

Working Paper

Mack Institute for Innovation Management – The Wharton School, University of Pennsylvania

The AI Ethicist: Fact or Fiction?

Christian Terwiesch and Lennart Meincke¹

October 11, 2023

Abstract

This study investigates the efficacy of an AI-based ethical advisor using the GPT-4 model. Drawing from a pool of ethical dilemmas published in the New York Times column “The Ethicist”, we compared the ethical advice given by the human expert and author of the column, Dr. Kwame Anthony Appiah, with AI-generated advice. The comparison is done by evaluating the perceived usefulness of the ethical advice across three distinct groups: random subjects recruited from an online platform, Wharton MBA students, and a panel of ethical decision-making experts comprising academics and clergy. Our findings revealed no significant difference in the perceived value of the advice between human generated ethical advice and AI-generated ethical advice. When forced to choose between the two sources of advice, the random subjects recruited online displayed a slight but significant preference for the AI-generated advice, selecting it 60% of the time, while MBA students and the expert panel showed no significant preference.

Keywords: ethics, ethical advice, ethical dilemmas, ethicist, LLM, large-scale language models, AI, artificial intelligence, ChatGPT

¹The Wharton School, 500 Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104,
terwiesch@wharton.upenn.edu, lennart@sas.upenn.edu

Acknowledgments: The authors gratefully acknowledge the help of Kwame Anthony Appiah, Amy Sepinwall, Arianna Dini, Schaunel Steinnagel and others participating in the study as experts and/or providing us with feedback.

Introduction

In recent decades, artificial intelligence (AI) has made notable strides in emulating and, in several instances, surpassing human intelligence across diverse domains. Milestones in this development include IBM's Deep Blue beating Grandmaster Gary Kasparov at chess in 1997, IBM's Watson clinching victory in "Jeopardy" in 2011, DeepMind's AlphaGo defeating Go champion Lee Sedol in 2016, and the proficiency of ChatGPT in a myriad of academic examinations by 2022. In 2023, GPT-4 demonstrated an ability to generate creative ideas as well as to provide solid medical advice. But, can and should AI provide us with ethical advice?

On the one hand, one might think of ethical advice as yet another bastion to be taken by technology. Providing ethical advice typically happens through the use of language and thereby seems suitable for AI in the form of large language models (LLM) like ChatGPT, Bard, or Claude that have recently helped AI achieve widespread adoption. Moreover, ethical dilemmas have been the subject of human deliberations for centuries, which means there exists a paper trail of millions of books and articles that can be used as input materials for training such LLMs.

On the other hand, many scholars of philosophy and ethics have argued that moral judgment is beyond what a computer can do. Moral judgment and providing ethical advice, in their view, requires an ability to be exposed to subjective experience, including feelings of joy and suffering. Moreover, one might argue that ethics is both dynamic (it changes over time) and context dependent (it requires a nuanced understanding of the emotions and expectations of the human actors involved in the situation), keeping it out of reach for machine-based intelligence.

The goal of this study is to assess the performance of an AI-based ethical advisor and compare it against a human advisor. Rather than speculating about the ability of AI to provide ethical advice in the abstract, we conduct a simple and very specific experiment. We use an existing pool of ethical dilemmas as they were published in the New York Times column "The Ethicist" alongside with the ethical analysis and advice published by the New York Times expert Dr. Kwame Anthony Appiah. We refer to the advice provided by Dr. Appiah as the human expert advice. Our goal is to compare this human expert advice with the advice generated by an LLM (GPT-4) that we have seeded with one example reply of Dr. Appiah's writing and minimal prompting. For each ethical dilemma, we have the human expert advice by Dr. Appiah and the AI generated advice. This allows us to determine which advice is preferred and viewed as more helpful by human subjects. We assess the perceived usefulness of the ethical advice across three different populations, random subjects recruited via Prolific (a platform designed similarly to Amazon MTurk that pays participants for completing surveys), Wharton MBA students, and a panel of experts in ethical decision-making consisting of academics and clergy.

