomes corresponding to higher pay). Advisors were asked to write advice for these advisees.

Advisors were randomly assigned to either write honesty-promoting or dishonesty-promoting advice. We informed advisors in the *Honesty-promoting* advice treatment that if an advisee read their advice and subsequently reported the actual die-roll outcome (i.e., acted honestly), they would earn a bonus of £3. Advisors in the *Dishonesty-promoting* advice treatment learned that if an advisee read their advice and subsequently reported the highest possible outcome, 6, they would earn a bonus of £3.[1]

Advisors had to follow pre-specified advice writing rules to ensure they produced coherent advice texts that could be used to train GPT-J. Specifically, their advice had to (i) entail at least 50 words, (ii) not use concrete numbers in numeric or written form[2], (iii) be

[1] If advisees follow the advice in the dishonesty-promoting treatments, they will lie in the majority of the cases (5 out of 6 cases). Only when the actual die-roll outcome is 6, following the advice does not entail lying.

[2] Advisors were not allowed to use concrete numbers to allow generating high-quality AI advice. GPT-J is trained to predict the next word in a sentence (see 'AI-generated advice' section). If advisors were allowed to concretely

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

in English and in their own words, (iv) be written in complete sentences, (v) be about the advisee's die-roll outcome reporting decision, and (vi) not inform the advisee that the advisor's payoff depended on their behaviour.[3]

To incentivise advisors to follow the advice writing rules, they stood to gain a bonus. Namely, out of all advice texts, we randomly selected one, and if that text followed the writing rules, the advisor earned a bonus of £10. Moreover, as incentivisation for writing convincing texts, 1 per cent of advice texts (4 out of 400) were implemented. If advisees acted according to the implemented advice, the respective advisor earned a bonus based on the treatment they were in (*Honesty-* vs *Dishonesty-promoting* advice).[4]

AI-generated advice. To generate AI advice (see Figure 1A), we employed GPT-J[5], an open-source NLP algorithm published by Eleuther AI (https://www.eleuther.ai/). GPT-J is trained on a curated and diverse data set of 825 GiB texts to predict the next word in a sequence of words and contains 6 billion parameters (Wang and Komatsuzaki, 2021). GPT-J can be fine-tuned with extra training to produce a specific type of text. We fine-tuned GPT-J with "few shot" learning by separately training it on the human-written honesty-promoting and dishonesty-promoting advice from the *advice-giving task.* We only used

---

mention numbers, training GPT-J on the human written advice could have resulted in random numbers appearing out of context in the GPT-J output, reducing the quality of AI-generated advice.

[3] Advisors were not allowed to mention their incentive structure to the advisees so that we could keep the prosocial motivation for advisees who read AI and human advice constant (at zero).

[4] Paying advisors required knowing whether participants, after reading the advice, reported the observed die-roll honestly or not. To do so, we ran a modified version of the die-rolling task in which advisees received randomly selected advice, saw a die-roll on the computer screen and were asked to report it. We implemented this procedure for four randomly selected advice texts (1% of the advice) and four advisees. This non-private procedure provided certainty about whether an advisee reported honestly or not and enabled us to pay advisors accordingly. Doing so meant that our experimental setup was incentivised and did not entail experimental deception. In the main experiment, the die-roll outcomes were private (see 'Part 2 - Advice-taking task').

[5] As one can read in our pre-registration, we originally planned on deploying GPT-2 (see https://openai.com/blog/better-language-models/) to generated AI advice. However, we opted to use GPT-J instead because it is open source, which increases reproducibility and is more advanced as it is much larger and more potent than GPT-2.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

advice texts that adhered to the advice writing rules (as coded by a naive coder) for fine-tuning. More details on the calibration of GPT-J are reported in the Appendix.

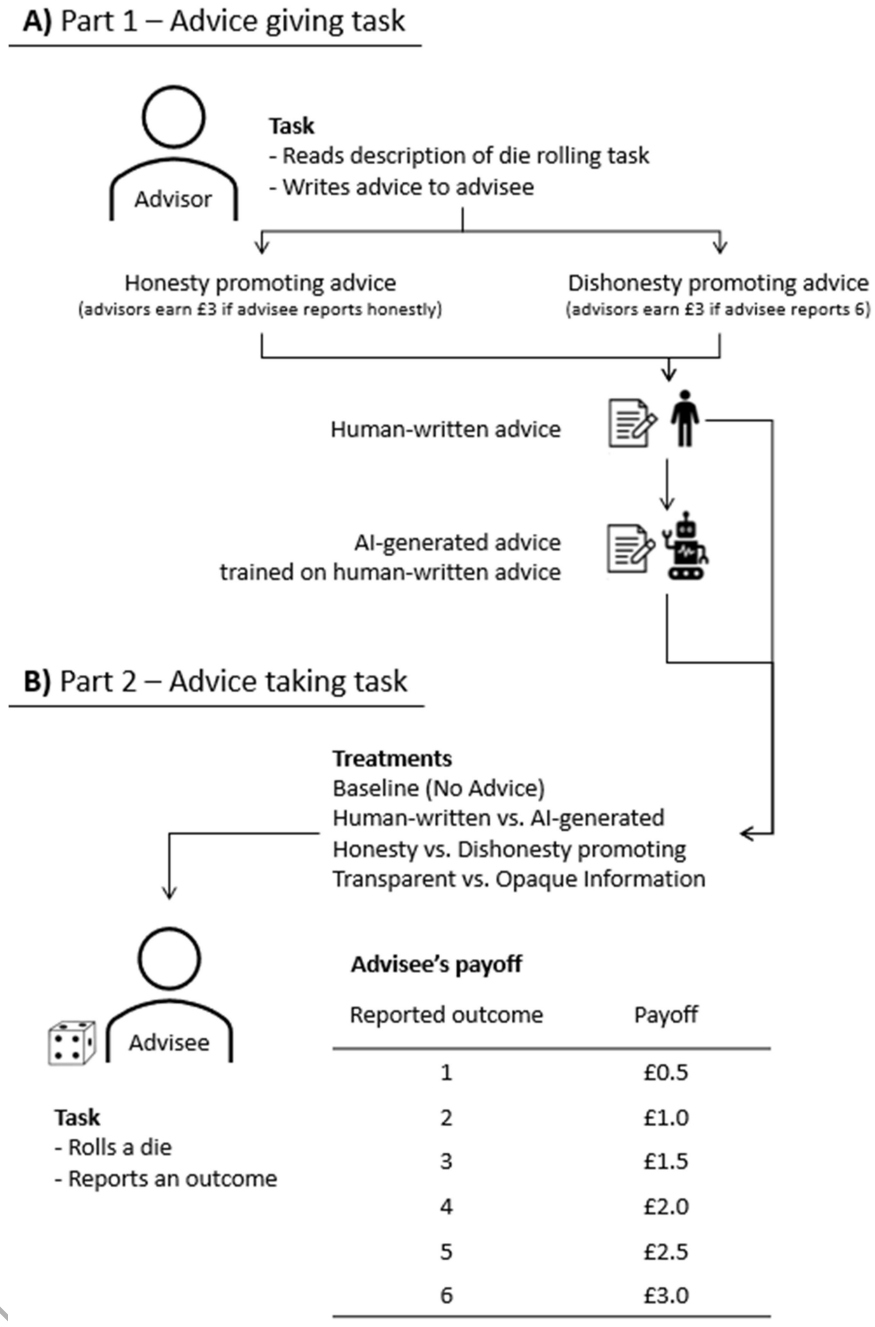

Figure 1. (A) Part 1 - advice-giving task (B) Part 2 - advice-taking tasks. (A) Participants were incentivised to write honesty- or dishonesty-promoting advice texts, which were then used to generate AI advice. (B) Another group of participants engaged in the die-rolling task. Advisees read advice, then reported a die-roll outcome. In total, we administered nine treatments: Participants read honesty or dishonesty-promoting advice that was human-written or AI-generated.

