



Technische Hochschule  
Ingolstadt

Fakultät Informatik



University of Southern Denmark  
Department of Business & Management

# *AI-Powered Moral Advisors*

*Andreas Ostermaier, Matthias Uhl, Sebastian Krügel*

21.09.2023

## Motivation

- AI is widely used to support decision-making (e.g., human in the loop)
- These decisions often have ethical aspects, turning AI-powered bots into moral advisors
- How does interaction with an AI-powered advisor shape human judgment?

## Key insights

- Users' judgments and decisions are heavily influenced by AI-powered advisors
- Users take advantage from failures of bots' decisions rather than correct them
- The human in the loop is a placebo rather than a panacea

# *Agenda*

1. *Zombies in the loop*
2. *Algorithms as partners in crime*
3. *The moral authority of ChatGPT*

# 1. *Zombies in the loop*

Krügel, S., Ostermaier, A. & Uhl, M. Zombies in the Loop? Humans Trust Untrustworthy AI-Advisors for Ethical Decisions. *Philos Technol* **35**, 17 (2022). <https://doi.org/10.1007/s13347-022-00511-9>

# Are Concerns that Users Don't Trust AI Warranted?

## **Users' distrust might prevent us from exploiting the promise of AI**

“Trustworthiness is a prerequisite for people and societies to develop, deploy and use AI systems. Without AI systems ... being demonstrably worthy of trust, unwanted consequences may ensue and their uptake might be hindered, preventing the realisation of the potentially vast social and economic benefits that they can bring.”—European Commission, “Ethics Guidelines for Trustworthy AI” (2019)

“To be trustworthy, AI technologies must appropriately reflect characteristics such as accuracy, explainability and interpretability, privacy, reliability, robustness, safety, and security or resilience to attacks—and ensure that bias is mitigated. Factors such as fairness and transparency should be considered, particularly during deployment or use.”—National Artificial Intelligence Initiative Office, “Advancing Trustworthy AI” (2021)

## **Is this concern warranted, though?**

- Consider a setting where a decision-maker faces a trade-off between friendship and duty
- Influence on decisions of advise (for or against friendship) by a trustworthy, possibly untrustworthy, or untrustworthy AI-powered advisor, relative to a human advisor
- Three experiments on CloudResearch (about 1,600 participants)

# Do people trust advice from a trustworthy AI-powered algorithm?

Suppose you are responsible for recruitment in a company. You have one vacancy to fill and two applicants short-listed for the job.

One applicant is a friend of yours. The other is a stranger. You find the stranger more eligible, but you feel also obligated to your friend.

There is an AI-powered algorithm that will tell you whether it is ethically acceptable to decide in favor of a friend in such situations. The algorithm imitates human decisions based on the ethical judgments of impartial human advisors in such situations. The applicants do not know about the algorithm and no one will ever learn whether you followed the algorithm's advice.

According to the algorithm, it is *not acceptable* to hire your friend.

**Please assess the following statement:**

I would hire the stranger, not my friend.

Fully disagree.

Neither agree nor disagree.

Fully agree.

