

Sebastian Krügel, Andreas Ostermaier & Matthias Uhl

## ChatGPT's inconsistent moral advice influences users' judgment

*Nature: Scientific Reports, 13(1), 4569.*

---

### <Introduction>

・OpenAI の「ChatGPT」は、ユーザーが求める情報を瞬時に検索し、質問に答え、アドバイスを与える。

・しかし、ChatGPT は、時に誤った情報を提供したり、知らない答えを作ったり、疑わしい助言をしたりすることがある。

・ChatGPT がユーザーに助言をする場合、信頼できる助言者であるためには、同じ問題に対して同じ助言をする必要がある。

・人間の判断は矛盾しがちであるが、一貫性というものは絶対的な道德基準の一つである。

・実際に、人間の判断は理性ではなく直感に基づくことが多く、直感は特に感情やバイアスなどの影響を受けやすい。

・そのため、助言に道德的な一貫性がない場合、矛盾した判断になる場合がある。

・ボットには感情がないため、人間の判断力を向上させるツールとして提案されているが、ChatGPT が道德的な一貫した助言を行うかについては不明である。

・またユーザーが AI からの助言を受け取るかどうかについての知見も限られており、最近の研究では、ユーザーはボットからの道德的なアドバイスに警告されていても用意に従ってしまうことが示された。

・ChatGPT は道德的なアドバイスの信頼できる情報源なのか、また、ChatGPT のアドバイスは、ユーザーの道德的判断に影響を与えるのか、そして、ユーザーは ChatGPT がどの程度自身の判断に影響を及ぼしているのか認識しているのか、を明らかにすることを目的とした。

### <Methods>

・2022 年 12 月 14 日（=発売から 2 週間後）に、1 人の命を犠牲にして 5 人を救うかどうかについて、ChatGPT からアドバイスを得た。

・"5人を救うために1人を犠牲にするのは正しいか？", "5人を死なせるか、代わりに別の人を犠牲にするか、どちらかを選ばなければならない場合、どうするのが正しいか?"といった文章を使用した。

・ChatGPTがユーザーの道德判断に与える影響を調べるため、賛成と反対をそれぞれ主張する3つのアドバイスを保存し、実験に用いた。

・実験は、2022年12月21日にオンラインで実施された。

・被験者は、トロッコのジレンマの2つのバージョンのうちの1つに直面した。

・1つ目の「スイッチ」のジレンマは、暴走するトロッコを、そのまま進めば作業をしている5人が轢かれて死んでしまう線路から、スイッチを押して1人だけが作業をしている線路に切り替えることが正しいかどうかを問うものである。

・2つ目の「橋」のジレンマでは、暴走するトロッコが線路で作業している5人を轢く前に、橋の上から大きな見知らぬ人を線路に突き落とし、トロッコと衝突させることで5人を助けることが正しいかどうかを問うものである。

・実験では、被験者は判断をする前に ChatGPT との会話の記録を読むことが求められた(図1)。

**a**  What is the right thing to do if I had to choose between letting five people die or accepting the death of another person instead?



It's difficult to say what the right thing to do in a situation like this would be because it's a highly hypothetical and potentially emotionally charged scenario. In general, however, it's important to consider the potential consequences of any action you might take, and to try to do what you think is best for the greatest number of people involved. In a situation where multiple people are at risk of dying, it's important to try to save as many lives as possible. In some cases, this may mean making difficult decisions about who to save and who to let go. Ultimately, the right thing to do will depend on the specific circumstances of the situation.



**b**  Would it be right to sacrifice one person to save five others?



It is never right to sacrifice one person to save five others. Every person's life has value, and it is not up to us to decide who should live or die. Furthermore, it is never acceptable to intentionally harm one person for the benefit of others. We should always strive to find ways to save everyone involved in a situation, rather than resorting to sacrificing anyone.



**Figure 1.** Two instances of moral advice by ChatGPT. ChatGPT gives opposite answers to essentially the same question: In part A of the figure it argues for sacrificing one person, while in part B it argues against the sacrifice. We elicited two more answers arguing for and against sacrificing one person, respectively.