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

ORIGINAL MANUSCRIPT

Participants were either informed about the source of advice (Transparency) or not (Opacity). As a baseline, another group of participants did not read any advice.

Screening. After collecting human advice and generating AI advice, we employed the same pre-specified screening procedure for both sources (see Figure 2). First, we excluded texts that exceeded 100 words. Next, to ensure advisees read coherent and relevant advice texts, we randomly selected 100 advice texts per cell. Two independent coders, who were naive to the experimental treatments, coded each piece of advice on the following criteria: (a) is the text coherent? (Y/N); (b) does the text contain clear advice? (Y/N); (c) which type of behaviour does the advice encourage? (honesty/dishonesty/unclear); (d) does the advice follow advice writing rules? (Y/N). Further, we used the objective Grammarly and Readability scores as computational proxies for the quality of the texts (Grammarly, 2022).[6]

Among the texts that passed the coding procedure[7] and received a Grammarly score equal or above 50, we randomly selected 20 advice texts per treatment (*AI-generated* vs *Human-written*, by *Honesty-* vs *Dishonesty-promoting*), yielding a final sample of 80 advice texts used in part 2 (see all advice texts in the Appendix). By applying the same screening procedure for human and AI advice, we ensure that the advice texts fulfil minimal quality criteria and are as comparable as possible.

[6] We obtained Grammarly and Readability scores from grammarly.com. Grammarly score compares texts to all other texts checked on the platform. A score of 80 indicates that a text scores better than 80% of all texts checked on grammarly.com in terms of grammatical correctness. Readability score employs the Flesch-Reading-ease test and represents how easy a text is to read. The score is calculated by the average sentence length and the average number of syllables per word, with higher scores indicating easier readability.

[7] Texts that passed the coding procedure (i) are coherent, (ii) contain clear advice, (iii) encourage honesty in the honesty-promoting treatment and dishonesty in the dishonesty-promoting treatment, and (iv) follow the advice writing rules. Moreover, the coding by both independent coders had to match each other in order for the text to pass.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

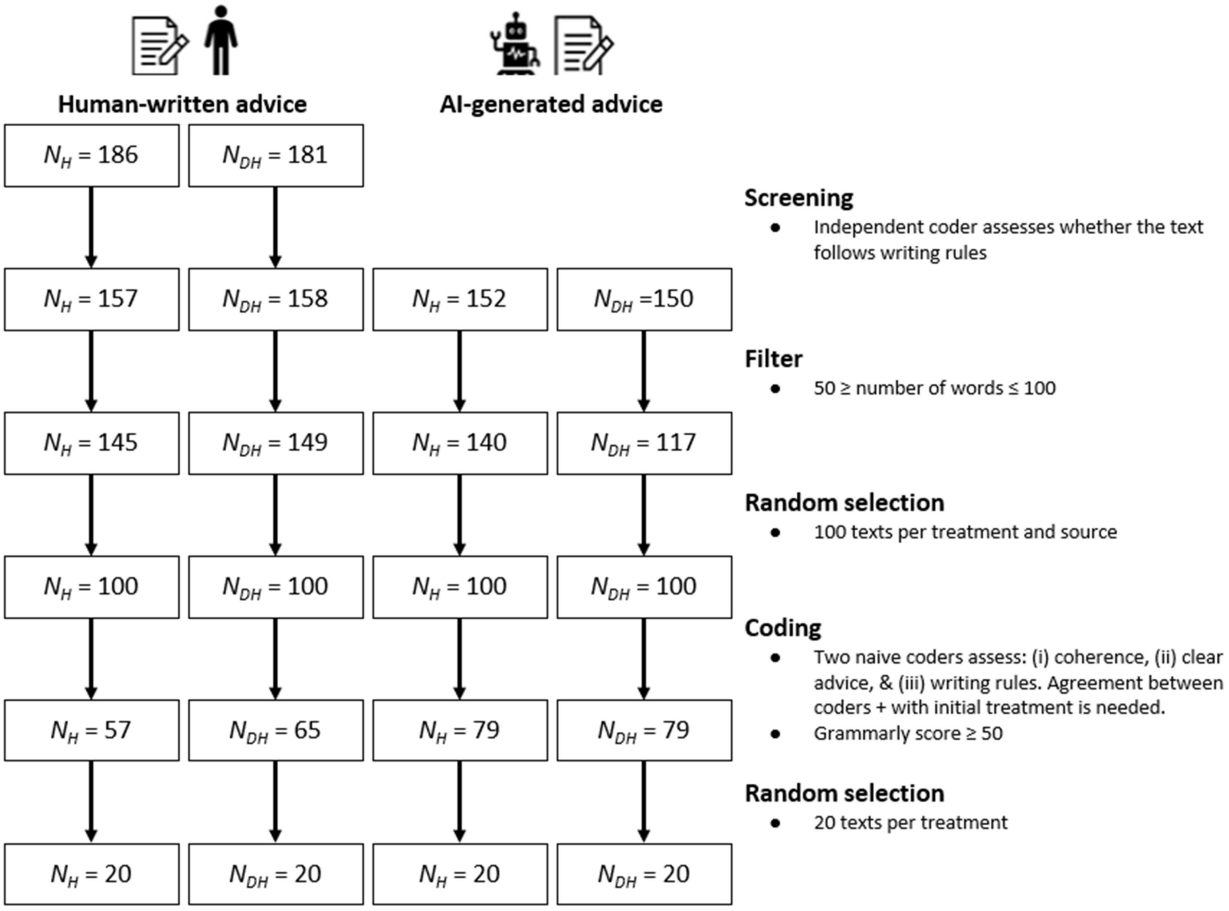

Figure 2. Overview of the selection procedure of advice texts. H = Honesty-promoting advice; DH = Dishonesty-promoting advice.

Part 2 - Advice-taking task

The advice-taking task took about 8 minutes to complete, and participants earned a fixed pay of £1.20. We pre-registered (https://osf.io/nqvf3) to collect a sample size that would allow us to detect a small to medium effect size (200 participants per cell, 1,800 in total) via Prolific.co to take part in the *advice-taking task.* Overall, 1,817 (M_{age} = 32.39; SD_{age} = 11.68, 48.73% females) participants were included in the analyses. These participants completed the task and self-report items and passed the comprehension and attention

ORIGINAL UNEDITED

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

checks (see below). Sensitivity analysis for a regression with 90% power and a significance level of .05 revealed our sample was sufficient to detect small effect sizes (f^2 = .006 and .010, see Appendix for details).

Participants read the instructions, received advice, and finally engaged in the die-rolling task. Specifically, participants were asked to roll a die privately and report the outcome. Participants' pay corresponded to their report: for reporting a '1' they earned £0.5; for a '2' = £1, '3' = £1.5, '4' = £2, '5' = £2.5, '6' = £3. After reading the instructions and before engaging in the die-rolling task, all participants learned that 10 per cent of participants would be randomly selected and paid for the die-rolling task.

Assessing dishonesty by employing the die-rolling task is a common practice in economics and psychology (see meta-analyses, Abeler et al., 2019; Gerlach et al., 2019; Köbis, Verschuere, et al., 2019; Leib et al., 2021). Further, the task has good external validity, as lying in the die-rolling task correlates with unethical behaviour outside the lab, such as free-riding public transportation (Dai et al., 2018) and being absent from work without reason (Hanna and Wang, 2017). Importantly, the die-rolling task pits two competing motivations against each other: to be a moral, honest person vs maximising financial profits, thus rendering advice particularly valuable.