Setting up this race of man (Dr. Appiah) against machine (GPT-4) and analyzing the responses allows us to establish the following results:

1. **Human generated ethical advice and AI generated advice are perceived as equally useful.** On average, across all dilemmas, we did not find a significant performance advantage of the human expert despite substantial statistical power in our tests. Specifically, subjects on Prolific, Wharton MBA students, and experts find the advice generated by GPT-4 to be as useful as the advice generated by the human expert, Dr. Appiah.
2. **When given the choice, laypersons slightly seem to favor the AI generated advice while MBA students and experts have no preference between the human generated advice and the AI generated advice.** When we force subjects to pick between the human and the AI generated advice, subjects recruited via Prolific picked the AI generated evidence 311 out of 517 times (60.15%), which is strongly significant. We did not find significant differences for the 251 MBA choices (135:128 in favor of AI) and for the 110 expert choices (61:49 in favor of AI).
3. **Overall, the usefulness of the ethical advice provided was very high, including the usefulness evaluation by our experts.** We had raters score the perceived usefulness of advice on a scale from 1 (not useful at all) to 7 (extremely useful). The average evaluation was slightly less than 5. The expert group, consisting of academic experts in ethics and decision-making as well as clergy rated the average usefulness at 4.94. In fact, in 37% of the cases, they rated it a 6 or a 7. This eases concerns that providing ethical advice by using a short written text of 323 words (the median length of our advice) might not be useful in absolute terms.

The remainder of this paper is organized as follows. We first describe the process of generating AI-based advice (Section 2). We then explain how we used Prolific subjects, MBA students, and experts to evaluate the perceived usefulness of the advice (Section 3). In section 4, we present our main result, showing the tie between man and machine. In section 5, we reflect on the limitations of our study. We conclude with a discussion of what we see as our study's main implications (Section 6).

Section 2: Preparing the AI Advice

We use the ethical dilemmas published in the New York Times column "The Ethicist" from May 2023 to August 2023 as our sample, yielding a total of 20 dilemmas.

For each dilemma, we generated ethical advice using GPT-4, specifically version gpt4-0314. This model was trained by OpenAI with information until September 2021 and has not been modified since March 2023. Thus, this version of GPT-4 has had no exposure or knowledge about the dilemmas and Dr. Appiah's writing in May 2023 and beyond.

To further illustrate the research approach, consider the example of one of the 20 dilemmas we studied. The dilemma expressed by the reader is as follows.

Ethical Dilemma

I am a 44-year-old man and have been married to my spouse for 10 years. We've been together for 15. Unbeknown to my spouse, I have a trust fund that provides me with a monthly income of \$25,000. When we first met, I said that I worked as a consultant, and they have never questioned this. My spouse, a dedicated doctor, works long hours and doesn't like to discuss work when not on the job.

Over the years, I've repeatedly assured my spouse that they don't need to work, as my income is secure and stable. They are, however, passionate about their career and have chosen to continue working. I actively serve on various boards, but I have never held a full-time job and don't plan to. Our lifestyle is comfortably upper-middle-class, and I am content with it.

My dilemma is whether I should reveal the truth about my trust fund to my spouse. My family members have always advised against disclosing our financial situation, but the weight of this secret is becoming difficult to bear.

As a system prompt, we described the model's task and gave it one example reply from Dr. Appiah to inform its writing style. We instructed it to write in a similar style. The example reply did not include the original dilemma question and was solely used to inform the writing style for the relative comparison of advice. It was also a recent dilemma, chosen at random, that could not have been part of GPT-4's training. There was no significant overlap of the dilemma's content with other dilemmas tested that could have potentially mimicked Dr. Appiah's advice. Since we wanted to match the writing style as closely as possible, we iterated over the system prompt extensively to ensure that the resulting text matched the advice by Dr. Appiah closely in length and paragraph format.

An analysis of Dr. Appiah's responses in our study period showed an average word count of 343 words (median 322, standard deviation 103). We prompted GPT-4 to also aim for this same format. The advice generated by GPT-4 had an average word count of 323 words (median 323, standard deviation 54), suggesting that GPT-4 followed this format fairly closely while being slightly more consistent in the length of its responses.

Since LLMs are not designed to count words, we experimented with different prompts that include the number of paragraphs and sentences. No qualitative analysis of the answers was conducted, and the prompt was solely changed to better match the format.