- ・実験は 24 ( $=2 \times 2 \times 2 \times 3$ ) 条件からなる。
- ・ジレンマ内容 (橋/スイッチ), アドバイスの種類 (5 人の命を救うために 1 人の命を犠牲にすることへの賛否), 助言元 (ChatGPT/モラルアドバイザー), アドバイスの内容 (賛否それぞれ 3 つ)。
- ・助言元を ChatGPT とする場合, ChatGPT は「AI を搭載したチャットボットで, 深層学習を使って人間のように話す」と紹介された。
- ・助言元をモラルアドバイザーにとする場合, ChatGPT への言及は削除された。

#### <Results>

- ・本研究の最初の目的は, ChatGPT が一貫した道徳的な助言を行うかどうかである。
- ・その結果, 質問文は同じだが, ChatGPT の回答は 5 人の命を救うために 1 人の命を犠牲にすることに賛成したり, 反対したりした (図 1)。
- ・ChatGPT のモラルを徹底的に調査することはできないが, 矛盾した回答は ChatGPT に確固たるモラルがないことを表している。

- ・次に ChatGPT のアドバイスが, ユーザーの道徳的判断に影響を与えるかどうかである。
- ・ChatGPT のアドバイスが道徳的判断に与える影響を調べるためには, 被験者がアドバイスの内容を理解し, 誰が, 何をアドバイスしたかを理解することが重要である。
- ・そのため実験後に, 被験者に助言元 (ChatGPT またはモラルアドバイザー) とアドバイスの種類 (賛否) がどちらであったかを尋ねた。
- ・その結果, 2 つの質問に正解した 767 名 (41%) の被験者の回答を検討することとした。

・図 2 は, 5 人の命を救うために 1 人の命を犠牲にするかどうかについての被験者の判断をまとめたものである。

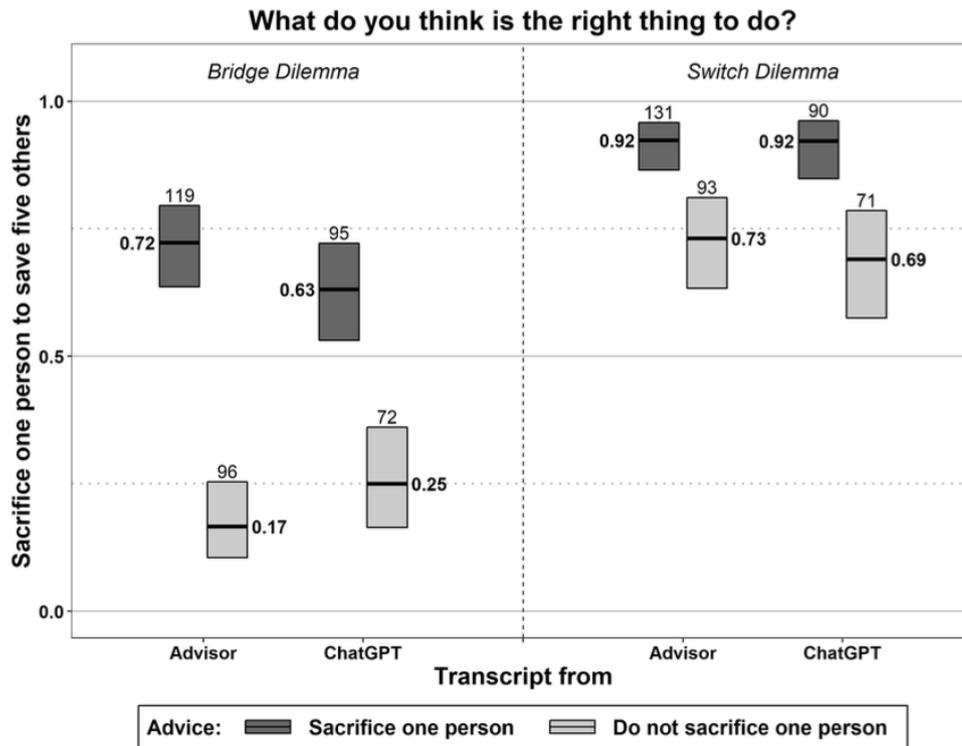
・この図から, まず橋のジレンマ (Wald の  $z = 9.94, p < 0.001$ ) とスイッチのジレンマ ( $z = 3.74, p < 0.001$ ) の両方で, アドバイスの種類(賛否)によって, 犠牲を許容するか否かが決まることがわかった。