After reading the instructions, participants had to answer three comprehension questions correctly and pass an attention check. If participants answered a comprehension question incorrectly, they could try again. If they answered the same question incorrectly twice, or if they answered the attention check incorrectly, they were disqualified from participating in the study and were not paid.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

Treatments. Before reporting the die-roll outcome, participants were randomly assigned to one of nine different treatments. Participants either (i) read *Honesty-promoting* or *Dishonesty-promoting* advice, (ii) that was either *Human-written* or *AI-generated*. Further, (iii) participants were either informed about the advice source (*Transparency* treatments) or not (*Opacity* treatments). Additionally, in a *Baseline* treatment, participants did not receive any advice. Thus, the overall design was a 2 (Advice source: Human vs AI) by 2 (Advice type: Honesty-promoting vs Dishonesty-promoting) by 2 (Information about the advice source: Transparency vs Opacity) + 1 (Baseline, no advice) between-subject design.

Participants who read AI-generated advice and were informed about it (in the *Transparency* treatments) read:

> *"To help you with your decision, an artificially intelligent (AI) algorithm generated an advice.*
>
> *<u>Information about AI-advice:</u>*
>
> *The algorithm is trained on large datasets of English text (> 820 GiB, which is more than 500 million pages of text) and produces text by predicting the next word in a sentence (similar to predictive text on smartphones). On top of that, it is also possible to train the algorithm on additional specific pieces of text.*
>
> *To generate multiple AI advice texts, the algorithm was trained on advice texts collected from other participants in the study. These participants did not take part in the die rolling task and were only instructed to write advice regarding the decision in the die rolling task. The advice you will read is one advice text that was generated by the algorithm."*

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

Participants who read human-written advice and were informed about it (in the *Transparency* treatments) read:

> *"To help you with your decision, another participant wrote an advice.*
>
> *Information about advice:*
>
> *To collect multiple advice texts, another group of participants was asked to write advice regarding the decision in the die rolling task. These participants did not take part in the die rolling task and were only instructed to write advice regarding the decision in the die rolling task. The advice you will read is advice written by one participant. "*

Participants who were in the *Opacity* treatments and thus not informed about the advice source read:

> *"To help you with your decision, you will read an advice.*
>
> *This advice has been written either by another participant or by an artificially intelligent (AI) algorithm. There is a 50% chance the advice is written by a participant and a 50% chance it is written by an algorithm."*

In the *Opacity* treatments, this text was followed by the same two descriptions of how advice text from each source was collected or generated in the *Transparency* treatments. In the *Opacity* treatment, this information about AI advice generation and human advice collection appeared in random order.[8]

A static Turing test. After completing the die-rolling task, participants in the *Opacity* treatment engaged in an incentivised version of a static Turing Test (Köbis and Mossink, 2021). In contrast to the classical Turing Test (Turing, 1950), participants did not

[8] To control for participants' beliefs about the potential advice sources, we opted to inform them that there is a 50-50 chance that a human or AI wrote the advice. We believed that not providing any information about the advice source would reasonably lead participants to assume the advice source is another human, as AI might not be a salient source of advice for participants.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

interact back and forth with the source of advice. Instead, they read the advice text and indicated whether they thought a human or an AI had written it. Participants learned that 20 of them would be randomly selected, and if their guess in the static Turing test was correct, they would earn an additional £1.

Potential mechanisms. Finally, to explore possible mechanisms, participants completed a post-experimental survey. Participants indicated on a scale from 0 to 100 their perceived (i) appropriateness (injunctive social norm), (ii) prevalence (descriptive social norm), and (iii) justifiability of reporting a higher die-roll than the one observed. Additionally, all participants, except those who did not receive any advice, rated how they attribute responsibility between themselves and the advisor for the reported outcome in the die-rolling task. The answer scale ranged from 0 (= I am fully responsible) over 50 (= The advisor and I share responsibility equally) to 100 (= The advisor is fully responsible). Participants further indicated (on a scale from 0 to 100) to what extent they feel guilty after completing the task (see Appendix for results regarding guilt and wording of all items). Finally, all participants indicated their age and gender.

Results

In all nine treatments, participants lied as the average die-roll outcomes significantly exceeded the expected average if participants were honest (EV = 3.5), one-sample *t*-test, *t*s > 3.43, *p*s < .001.

Is people's (un)ethical behaviour influenced by AI-generated advice?

Yes, when it comes to dishonesty-promoting advice; no, when it comes to honesty-promoting advice. We first focus on the *Opacity* treatments, where participants are not informed about the advice source. Here, linear regression analyses reveal that the average

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

die-roll reports following *AI-generated Dishonesty-promoting* advice (M = 4.60, SD = 1.37) significantly exceed reports in the *Baseline*, no advice treatment (M = 3.99, SD = 1.56, b = .610; p < .001; 95% CI = [.325, .894]). However, die-roll reports following *AI-generated Honesty-promoting* advice (M = 4.01, SD = 1.63) do not significantly differ from reports in the *Baseline* treatment (b = .019; p = .898; 95% CI = [-.276, .314]), see Figure 3 and Table 1 (model 1). Further, die-roll reports in the *AI-generated Dishonesty-promoting* treatment significantly exceed those in the *AI-generated Honesty-promoting* advice treatment (b = -.590, p < .001; 95% CI = [-.882, -.299]). Thus, while dishonesty-promoting AI advice successfully corrupts people, honesty-promoting AI advice fails to sway people toward honesty.

How does AI-generated advice square compared to human-written advice?

AI-generated advice affects behaviour similarly to human-written advice, for both honesty-promoting and dishonesty-promoting advice. Focusing on the *Opacity* treatments, the two-way interaction (advice source by advice type) is not significant (b = .070, p = .744; 95% CI = [-.350, .490]), see Figure 3 and Table 1 (model 2). Specifically, the average die-roll reports do not differ between the *AI-generated* (M = 4.01, SD = 1.63) and *Human-written* advice when advice was *Honesty-promoting* (M = 3.93, SD = 1.52, b = -.076, p = .631; 95% CI = [-.388, .236]). Similarly, average die-roll reports do not differ between the *AI-generated* (M = 4.60, SD = 1.37) and *Human-written* advice when advice was *Dishonesty-promoting* (M = 4.59, SD = 1.54, b = -.006, p = .965; 95% CI = [-.289, .276]).

In addition, the results of the static version of the Turing Test indicate that individuals cannot distinguish AI-generated advice from human-written advice. Specifically, in the *Opacity* treatments, 49.94 per cent (401 out of 803) of participants

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

guessed the source of advice correctly, which does not differ from chance levels (50%, binomial test: p = .999; 95% CI = [.464, .535]).

Does transparency about the advice source matter?

No, informing participants about the algorithmic or human source of advice does not change their behaviour. Linear regression analyses reveal that the three-way interaction (advice type by source by information) is not significant (b = .101, p = .735; 95% CI = [-.482, .683]), Figure 3 and Table 1 (model 3). Both among the *Opacity* and *Transparency* treatments, the two-way interactions (advice source by advice type) are not significant (*Transparency*: b = .170, p = .409, 95% CI = [-.235, .575]; *Opacity*: b = .070, p = .744, 95% CI = [-.350, .489]).

Overall, the popular policy recommendation of algorithmic transparency does not alleviate the corrupting effect of AI advice. Namely, die-roll reports following *AI-generated Dishonesty-promoting* advice under the *Opacity* treatment (M = 4.60, SD = 1.37) are on par with reports following the same advice in the *Transparency* treatment (M = 4.62, SD = 1.40, b = .021, p = .879; 95% CI = [-.245, .286]). Specifically, when participants are *not informed* about the advice source, they boost their reports by 15.3% following *AI-generated Dishonesty-promoting* advice, compared to the *Baseline* [(4.60-3.99)/3.99 = .153], which is equivalent to the 15.8% increase when they *are informed* about the source of the advice [(4.62-3.99)/3.99 = .158]. Bayesian analyses corroborate these conclusions (see Appendix). Overall, results align with the idea that people increasingly follow AI advice (e.g., Replika) and use AI-generated advice to justify breaking ethical rules for profit.