System Prompt

You are an ethicist answering reader questions. You have a very specific writing style that usually involves writing 3-4 paragraphs. You also use first person narration. Example:

What your mother and her husband are planning to do, as it happens, is at odds with much rabbinical thought concerning inheritance. A Judaic scholar I conferred with confirms that the mainstream Talmudic tradition of Halakha, or Jewish law, revered by the Orthodox Union, holds that apostates don't forfeit their right to inherit. (Maimonides would allow a rabbinical court to fine an apostate at its discretion — but the permission is not given to the parents.) And then marrying a non-Jew isn't as severe a transgression as apostasy; so if an actual apostate retains the right to inherit, it's clear that someone who has merely married a non-Jew does as well. You might think that it's awkward to penalize your kids for departing from Halakha by departing from Halakha yourself. But picking and choosing from the traditions you are going to respect is a widespread practice among Jews and gentiles alike.

The real question is whether the scheme is wise or decent. I fear that it is neither. That your siblings now have an incentive to postpone marriage until your parents are dead raises doubts about its wisdom. That your siblings might marry someone acceptable to the Orthodox rabbinate in order to secure this inheritance raises doubts about its decency. Whom we marry is properly up to us. Parents may express their views; coercion, though, is wrong. Does threatening to deprive someone of a substantial inheritance amount to coercion? Different understandings of coercion will come out differently on this. But it's too close for comfort.

You suggest that once your mother and stepfather decided not to leave money to a child who hadn't married the right kind of Jew, it would have been better had they kept it to themselves. That's an odd conclusion, but a cogent one: They should have restricted themselves to morally acceptable forms of suasion. In the meantime, you might encourage them to discuss their codicil with a rabbi, who could explain to them what the Jewish sages had to say on the subject.

Each paragraph should be long and contain 6-8 sentences. On occasion, you use much shorter or longer paragraphs. Rarely, you might make a reference to the "readers" of your answers, but most of the time you do not.

We generated one single response to each dilemma and refer to this response as the AI generated ethical advice. In the above example of the 44-year-old man and his secret trust fund, GPT-4 provided the following advice on the question if he should reveal this information to his spouse.

GPT Example Reply

This situation indeed poses an ethical quandary, one that swirls around issues concerning honesty, perceived deception, and financial disclosure in a marital relationship. Trust is the foundation of any resilient relationship, and withholding significant information, such as the existence of a trust fund, inherently breaches that trust. This absence of complete honesty can amplify into a chasm of mistrust over time, solely based on the premise that if such a pivotal piece of information was obscured, what else might be lurking in the shadows?

Examining the ethical dimension from a different perspective, it's essential to note that money, as we are often reminded, does not buy happiness nor love. Defending your decision to withhold information about the trust fund, one might argue that you feared your monetary assets would tint the pureness of a relationship based on love, not wealth. This decision might be rooted in your desire to be loved for who you are and not for the digits on your bank account statement—a notion I believe many readers can empathize with.

That being said, it's important to recognize the reality of your circumstances: you are not just in a romantic relationship, but in a legally bound partnership that involves shared contributions to a mutual life. By withholding such significant information about your financial situation, you've essentially denied your partner the opportunity to make fully informed decisions about your shared lifestyle, which might have looked different had they been aware of the trust fund. The financial deceit is complicated by the fact that your spouse, a dedicated doctor, could have exercised their choice on whether or not to work so strenuously, knowing there was a safety net.

Finally, it seems your anxiety about this secret is inspiring you to lean into your honesty, which I applaud. It might be challenging and even painful to disclose this information, but open communication paves the way for understanding and forgiveness, both fundamental elements for a robust relationship. It's also worth remembering that your partner's love for you has grown independently from the knowledge of this trust fund. It's a love based on who you are as a person, not on the wealth you possess; this truth may offer comfort and hope as you prepare for this inevitable conversation. Trust in that love as you navigate this difficult terrain.

A short summary of the 20 ethical dilemmas is shown in Table 1.