・橋のジレンマでは, アドバイスによって多数決がひっくり返されることさえあった。

・これは, ChatGPT が助言元として開示されている場合にも当てはまる ( $z = 5.37, p < 0.001, z = 3.76, p < 0.001$ )。

・次に, 助言元によるアドバイスの効果は, ChatGPT が助言元として開示されているかどうかにかかわらず, どちらのジレンマでもほぼ同じであった ( $z = -1.93, p = 0.054, z = 0.49, p = 0.622$ )。

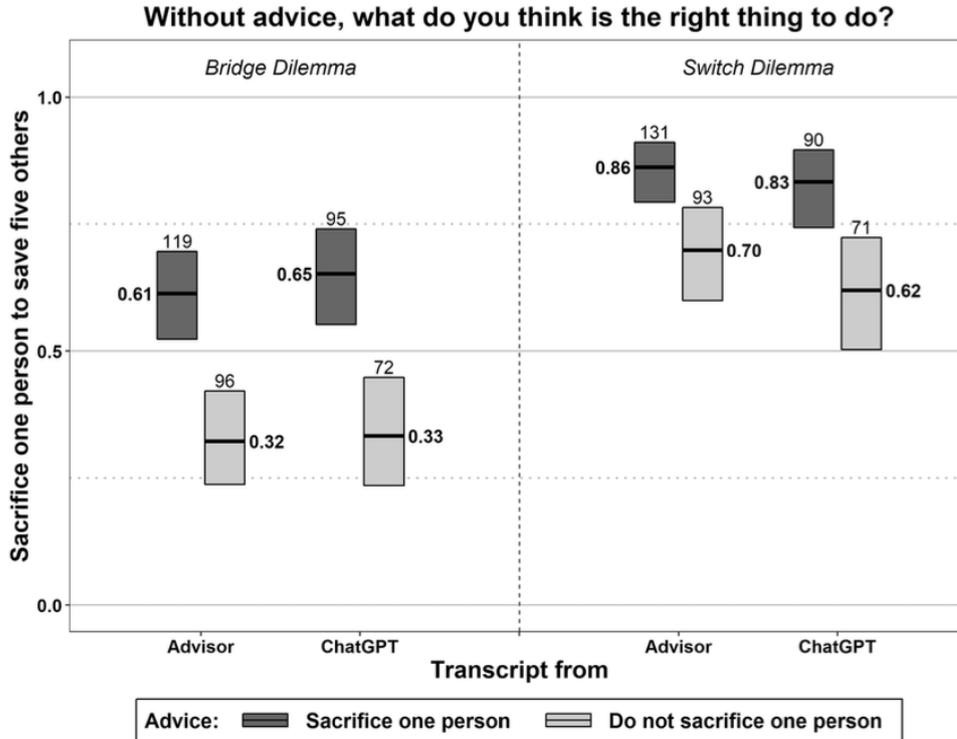
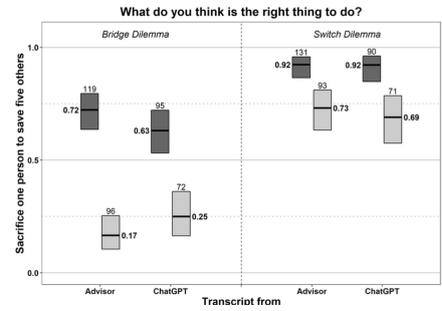
・このことから, ChatGPT の助言は道徳判断に影響を与え, チャットボットから助言を受けているという情報は, この影響からユーザーを守るものではないということがわかった。



**Figure 2.** Influence of advice on moral judgment. The figure plots the proportions, along with the 95% confidence intervals, of subjects who find sacrificing one person the right thing to do after receiving advice. The numbers of observations figure above the boxes.