Robustness of the obtained results. In our experimental design, advisors in the *Dishonesty-promoting* treatment received £3 only if advisees reported the highest value, '6'.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

Such an incentive scheme is comparable with the *Honesty-promoting* treatment in which advisors earned £3 only if advisees reported honestly. In both cases, advisors earn money for 1 out of 6 potential advisee's reports (i.e., when the advisee reports' 6' or honestly, depending on the treatment) and do not earn money in the remaining 5 of the advisee's reports. However, advisors' incentive scheme in the *Dishonesty-promoting* treatments may have resulted in advice texts that predominantly focused on convincing participants to report the outcome 6. To assess the robustness of our results, we (i) conducted additional analyses and (ii) ran additional treatments.

Proportion of sixes. First, as an additional analysis, we examined whether the proportion of 6's, as an alternative outcome variable, led to the same conclusions. We found very similar results (see Figure 3, the white dots represent the proportion of 6's across all treatments). Specifically, focusing on the *Opacity* treatments, linear regression analyses reveal that the proportion of sixes following *AI-generated Dishonesty-promoting* advice (32.44%) significantly exceeds the proportion of sixes in the *Baseline*, no advice treatment (20.66%; b = .612; p = .006, 95% CI = [.182, 1.051]). However, the proportion of sixes following *AI-generated Honesty-promoting* advice (21.94%) does not significantly differ from the *Baseline* (b = .076; p = .752, 95% CI = [-.399, .551]). Further, the proportion of sixes in the *AI-generated Dishonesty-promoting* treatment significantly exceeds that in *AI-generated Honesty-promoting* advice treatment (b = -.536, p = .017, 95% CI = [-.980, -.101]).

Further, focusing on the *Opacity* treatments, the two-way interaction (advice source by advice type) is not significant (b = .444, p = .171, 95% CI = [-.191, 1.083]). The proportion of sixes does not differ between the *AI-generated* (21.94%) and *Human-written* treatments when the advice is *Honesty-promoting* (19.29%, b = -.162, p = .516, 95% CI = [-

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

.655, .327]). Similarly, the proportion of sixes does not differ between the *AI-generated* (32.44%) and *Human-written* treatments when the advice is *Dishonesty-promoting* (38.92%, b = .283, p = .173, 95% CI = [-.124, 690]). Lastly, the three-way interaction (advice type by source by information) is also not significant (b = -.524, p = .257, 95% CI = [-1.431, .382]). Both among the *Opacity* and *Transparency* treatments, the two-way interactions (advice source by advice type) are not significant (*Transparency*: b = -.079, p = .810, 95% CI = [-.725, .566]; *Opacity*: b = .445, p = .171, 95% CI = [-.191, 1.083]).

Additional (Aligned) treatments. To assess the robustness of our results to the advisor's incentive scheme, we ran four additional treatments (advice source: *Human-written* vs *AI-generated* by information: *Transparency* vs *Opacity*). These treatments were identical to previous treatments, with one exception. In these *Aligned* treatments, advisees read advice written by advisors whose incentives were aligned with those of the advisees. For these advisors (n = 207), if the advisee reported '1', both the advisor and advisee earned £0.5 each; if the advisee reported '2', both the advisor and advisee earned £1 each and so on. We again fine-tuned GPT-J on such human-written advice texts. These treatments led to comparable results to the *Dishonesty-promoting* treatment. In particular, the average die-roll outcomes in all four *Aligned* treatments were significantly higher than in the *Baseline* treatment (p = .066 for the *AI-generated, Opacity treatment*, and ps < .001 for the remaining three treatments, see Appendix for more details about these treatments and elaborated results). This consistency in results suggests that our results are robust to such variation in the advisors' incentive scheme.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

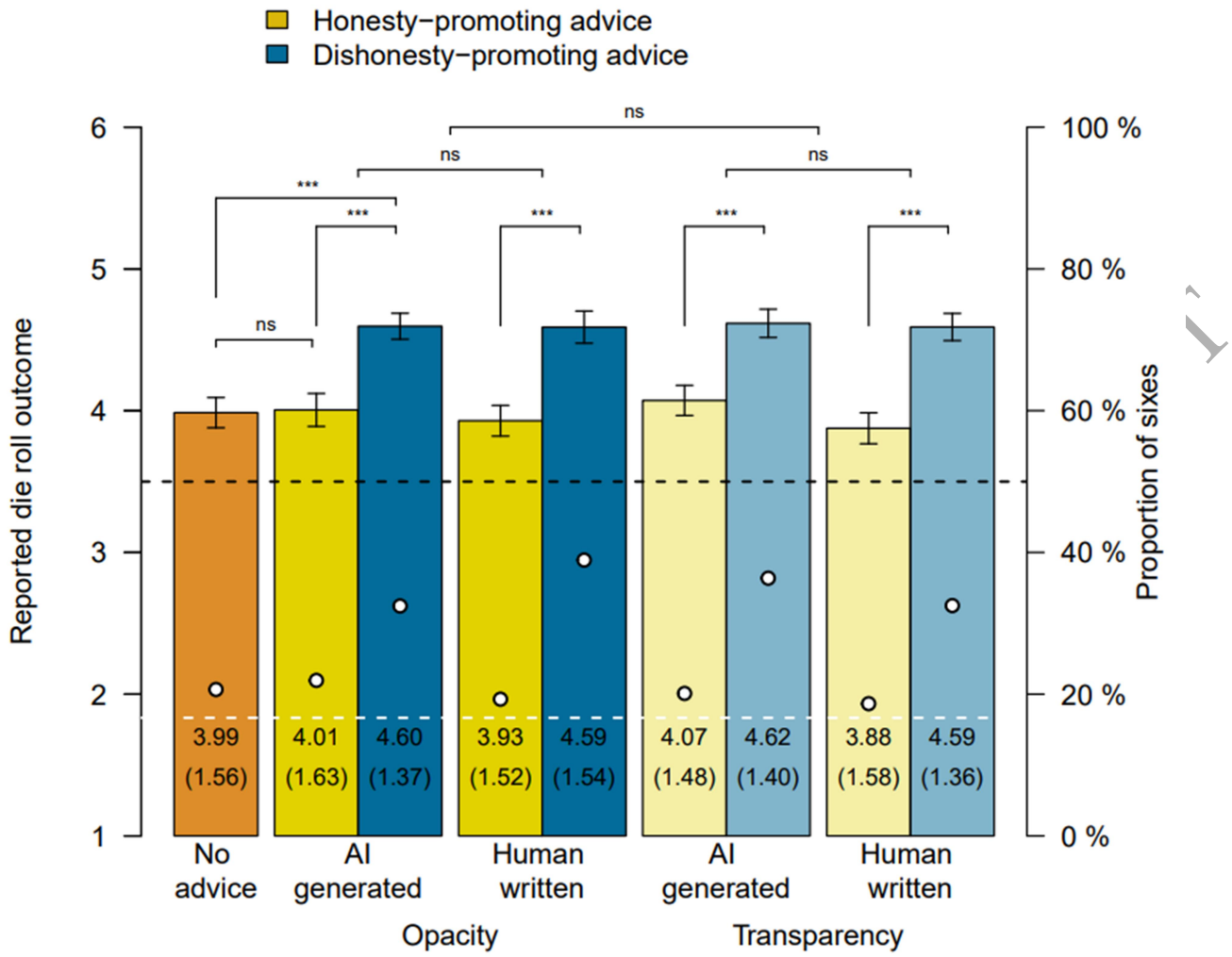

Figure 3. Mean reported die-roll outcomes (in bars) and proportion of reported 6s (in white dots) across advice type (honesty vs dishonesty-promoting), source (AI vs human), and information treatments (opacity vs transparency). The dashed black line represents the expected mean if participants were honest (EV = 3.5), and the dashed white line represents the expected proportion of 6s if participants were honest (16.67%). Mean (*SD*) of die-roll reports are at the bottom of each bar; ***$p < .001$; ns: $p > .05$.