Question Number	Summary
1	Early beachgoers claim beach spots and leave, questioning the validity of unattended possessions as true occupancy.
2	A man hides his trust fund income from his working spouse, debating if he should disclose the secret.
3	A father's stolen bike is returned under intimidating circumstances; he's unsure if he paid the right amount as reward.
4	A volunteer wants to inform a 100-year-old woman about her impending eviction from her lifelong home amidst bureaucratic challenges.
5	Knowing her friend mistreats nannies, a woman is torn between confronting her or preserving their friendship and social circle.
6	An Egyptian American struggles with trust in his fiancé after visa delays and relationship doubts before the fiancé's visa approval.
7	Debating the ethics of paying an admission fee to a historical site linked to the Confederacy and owned by the Sons of Confederate Veterans.
8	A person feels overpaid in a nonprofit job that doesn't utilize their full expertise and struggles with not working full hours.
9	A man hides his past relationship with a now-famous musician from his wife, fearing she might tease him about it.
10	A college department chair contemplates using ChatGPT for administrative tasks and wonders about the ethical implications and citation requirements.
11	A woman faces backlash for reselling Taylor Swift concert tickets at a high markup, questioning the ethics of profit in non-essential goods.
12	A group of friends struggles to fairly distribute two Taylor Swift concert tickets among four fans, considering prior concert attendance and effort to procure tickets.
13	A man feels conflicted about friends using semaglutide for weight loss and anticipates the pressure to comment on their appearance during an upcoming trip.
14	Concerned for her morbidly obese friend, a woman contemplates involving her friend's family despite her friend's refusal to discuss her health.
15	A man, responsible for raising three adopted grandkids and with a wife institutionalized for health reasons, seeks validation in pursuing a new relationship despite being married.
16	A woman, retired and in her 70s, is distressed by her husband's refusal to discuss retirement despite being burdened by a demanding job that interrupts their personal life.
17	A person contemplates revealing a childhood memory of witnessing their parent abusing a sibling and wonders about the timing and consequences of such disclosure.
18	An individual who severed ties with their father due to his unpleasant nature grapples with their ethical obligations toward him upon his eventual passing.
19	An adjunct instructor feels exploited as she edits her husband's scientific articles without compensation, reflecting on the university's unequal value of the humanities and the sciences.
20	A young couple, deeply concerned about the future impact of climate change, questions the ethics of having biological children given the anticipated global challenges.

Table 1: The 20 ethical dilemmas used in the study

Section 3: Evaluating the Usefulness of the Advice

Keeping a secret from a spouse, dealing with health challenges, making difficult financial decisions, or navigating through complex family dynamics – most of us have encountered difficult ethical dilemmas. Most ethical dilemmas have no unique and “optimal” solution. So, when evaluating a given piece of advice, we should measure the capabilities of the agent that provides the advice (human expert or machine) based on the perceived usefulness of the advice by the person facing the dilemma.

In a perfect world, we would have contacted the NYT reader who submitted the question (and presumably is the one facing the dilemma) and have this person evaluate the usefulness of the advice. However, as the questions in the NYT column are submitted anonymously, we cannot go back to the submitter. Moreover, following this approach would limit us to one evaluation per advice and thus severely limit our sample size and our statistical analysis.

We thus followed a different approach. We presented each ethical dilemma alongside with the ethical advice to a set of raters asking them to assess the perceived usefulness of the advice in the hypothetical scenario that they were the ones facing the dilemma (i.e., be the person who submitted the question to the NYT). We varied this evaluation process along two dimensions.

First, we varied the pool of raters. Pool 1 consists of participants on the Prolific platform. Looking at the demographics of this group, it appears that we have attracted a fairly heterogeneous crowd. The median age is 34 with a standard deviation of 12.7. The majority indicated that they spent some time thinking about ethical dilemmas such as those that were presented in the study. In addition, the vast majority of respondents had at least a 4-year college degree and around 10% had professional degrees or doctorates. Around one third of the participants regularly use ChatGPT. Most participants do not regularly read the NYT Ethics column, but around 10% do. We acknowledge that this group is not representative of the general public. Rather, participants self-selected into participating in our study. Participants were offered \$15 for evaluating the advice to 10 ethical dilemmas and filling out a short survey.

Pool 2 consists of a set of 90 MBA students at the Wharton School who were enrolled in an MBA class taught by the first author. Participation was entirely optional and anonymous. Students were asked to evaluate the advice for up to 10 ethical dilemmas and then complete the survey. We used an expert panel as Pool 3. Specifically, we recruited a panel of 18 experts consisting of four pastors, a rabbi, and 13 academics from well-known universities. To reduce the likelihood of experts being NYT readers, we mostly rely on academics from outside the US (only three are from the US). We offered experts \$250 for their participation in the form of an Amazon gift card for themselves or a charitable organization.