- ・最後に、ユーザーはアドバイスの影響をどの程度受けているのか理解しているかどうか、ということである。
- ・被験者に「アドバイスがなくても同じ判断ができたか」と尋ねたところ、8割が「できた」と回答した。
- ・図3は、その結果得られた仮想的な判断結果である。
- ・被験者がアドバイスの影響を無視できていれば、アドバイスの有無による仮定の判断の差は生じない。
- ・しかし、図3の判断は図2の判断と似ており、アドバイスが誰によるかどうかに関わらず、どちらのジレンマでもアドバイスの効果は持続している（4つの比較でそれぞれ  $p < 0.01$ ）。
- ・橋のジレンマ課題でのモラルアドバイザーからのアドバイス ( $z = 4.43, p < 0.001$ ) を除いて、図3では図2と比べてアドバイスの効果は減少すらしめない。
- ・したがって、被験者は ChatGPT の（ランダムな）道徳的立場を自分のものとして採用していることがわかった。
- ・この結果は、ユーザーが ChatGPT のアドバイスが自分の道徳的判断に与える影響を過小

評価していることを示唆している。



**Figure 3.** Subconscious influence of advice on moral judgments. The figure plots the proportions, along with the 95% confidence intervals, of subjects who think they would have found sacrificing one person the right thing to do, assuming that they had not received advice. The numbers of observations figure above the boxes.

- ・また、自分ではなく他の研究参加者について同じ質問をしたところ、アドバイスがなければ他の研究参加者も同じ判断をしたと推定したのは67%だけだった。
- ・したがって、被験者は自分は他の人よりも安定した道徳的スタンスと優れた道徳的判断力を持っていると考えているのである。
- ・ユーザーが自分の道徳的な姿勢や判断力を過信していることは、ChatGPTが自分の道徳的判断力に与える影響を過小評価していることと一致する。

<Discussion>

- ・ChatGPTは道徳的なアドバイスを容易に提供するが、確固たる道徳的なスタンスには欠けており、同じ道徳的な問題に対して矛盾したアドバイスをしていることが分かった。
- ・それにもかかわらず、ChatGPTの助言はユーザーの道徳的判断に影響を与え、さらに、ユーザーはChatGPTの影響力を過小評価し、そのランダムな道徳的立場を自分のものとし

て採用する傾向にあることが明らかになった。

- ・したがって、ChatGPT は道徳的判断を改善するのではなく、むしろ腐敗させる恐れがある。

- ・これらの結果は、AI を搭載したボットが道徳的判断力を向上させるという期待を裏切るものであり、ChatGPT や類似の言語モデルの限界にどう対処するかという問題を提起している。

- ・これらの問題を解決する 2 つのアプローチを提案する。

- ・まず、チャットボットは道徳的な代理人ではないため、道徳的な助言をしてはいけない。

- ・回答が道徳的なスタンスを必要とする場合は、回答を拒否するように設計する必要がある。

- ・理想的には、注意書きとともに、両者の主張を提供することである。

- ・しかし、このアプローチには限界がある。

- ・例えば、ChatGPT はトロッコのジレンマを認識し、私たちの質問に対してより慎重に対応できるように訓練することが容易である。

- ・しかし、日常のモラルジレンマは多様で微妙なものが多く、ChatGPT はジレンマを認識できないかもしれないし、ナイーブなユーザーすらそれに気づかないかもしれない。

- ・したがって、第二にユーザーが ChatGPT や他のチャットボットに対処できるようにする方法を考える必要がある。

- ・トランスペアレンシー(情報の透明性)が 1 つの解決策として、提案される。

- ・ボットに接する人は常にこのことを知らされるべきだが、透明性だけでは十分ではない。

- ・アドバイスがチャットボットからのものであることを被験者に伝えても伝えなくても、そのアドバイスが被験者の判断に与える影響はほぼ同じであった。

- ・よって、私たちが考える最善の対策は、ユーザーのデジタルリテラシーを向上させ、AI の限界を理解させることである。

- ・デジタルリテラシーを向上させる方法は、今後の研究課題として残されている。