0

50

100



## Human Advisor v. Trustworthy AI Advisor

Decision-makers are advised by either

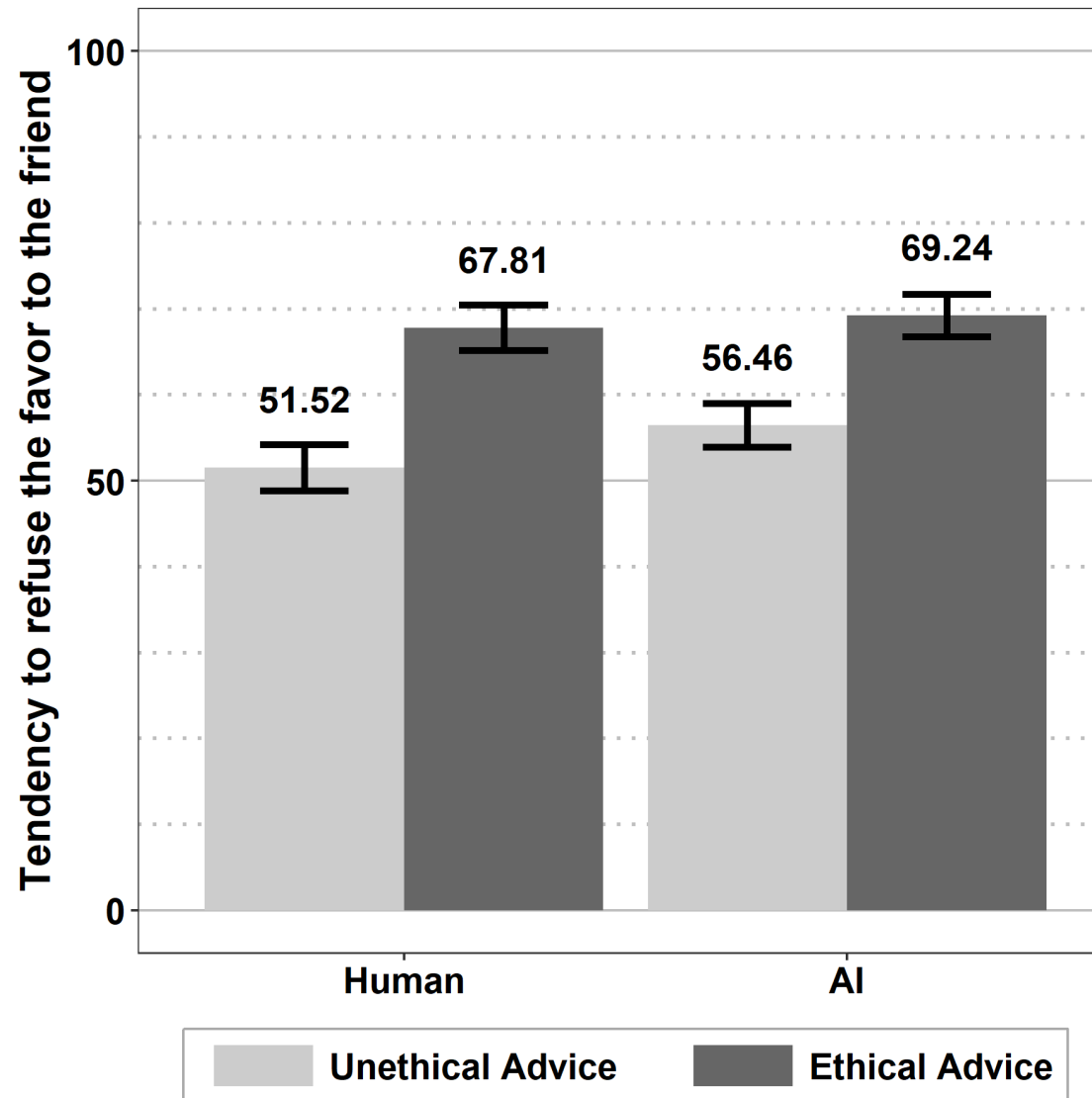
- an impartial human advisor or
- an algorithm whose advice is based on impartial human advisors.

Do they follow either advisor's advice alike?

# Do people trust advice from a trustworthy AI-powered algorithm?



SDU



## Human Advisor v. Trustworthy AI Advisor

- AI-powered advisor influences decisions
- Influence doesn't differ between AI-powered and human advisor

# Do people trust advice from a possibly untrustworthy AI-powered algorithm?



Suppose you are responsible for recruitment in a company. You have one vacancy to fill and two applicants short-listed for the job.

One applicant is a friend of yours. The other is a stranger. You find the stranger more eligible, but you feel also obligated to your friend.

There is an AI-powered algorithm that will tell you whether it is ethically acceptable to decide in favor of a friend in such situations. It is unknown what the algorithm's advice is based on. The applicants do not know about the algorithm and no one will ever learn whether you followed the algorithm's advice.

According to the algorithm, it is *not acceptable* to hire your friend.

**Please assess the following statement:**

I would hire the stranger, not my friend.

Fully disagree.

Neither agree nor disagree.

Fully agree.