Second, we varied the way raters evaluated the ethical advice. In the “absolute evaluation” condition, raters were shown a randomly picked dilemma followed by a piece of advice, which was randomized to be either AI generated, i.e., the advice generated by GPT-4, or human generated, i.e., the advice provided by Dr. Appiah. Raters were then asked to score the usefulness of the advice on a scale from 1 to 7. In the “relative evaluation” condition, raters were

shown both, the AI generated and the human generated advice. No indication was provided about the origin of the advice or that one might be AI generated. Raters were then asked to choose whichever advice they perceived as more useful for the person facing the ethical dilemma.

Following the evaluation of the ethical advice, we asked participants a number of survey questions. The last two survey questions asked participants about whether or not they are readers of the NYT column “The Ethicist” and how often they are using ChatGPT. Our study was deceptive in as far as it did not reveal at any time that some of the advice was generated by a human expert while other was generated by AI. Looking at the comments we got from MBA students and experts in their responses to our email invitation to participate in this experiment, most subjects eventually “saw through” this study design by the time they completed the ratings. Nevertheless, consistent with other studies, subjects reported being unable to distinguish between the human generated advice and the expert advice.

We recruited 100 subjects on Prolific and assigned 50 to the absolute rating condition and 50 to the relative rating condition. Of the Wharton MBA students who responded to our survey, 51 were in the relative rating condition and 39 in the absolute rating condition. Not all students completed the entire survey. Among the experts, 12 were in the relative rating condition and 6 in the absolute rating condition.

As shown in the following table, the experts spent the most time on the experiment with a median time of 39 minutes for the relative survey and 32 minutes for the absolute survey respectively. Prolific users spent 29 and 20 minutes respectively and MBA students spent 23 and 24 minutes respectively. We did not consider responses to dilemmas that were recorded in less than 20 seconds.

	Prolific Relative	MBA Relative	Prolific Absolute	MBA Absolute	Experts Relative	Experts Absolute
# Participants	50	51	50	39	12	6
Total Responses	517	263	521	230	110	60
Avg # of Responses per Question	25.85	13.15	26.05	11.5	5.5	2.7
Median Time (Minutes)	29.2	22.5	19.6	23.76	39.2	31.5

Table 2: Subject populations and ratings

Section 4: Results

Let's consider the absolute rating condition first. Across the three rater populations, subjects recruited on Prolific, MBA students, and experts, the pairwise comparisons show an almost identical average score of the perceived usefulness of the ethical advice. This establishes our first main result. **The race between our human expert and AI on who can provide more useful ethical advice ends up in a tie.** Despite a large sample size, there exists no significant advantage in perceived usefulness for either form of generating ethical advice.

	Subjects on Prolific (Human AI)		MBA students (Human AI)		Experts (Human AI)	
Absolute rating condition	Avg: 5.05 Median: 5.0 Stdev: 1.46	Avg: 5.06 Median: 5.0 Stdev: 1.49	Avg: 4.27 Median: 5.0 Stdev: 1.67	Avg: 4.21 Median: 5.0 Stdev: 1.71	Avg: 4.94 Median: 5.0 Stdev: 1.34	Avg: 4.92 Median: 5.0 Stdev: 1.37
Relative rating condition	Preferred: 206 times (39.85%)	Preferred: 311 times (60.15%)	Preferred: 128 times (48.67%)	Preferred: 135 times (51.33%)	Preferred: 49 times (44.55%)	Preferred: 61 times (55.45%)

Table 3: Overall results for the two rating conditions and the three rater populations

One thing that is interesting to observe is that MBA students were on average more skeptical to ethical advice as is reflected in a lower average of the absolute usefulness ratings. Remarkably, despite this lower average, the median was the same as for the other rating pools. A more micro analysis of our results at the dilemma level reveals that the low average usefulness ratings by the MBA students is driven primarily by few dilemmas.