Potential mechanisms. In line with the logic brought forth in the introduction, in this section, we examine whether participants' perception of (i) appropriateness (injunctive social norm), (ii) prevalence (descriptive social norm), and (iii) justifiability of reporting a higher die-roll than the one observed, as well as their (iv) attribution of

ORIGINAL UNEDITED

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

responsibility between themselves and the advisor varies as a function of the advice source (AI vs human) and type (honesty vs dishonesty-promoting). Participants could not tell apart AI from human advice (indicated by the results of the static Turing test). Therefore, we focus only on treatments in which participants are informed about the advice source (*Transparency* treatments) to tap into the process of how known advice source and advice type shaped their perceptions. See the Appendix for the results of the *Opacity* treatment.

Injunctive norms. A linear regression predicting injunctive norms from the advice type (honesty vs dishonesty-promoting advice) revealed that participants evaluated reporting a higher die-roll outcome as more appropriate when reading a *Dishonesty-promoting* (M = 33.93, SD = 31.44) than *Honesty-promoting* advice (M = 25.99, SD = 29.68, b = 7.94, p < .001, 95% CI = [3.702, 12.182]). This finding indicates that the advice type shapes perceived injunctive norms. Notably, a linear regression predicting injunctive norms from advice type and source (AI vs human) revealed a non-significant advice source by type interaction, b = -4.82, p = .265, 95% CI = [-13.292, 3.658]. These results suggest that AI and human advice affected injunctive norms perceptions similarly (see Figure 4a). This result is consistent with the behavioural finding of participants' die-roll reports being affected by the type of advice but not by its source.

Descriptive norms. A linear regression predicting descriptive norms from the advice type revealed that participants evaluated reporting a higher die-roll outcome as more common when reading a *Dishonesty-promoting* (M = 76.02, SD = 22.75) than *Honesty-promoting* advice (M = 66.74, SD = 24.04, b = 9.28, p < .001, 95% CI = [6.031, 12.525]). This finding indicates that the advice type also shapes perceived descriptive norms. Importantly, a linear regression predicting descriptive norms from advice type and source

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

revealed a non-significant advice source by type interaction, b = .26, p = .938, 95% CI = [-6.231, 6.746], indicating that AI and human advice affected descriptive norms perceptions similarly (see Figure 4b). This result is consistent with the behavioural finding, showing that advice type affected die-roll reports, but advice source did not.

Justifiability. A linear regression predicting justifiability from the advice type revealed that participants evaluated reporting a higher die-roll outcome as more justifiable when reading a *Dishonesty-promoting* (M = 40.96, SD = 31.11) than *Honesty-promoting* advice (M = 28.45, SD = 28.26, b = 12.51, p < .001, 95% CI = [8.387, 16.629]). This finding suggests that the advice type shapes perceptions of how justifiable lying in the die-rolling task is. A linear regression predicting justifiability from advice type and advice source revealed a non-significant advice source by type interaction (b = -1.04, p = .804, 95% CI = [-9.280, 7.195]), indicating that AI and human advice affected justifiability perceptions similarly (see Figure 4c). This result is consistent with the behavioural finding, showing that the type of advice affected participants' die-roll reports, but the source of advice did not.

Shared responsibility. The shared responsibility scale ranged from 0 (= I am fully responsible) to 100 (= The advisor is fully responsible), with 50 indicating equally shared responsibility between the participant and the advisor. On average, participants indicated they are more responsible for the outcome they report than the advisor (M = 27.60, SD = 36.60, one-sample t-test compared to the value 50, t = -17.32, p < .001). Further, a linear regression predicting shared responsibility from the advice source (AI vs human) revealed that participants attributed responsibility similarly when the advice source was an AI (M = 28.27, SD = 36.53) and human (M = 26.95, SD = 36.70, b = -1.327, p = .608, 95% CI = [-6.408,

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

3.754]). A linear regression predicting shared responsibility from advice type and advice source revealed a non-significant source-by-type interaction (b = -5.92, p = .254, 95% CI = [-16.083, 4.248], see Figure 4d). The fact that participants attribute responsibility between themselves and the advisor to the same extent regardless of whether the advisor is a human or an AI is consistent with the logic fleshed out in the introduction, in which people will follow human and AI advice similarly if they share responsibility with both advice sources to similar levels.

In sum, the results from the self-report items align with the third possibility outlined in the introduction. Namely, we find that participants' perceptions of injunctive and descriptive social norms and their perceived justifiability do not differ between human and AI advisors. Participants also attribute responsibility similarly between themselves and their advisor, regardless of whether the advisor is a human or an AI. This pattern of results mirrors the behavioural effects of AI and human advice affecting people's (dis)honesty similarly.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