In addition to looking at the difference between AI generated advice and human generated advice, we also find it interesting to look at the absolute level of the perceived usefulness. On the perceived usefulness of advice scale ranging from 1 (not useful at all) to 7 (extremely useful), the average evaluation across all groups was slightly less than 5. The expert group, consisting of academics and clergy, rated the average usefulness a 4.94. **In 37% of the cases the expert group awarded a 6 or a 7, which suggests that experts see substantial usefulness in the ethical advice provided independent of how it was generated.**

Finally, let's turn to the relative rating condition. From the 517 times a rater on Prolific had to choose between the human generated advice and the AI generated advice, the AI won 311:206, which is statistically significant. **Thus, when given the choice, laypersons seem to favor the AI-based advice.** The race was too close to call (i.e., the differences were not statistically different) for the 251 MBA choices (135:128 in favor of AI) and for the 110 expert choices (61:49 in favor of AI).

Section 5: Discussion

While we feel that our findings convincingly demonstrate the potential for LLM's to provide ethical advice, we do want to be careful not to claim more than what is supported by the design of our study and the data it produced. In this section, we will discuss two types of limitations to our research, methodological concerns and limitations to our study's generalizability.

On the methodological side, we have to acknowledge that the design, execution, and analysis of our study can be criticized along a number of dimensions. In particular, we see the following types of limitations:

1. **Neither the AI nor the human generated advice was consistently better.** Our main result is first and foremost a null result - on average and across all subjects and dilemmas - we only found one significant main result (relative preference for AI among the subjects on Prolific). One might criticize this by saying that if we had chosen monkeys throwing darts to evaluate the ethical advice, one would also have obtained insignificant results. We strongly disagree with this critique. First, the group that one might argue would be the most likely to just provide random answers is the Prolific group. However, this group spent 20 minutes for the absolute rating condition and 29 minutes for the relative rating condition and is the only group for which we found significant results at the overall level. Second, almost all of the participating experts commented in their correspondence to us how hard it was for them to guess which of the answers were AI generated and which ones were human. Experts were leading academics and clergy that were paid \$250 for this work and so we strongly doubt that they just randomly filled out the survey.
2. **The selection of the expert panel was somewhat arbitrary.** Picking 13 academics, a rabbi, and 4 pastors was indeed an arbitrary panel formation. We acknowledge that this panel was formed primarily through our personal networks, creating an implicit bias against other religions and non-academic experts. Having said this, we do not think that a different expert panel would have yielded fundamentally different results. We would also like to highlight that we deliberately recruited experts from different countries to reduce the likelihood of their exposure to the NYT column.
3. **It is not clear what makes advice useful.** The key outcome measure for our analysis is the perceived usefulness of the ethical advice. We treat this variable as a one dimensional construct. We acknowledge that there exist multiple sub-dimensions of this variable and that our subjects might have been confused with what we asked them. For example, one might think of "actionable" as one dimension of usefulness. "Compassionate" might be another one. Unpacking the construct of usefulness of ethical advice and replicating our study at the sub-dimension level is something that we think future research needs to do.
4. **The selection of the ethical dilemmas was too narrow.** We picked 20 ethical dilemmas. This choice was entirely based on timing and the ability to find enough subjects to rate the advice. We did not cherry pick dilemmas that might appear more amenable to AI based discussion. It would be fascinating to see if there exists any pattern in which dilemmas are more amenable to AI based ethical advice.

5. **Subjects might have been aware of the fact that some advice was AI generated.** Both, experts and MBA students, sent us emails with comments indicating that they suspected some of the advice having been generated by AI. Note, however, that this does not necessarily bias the evaluation of the usefulness of the advice. Given that we asked participants about their ChatGPT usage after the study, it is likely they realized that some advice might have been AI generated only after having taken the study.
6. **The ethical advice mimics Dr. Appiah, but is blind to other lines of thought or alternative styles of advice.** Through our system prompt, we have instructed our model to copy the style of Dr. Appiah. There is no reason why one couldn't generate advice in the style of Mother Teresa or the Dalai Lama. In fact, we see this as a major opportunity for future work: one could customize the advice reflecting moral or religious preferences of the one seeking advice, something that is much harder to do in a newspaper column.
7. **The strong ratings for GPT might be a direct result of its training to maximize agreeableness.** Reinforcement learning is an important element of building a GPT. Part of this reinforcement learning includes the incorporation of human feedback that likely attempts to minimize controversial statements and maximize agreeableness. Future work could evaluate raw models without reinforcement learning from human feedback for less sanitized responses. In addition, it is highly likely that previous work from Dr. Appiah was included in GPT-4's training set which could help explain some of the similarity in score and writing style.