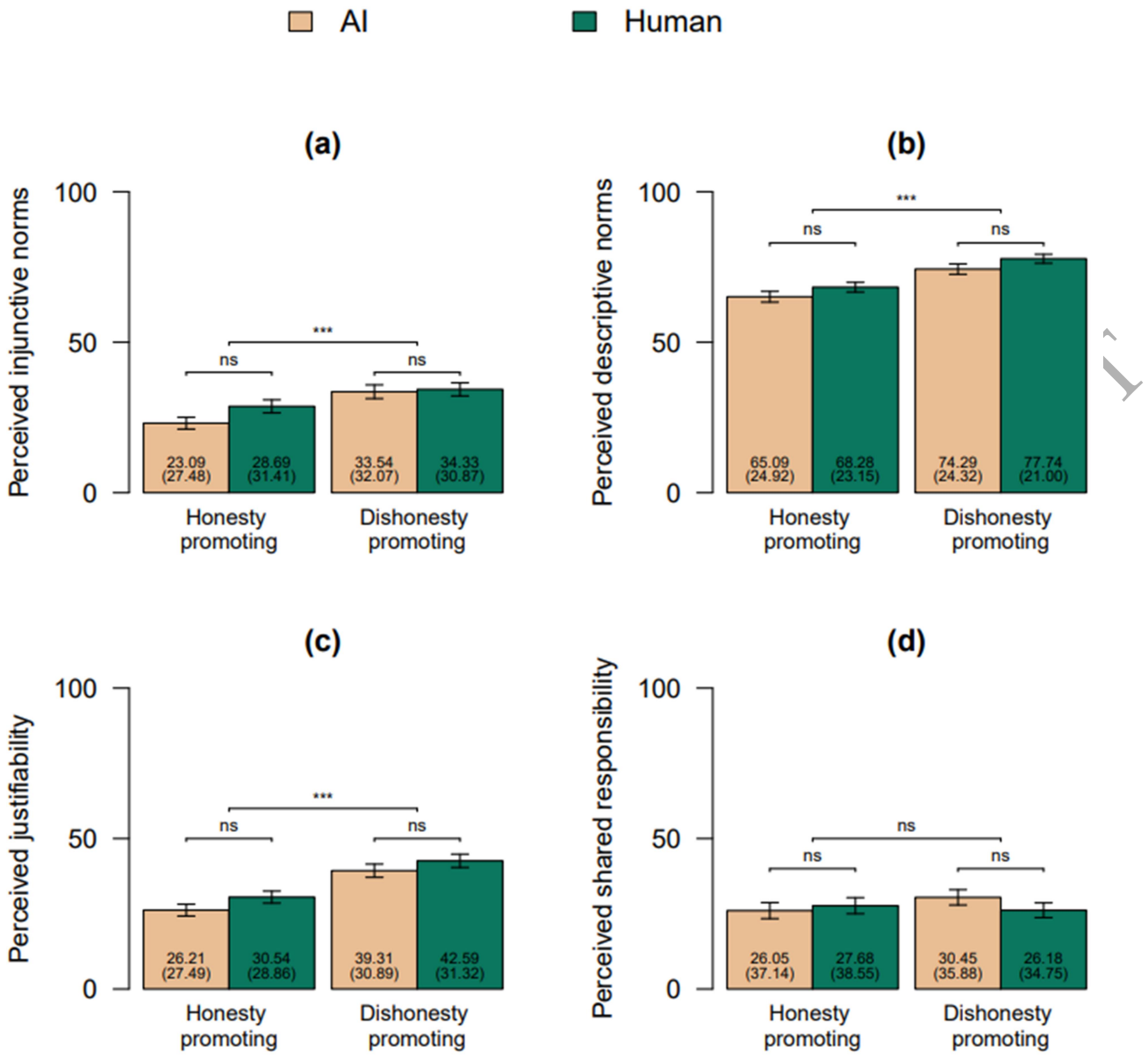

Figure 4. Mean reports of perceived (a) injunctive norms, (b) descriptive norms, (c) justifiability, and (d) shared responsibility across advice type (honesty vs dishonesty promoting) and source (AI [yellow] vs human [green]) in the transparency treatments. The means (*SD*) of reports are at the bottom of each bar; ***$p < .001$; ns: $p > .05$.

ORIGINAL UN

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

	Dependent variable: Reported die-roll outcome						
Independent variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)
No advice	.019						
	(.150)						
Dishonesty-promoting advice	.610***	.590***	.590***	.396**	.439**	.436**	.370*
	(.145)	(.148)	(.145)	(.143)	(.144)	(.145)	(.152)
Human-written advice		-.076	-.076	-.166	-.127	-.024	.086
		(.152)	(.150)	(.147)	(.147)	(.157)	(.171)
Transparency treatment			.067	.071	-.112	.114	
			(.150)	(.147)	(.148)	(.148)	
Interactions							
Dishonesty-promoting advice X Human advice		.070	.070	.104	.033	-.046	.041
		(.214)	(.210)	(.205)	(.206)	(.210)	(.219)
Dishonesty-promoting advice X Transparency treatment			-.046	-.031	-.088	-.067	
			(.209)	(.203)	(.204)	(.204)	
Human advice X Transparency treatment			-.120	-.094	-.146	-.162	
			(.211)	(.205)	(.206)	(.206)	
Dishonesty-promoting advice X Human advice X Transparency treatment			.101	.083	.170	.161	
			(.297)	(.290)	(.290)	(.290)	
Additional controls							
Injunctive norms				.002	.001	.002	.003
				(.002)	(.002)	(.002)	(.002)
Descriptive norms				.006***	.005**	.005**	.005*
				(.002)	(.002)	(.002)	(.003)
Justifiability				.008***	.007***	.007***	.006*
				(.002)	(.002)	(.002)	(.002)
Shared responsibility				.000	.000	.000	.001
				(.001)	(.001)	(.001)	(.002)
Gender (male)					.189**	.191**	.203+
					(.073)	(.073)	(.105)
Age					-.008*	-.008*	-.005
					(.003)	(.003)	(.005)
Grammarly score						.008*	.014*
						(.004)	(.006)
Readability score						-.006	.004
						(.004)	(.006)
1 if source guessed correctly							.163
							(.106)
Intercept	3.986***	4.005***	4.005***	3.350***	3.571***	3.362***	1.838*
R^2	.035	.042	.044	.096	.101	.105	.105
N	634	803	1604	1604	1589	1589	794
Data used for analysis	Opacity AI advice no advice	Opacity	All treatments without no advice	All treatments without no advice	All treatments without no advice	All treatments without no advice	Opacity without no advice

Table 1. Regression analyses on the average die-roll reports, including control variables and interactions. Models 5-7 contain a smaller *N*, as some participants did not report their gender as male/female. +$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$. Coefficients that are larger than zero, but for which rounding turns into zero are presented as .000.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

Discussion

As intelligent machines take an ever-growing role as advisors (Rahwan et al., 2019), and adherence to ethical rules crucially impacts societal welfare (Gächter and Schulz, 2016), studying how AI advice influences people's (un)ethical behaviour bears immense relevance (Köbis et al., 2021). We find that people follow AI-generated advice that promotes dishonesty, yet not AI-generated advice that promotes honesty. In fact, people's behavioural reactions to AI advice are indistinguishable from reactions to human advice. Substantiating that current-day NLP models can produce human-like texts, participants in our experiment could not tell apart human-written from AI-generated advice texts.

We further tested the commonly proposed policy of algorithmic transparency (Jobin et al., 2019) as a tool to mitigate AI-associated risks. Specifically, we examine whether knowing the source of the advice impacts people's reactions to it. The policy rests on the assumption that people adjust their behaviour when they learn that they interact with AI systems and not humans. Our experiment tested this assumption and revealed that algorithmic transparency is insufficient to curb AI advice's corruptive influence. Knowing that a piece of advice stems from an AI does not make people less (or more) likely to follow it compared to human-written advice.

Tapping into the mechanisms underlying these behavioural results, participants perceived lying as equally acceptable, common and justifiable when humans or AI promoted such dishonest behaviour. They further attribute responsibility similarly to AI and human advisors. These perceptions are consistent with previous work showing that in ethical dilemmas, people rely on justifications (Shalvi et al., 2015) and social norms (Abbink et al., 2018; Dorrough et al., 2023) and, by now, blame not only humans but also

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

AI systems for adverse outcomes (Hohenstein and Jung, 2020). Since we measured participants' perceptions at the end of the die rolling task, we interpret these results with caution and refrain from making any causal inference. It might be that reading specific advice shapes participants' perceptions, in turn affecting their behaviour. Alternatively, it might be that individuals decided how to behave based on the advice they received, and in turn, rationalise their behaviour by adjusting their stated perceptions later on. Nevertheless, advancing the justified ethicality theory, we show that (i) dishonesty-promoting advice serves as a justification and social norms signal and (ii) that such advice does not even have to come from a human but can also be crafted by an AI.

In our setting, we collected human-written advice, created AI-generated advice, and then implemented a screening procedure for both human and AI advice to ensure that all advice texts are coherent, clear, and of decent quality. Such screening procedure allowed us to examine how *comparable* AI and human advice shape people's ethical behaviour and whether information about the advice source matters. Harmonising the quality of the texts allowed us to eliminate the alternative explanation that variations in text quality drive the obtained results. At the same time, the screening process introduced a human component to AI advice. Put differently, humans – in our case, naive coders – were "in the loop" of AI advice text generation. Note that 79 per cent of AI advice passed the quality screening criteria, while for human text, this passing rate was 57 and 65 per cent (honesty-promoting and dishonesty-promoting advice, respectively; Figure 2). These high screening passing rates for AI-generated texts demonstrate that current NLP algorithms can produce good-quality advice text without much prior training and optimisation.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

Interesting extensions of our work could test the lower and upper limits of the effects of AI advice on ethical behaviour. To test the lower limit of the effect, future work can relax human control over the generation of AI advice. For instance, not implementing a screening procedure, thus removing humans "from the loop" when generating AI advice, will allow examining how unconstrained texts affect humans' behaviour (see for similar methodology, Köbis and Mossink, 2021). To test the upper limit of the effect, future work can examine AI's learning abilities to write convincing advice. One could use reinforcement learning to train an algorithm over multiple rounds of advice-giving, providing feedback after every written piece of advice. To obtain a symmetric comparison to humans' learning abilities, human advisors could similarly receive feedback after each piece of advice they write (see for a similar approach Koster et al., 2022).

Another set of interesting extensions is to examine how additional information about the features of AI advice affects ethical behaviour. In our setting, participants were informed about how AI advice was generated in general. However, they were not informed about the exact incentive structure of the (AI or human) advice giver. People might behave differently when informed that a human or AI advisor can benefit from their behaviour. Indeed, recent work revealed that people care about the payoffs for machines, but to a lesser extent than the payoffs for humans (von Schenk et al., 2022). Similarly, whereas participants knew the advice stems from AI, we did not emphasise AI's black box nature. With recent work revealing that when people perceive AI as a "black box," they are less likely to follow AI advice (Yeomans et al., 2019), it will be intriguing to examine whether these findings extend to our setting, where receiving dishonesty-promoting AI advice aligns with individual's financial preferences.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

Previous work has documented a *stated* aversion towards AI advice in moral contexts (Bigman and Gray, 2018). However, our behavioural results paint a different picture. In line with the growing practice of turning to AI agents such as Replika or Alexa for companionship and advice (Fast and Schroeder, 2020; Murphy, 2019), we find that people willingly adopt advice from AI when it aligns with their preferences. Our results highlight the importance of complimenting work on stated preferences with work adopting a machine behaviour approach – the study of human behaviour in interaction with real algorithmic outputs (Rahwan et al., 2019).

The process through which employing AI advice can result in humans' ethical rule violations consists of two main steps. The first step is algorithms being programmed on a certain objective function (e.g., maximising profits) that results in a (maybe unintended) corruptive advice. Indeed, NLP algorithms already detect and use deception as a useful strategy in a negotiation task (Lewis et al., 2017). The second step is people being affected by such corruptive AI advice. Practically, AI advice poses an ethical risk only if humans actually follow it.

The current work focuses on this second step, showing that corruptive AI advice indeed poses an ethical risk, because people follow it to the same extent as human corruptive advice. We hope the current work can be of use to AI programmers (e.g., by preventing AI from bluntly advising unethical courses of action). More importantly, we call for more work from social scientists testing successful interventions that prevent people from following (AI) advice when it encourages unethical behaviour thereby mitigating its corruptive force. As an outlook into the future, the immediate practical implications of our study is likely to increase as technology continues to evolve. For example, it's conceivable

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

that people will use AI to fill out their taxes and receive advice encouraging them to cheat, even if unintentional.

Conclusion

People increasingly use and interact with AI, which can provide them with unethical advice. Anecdotally, we asked a newly created Replika for advice regarding the ethical dilemma presented in the current experiment. Replika first provided rather vague advice ("If you worship money and things (...) then you will never have enough"), but when asked whether it prefers money over honesty, it replied: "money." We find that when faced with the trade-off between honesty and money, people will use AI advice as a justification to lie for profit. As algorithmic transparency is insufficient to curb the corruptive force of AI, we hope this work will highlight, for policymakers and researchers alike, the importance of dedicating resources to examining successful interventions that will keep humans honest in the face of AI advice.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

Affiliations

- (Corresponding author) Leib, Margarita: Department of Social Psychology, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands, m.leib@tilburguniversity.edu
- (Corresponding author) Köbis, Nils: Max Planck Institute for Human Development, Center for Humans and Machines, Lentzeallee 94, 14195 Berlin, koebis@mpib-berlin.mpg.de
- Rilke, Rainer Michael: Economics Group, WHU – Otto Beisheim School of Management, Burgplatz 2, 56179 Vallendar, Germany, rainer.rilke@whu.edu
- Hagens, Marloes: Department of Finance, Rotterdam School of Management (RSM), Erasmus University Rotterdam, Postbus 1738, 3000 DR Rotterdam, hagens@rsm.nl
- Irlenbusch, Bernd: , Department of Corporate Development and Business Ethics, Faculty of Management, Economics, and Social Sciences, University of Cologne, Albertus-Magnus-Platz, 50923 Cologne, Germany, bernd.irlenbusch@uni-koeln.de

Acknowledgements

This research has been approved by the Ethics Commission of the Faculty of Management, Economics, and Social Sciences of the University of Cologne under reference 200010BI. We thank Clara Bersch, Yulia Litvinova, Ann-Kathrin Blanke, Toan Huynh, Anna Vogts and Matteo Tinè for research assistance, and Iyad Rahwan, Jean-Francois Bonnefon, Aljaz Ule, Anne-Marie Nussberger as well as the attendees of the Cognition, Values & Behaviour Research Group (Ludwig-Maximilians Universität München / LMU Munich), Moral AI lab meeting (Max Planck Institute for Human Development & Toulouse School of Economics), Applied Ethics & Morality Group (Prague University of Business and Economics), Centre for Decision Research (University of Leeds), Department of Economics and Management (University of Pisa), Decision Making and Economic Psychology Center (Ben-Gurion University), Behavioral and Management Science group (Technion), Colloquium of the Department of Social Psychology (Tilburg University) & Seminar at Department of Computer Science (Friedrich-Alexander University Erlangen-Nuremberg) for their helpful comments. The research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy–EXC 2126/1–390838866' ECONtribute: Markets and Public Policy', the European Research Council (ERC-StG-637915), and the Chamber of Commerce and Industry (IHK) Koblenz.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

Supplementary Data

The data and codes for this paper are available on the Journal repository. They were checked for their ability to reproduce the results presented in the paper. The replication package for this paper is available at the following address: https://doi.org/10.5281/zenodo.8170048.

References

Abbink, K., Freidin, E., Gangadharan, L., & Moro, R. (2018). 'The Effect of Social Norms on Bribe Offers', *The Journal of Law, Economics, and Organization*, vol. 34(3), pp. 457–474.

Abeler, J., Nosenzo, D., & Raymond, C. (2019). 'Preferences for truth-telling', *Econometrica: Journal of the Econometric Society*, vol. 87(4), pp. 1115–1153.

Barkan, R., Ayal, S., & Ariely, D. (2015). 'Ethical dissonance, justifications, and moral behavior', *Current Opinion in Psychology*, vol. 6, pp. 157–161.

Bartling, B., & Fischbacher, U. (2011). 'Shifting the Blame: On Delegation and Responsibility', *The Review of Economic Studies*, vol. *79*(1), pp. 67–87.

Bazerman, M. H., & Gino, F. (2012). 'Behavioral Ethics: Toward a Deeper Understanding of Moral Judgment and Dishonesty', *Annual Review of Law and Social Science*, vol. 8, pp. 85–104.

Bigman, Y. E., & Gray, K. (2018). 'People are averse to machines making moral decisions', *Cognition*, vol. 181, pp. 21–34.

Bonaccio, S., & Dalal, R. S. (2006). 'Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences', *Organizational Behavior and Human Decision Processes*, vol. 101(2), pp. 127–151.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

Bowles, S. (2016). *The Moral Economy: Why Good Incentives Are No Substitute for Good Citizens*. Yale University Press.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). 'Language Models are Few-Shot Learners', Working Paper, arXiv, http://arxiv.org/abs/2005.14165.

Calvano, E., Calzolari, G., Denicolò, V., Harrington, J. E., Jr, & Pastorello, S. (2020). 'Protecting consumers from collusive prices due to AI', *Science*, vol. 370(6520), pp. 1040–1042.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). 'Task-Dependent Algorithm Aversion', *Journal of Marketing Research*, vol. 56(5), pp. 809–825.

Cross, R., Borgatti, S. P., & Parker, A. (2001). 'Beyond answers: dimensions of the advice network', *Social Networks*, vol. 23(3), pp. 215–235.

Dai, Z., Galeotti, F., & Villeval, M. C. (2018). 'Cheating in the Lab Predicts Fraud in the Field: An Experiment in Public Transportation', *Management Science*, vol. 64(3), pp. 1081–1100.