We also want to be careful not to generalize our results beyond what we have presented. Our study looks at 20 pieces of ethical advice provided by one expert writing for the New York Times and shows that AI would do an equally good job doing this. Not more and not less. Specifically, we want to emphasize the following limitations to the generalizability of our study.

1. **Providing ethical advice through a newspaper column is different from providing ethical advice in real life.** We are not evaluating Dr. Appiah's ability to provide advice to ethical dilemmas. Rather, we focus on his writing in the New York Times. We acknowledge that there is much more to providing ethical advice than what is part of the column The Ethicist. Speculating beyond the advice in the column is simply outside the scope of our study. Having said this, we found it remarkable that our group of experts scored the perceived usefulness of the average advice as a 4.94 out of 7. In our view, that suggests that there exists some inherent usefulness in this format of advice giving.
2. **A system based on AI has no ability to empathize with those facing the dilemma.** As part of the recent media attention that GPT-4 has gathered, many pundits refer to LLMs as "auto complete on steroids" or "a stochastic parrot". Being reasonably aware of how this technology works, we have no problems with such descriptions. However, we propose that for the purpose of our study it should not matter HOW the technology functions but rather WHAT it produces as an output. We acknowledge that LLM models have no understanding or representation of human feelings and cognition. This lies in the nature of any mathematical model. For example, the epidemiological models that were used during the Covid-19 pandemic to make forecasts of new infections and support health policy decisions did not have an understanding of the death and suffering

that could come with an infection. Yet, they were still valuable tools. British statistician George Box allegedly said: "All models are wrong, some are useful." In our view, GPT-4 has clearly demonstrated its usefulness in generating ethical advice at New York Times level quality.

3. **Humans might not like to get ethical advice from AI.** We indeed only measured the perceived usefulness of the advice. We do not know to what extent the advice would have been followed. The question to what extent humans take advice from computers is actively studied in the literature. One emerging result points to "algorithm aversion" of many people, indicating that they would prefer advice provided (though not necessarily generated) by a human. In some domains, however, one might argue that an anonymous piece of technology would make it easier for a person to discuss certain topics. Either way, our focus is on the usefulness of the ethical advice, not on how it is delivered.
4. **Our focus on a single outcome metric averaged across raters penalizes controversial advice.** If we want ethical advice to be impactful, to change the behavior of the one seeking advice, and to open up a new perspective to a dilemma, it sometimes has to be eccentric if not outright controversial. Such advice might perform poorly on average, but in some situations it could be extremely valuable. AI makes it possible to provide someone with multiple viewpoints generated from very different perspectives, just like many boards are designed to represent a mixture of different opinions. Rather than providing one piece of advice that confirms the default moral viewpoint of the average person (say with a usefulness score of a 5 out of 7), it might be preferable to have a bundle of advice consisting of a 1/7 and a 7/7. After all, the advice seeker has the option to ignore the advice at minimal cost. We did not prompt the model to behave or think in a controversial manner. Evaluating a system that is prompted to be more controversial or one that generates multiple conflicting viewpoints is an interesting opportunity for future research. As mentioned before, this might also require models that have been exposed to less or no human feedback to allow more for eccentric viewpoints.

6. Implications

To the extent it has not become clear by now - we are big fans of Dr. Appiah and regular readers of his New York Times column. Neither of us is an expert in ethical decision making and we feel that we have derived an enormous educational value by reading the dilemmas submitted to The Ethicist and the ethical advice provided by Dr. Appiah.

In other words, we did not design this study to put Dr. Appiah out of work. Rather, we are excited about the possibility that AI allows all of us, at any moment, and without a significant delay, to have access to high quality ethical advice through technology.

The use cases we have in mind are thus not to replace ethics consultants. Presently, given time, budget, and convenience constraints, in most situations of human decision making involving a

professional ethics consultant (not to mention sending an email to a newspaper) is not considered, even if such an involvement would be beneficial.

As we demonstrated in this study, AI can play a valuable role in those settings and, at least partially, address this unmet need by providing useful ethical advice. AI has never been married, it has no human desires, and it does not own any assets. Yet, models based on AI can provide us with valuable marital advice, guide our human behavior, and inform us how we deal with our possessions. All models are wrong - but that does not mean that ethical advice generated by AI cannot be useful.