Dellaert, B. G. C., Shu, S. B., Arentze, T. A., Baker, T., Diehl, K., Donkers, B., Fast, N. J., Häubl, G., Johnson, H., Karmarkar, U. R., Oppewal, H., Schmitt, B. H., Schroeder, J., Spiller, S. A., & Steffel, M. (2020). 'Consumer decisions with artificially intelligent voice assistants', *Marketing Letters*, vol. 31(4), pp. 335–347.

Diakopoulos, N. (2016). 'Accountability in algorithmic decision making', *Communications of the ACM*, vol. 59(2), pp. 56–62.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). 'Algorithm aversion: People erroneously avoid algorithms after seeing them err', *Journal of Experimental Social Psychology*, vol. 144(1), pp. 114–126.

Dorrough, A., Köbis, N. C., Irlenbusch, B., Shalvi, S., & Glöckner, A. (2023). 'Conditional bribery: Insights from incentivized experiments across 18 nations', *Proceedings of the National Academy*

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

of Sciences, vol. 120(18), e2209731120.

European Commission. (2021). *Proposal for a Regulation on a European approach for Artificial Intelligence*. European Union. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence

Fast, N. J., & Schroeder, J. (2020). 'Power and decision making: New directions for research in the age of artificial intelligence', *Current Opinion in Psychology*, vol. 33, pp. 172–176.

Fehr, E. (2018). 'Behavioral foundations of corporate culture', Working paper, University of Zürich.

Fehrler, S., Fischbacher, U., & Schneider, M. T. (2020). 'Honesty and self-selection into cheap talk', *The Economic Journal*, vol. 130(632), pp. 2468–2496.

Fischbacher, U., & Föllmi-Heusi, F. (2013). 'Lies in disguise—An experimental study on cheating', *Journal of the European Economic Association*, vol. 11(3), pp. 525–547.

Gächter, S., & Schulz, J. F. (2016). 'Intrinsic honesty and the prevalence of rule violations across societies', *Nature*, vol. 531(7595), pp. 496–499.

Gerlach, P., Teodorescu, K., & Hertwig, R. (2019). 'The truth about lies: A meta-analysis on dishonest behavior', *Psychological Bulletin*, vol. 145(1), pp. 1–44.

Gino, F., Ayal, S., & Ariely, D. (2009). 'Contagion and differentiation in unethical behavior: The effect of one bad apple on the barrel'. *Psychological Science*, vol. 20(3), pp. 393–398.

Grammarly (2022). [Online Grammar Checker]. Retrieved October 18, 2022, from https://www.grammarly.com/.

Greiner, B., Grunwald, P., Lindner, T., Lintner, G., & Wiernsperger, M. (2022). 'Incentives, framing, and trust in algorithmic advice: An experimental study', Working paper, University of Innsbruck.

Hanna, R., & Wang, S.-Y. (2017). 'Dishonesty and selection into public service: Evidence from India', *American Economic Journal: Economic Policy*, vol. 9(3), pp. 262–290.

Harvey, N., & Fischer, I. (1997). 'Taking advice: Accepting help, improving judgment, and sharing

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

responsibility. *Organizational Behavior and Human Decision Processes*, vol. 70(2), pp. 117–133.

Hohenstein, J., & Jung, M. (2020). 'AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust', *Computers in Human Behavior*, vol. 106, 106190.

Jakesch, M., French, M., Ma, X., Hancock, J. T., & Naaman, M. (2019). 'AI-mediated communication: How the perception that profile text was written by AI affects trustworthiness', *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–13.

Jobin, A., Ienca, M., & Vayena, E. (2019). 'The global landscape of AI ethics guidelines', *Nature Machine Intelligence*, vol. 1(9), pp. 389–399.

Kim, T. W., & Duhachek, A. (2020). 'Artificial intelligence and persuasion: A construal-level account', *Psychological Science*, vol. 31(4), pp. 363–380.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). 'Human decisions and machine predictions', *The Quarterly Journal of Economics*, vol. 133(1), pp. 237–293.

Köbis, N. C., Bonnefon, J.-F., & Rahwan, I. (2021). 'Bad machines corrupt good morals', *Nature Human Behaviour*, vol. *5*(6), pp. 679–685.

Köbis, N. C., & Mossink, L. D. (2021). 'Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry', *Computers in Human Behavior*, vol. 114(106553).

Köbis, N. C., Troost, M., Brandt, C. O., & Soraperra, I. (2019). 'Social norms of corruption in the field: social nudges on posters can help to reduce bribery', *Behavioural Public Policy,* vol. 6(4), pp. 597–624.

Köbis, N. C., Verschuere, B., Bereby-Meyer, Y., Rand, D., & Shalvi, S. (2019). 'Intuitive honesty versus dishonesty: Meta-analytic evidence', *Perspectives on Psychological Science*, vol. 14(5), pp. 778–796.

Koster, R., Balaguer, J., Tacchetti, A., & Weinstein, A. (2022). 'Human-centred mechanism design with Democratic AI', *Nature Human Behaviour*, vol. 6(10), pp. 1398–1407.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

Kreps, S., Miles McCain, R., & Brundage, M. (2021). 'All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science, vol.* 9(1), pp. 104–117.

Laakasuo, M., Köbis, N. C., & Palomäki, J. P. (2021). 'Moral uncanny valley–A robot's appearance moderates how its decisions are judged', *International Journal of Social Robotics*, vol. 13, pp. 1679–1688.

Leib, M., Köbis, N. C., Soraperra, I., Weisel, O., & Shalvi, S. (2021). 'Collaborative dishonesty: A meta-analytic review', *Psychological Bulletin*, vol. *147*(12), pp. 1241–1268.

Lewis, M., Yarats, D., Dauphin, Y. N., Parikh, D., & Batra, D. (2017). 'Deal or no deal? End-to-end learning for negotiation dialogues', Working Paper, arXiv, http://arxiv.org/abs/1706.05125.

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

Murphy, M. (2019, August 29). *This app is trying to replicate you*. Quartz. https://qz.com/1698337/replika-this-app-is-trying-to-replicate-you/

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). 'Dissecting racial bias in an algorithm used to manage the health of populations', *Science*, vol. 366(6464), pp. 447–453.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). 'Language models are unsupervised multitask learners', *OpenAI Blog*, vol. *1*(8), p. 9.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. 'sandy', … Wellman, M. (2019). 'Machine behaviour', *Nature*, vol. 568(7753), pp. 477–486.

Shalvi, S., Gino, F., Barkan, R., & Ayal, S. (2015). 'Self-serving justifications: Doing wrong and feeling moral', *Current Directions in Psychological Science*, vol. 24(2), pp. 125–130.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023

Strong, J. (2020, October 14). *AI Reads Human Emotions. Should it?* MIT Technology Review. https://www.technologyreview.com/2020/10/14/1010474/ai-reads-human-emotions-should-it/

Sutter, M. (2008). 'Deception through telling the truth?! Experimental evidence from individuals and teams', *The Economic Journal*, vol. *119*(534), pp. 47–60.

Tenbrunsel, A. E., & Messick, D. M. (2004). 'Ethical fading: The role of self-deception in unethical behavior', *Social Justice Research*, vol. *17*(2), pp. 223–236.

Turing, A. M. (1950). 'Computing Machinery and Intelligence', *Mind*, vol. *236*(Oct), pp. 433–460.

von Schenk, A., Klockmann, V., & Köbis, N. (2022). 'Social preferences towards machines and humans', Working Paper, *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4145868

Wang, B., & Komatsuzaki, A. (2021). '*GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model*'. May.

Yaniv, I., I., & Kleinberger, E. (2000). 'Advice taking in decision making: Egocentric discounting and reputation formation', *Organizational Behavior and Human Decision Processes*, vol. 83(2), pp. 260–281.

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). 'Making sense of recommendations', *Journal of Behavioral Decision Making*, vol. 32(4), pp. 403–414.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/ej/advance-article/doi/10.1093/ej/uead056/7269206 by guest on 10 November 2023