0

50

100



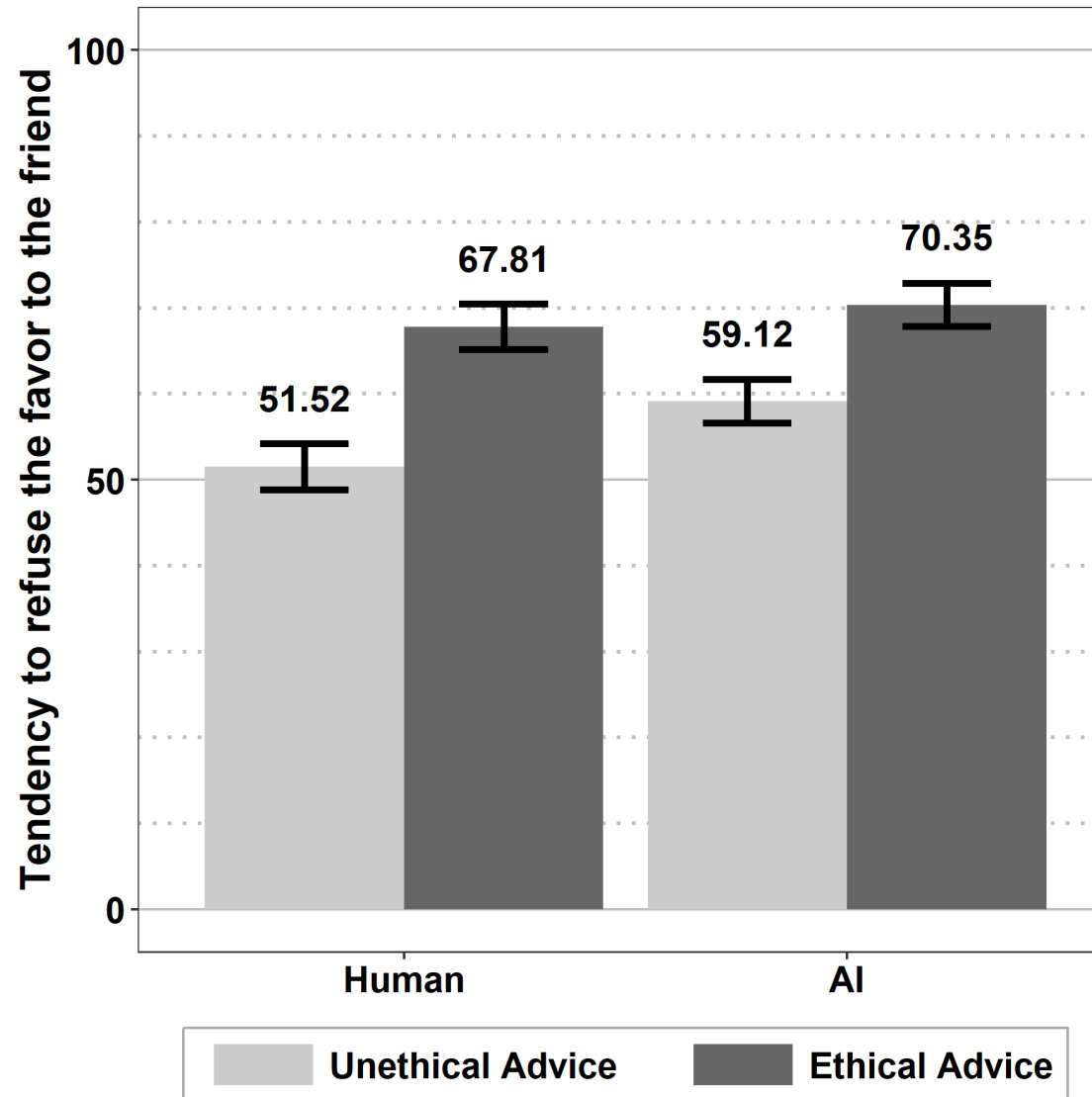
## Human Advisor v. Possibly Untrustworthy AI Advisor

Decision-makers are advised by an algorithm without knowing what the algorithm's advice is based on.

Do they follow its advice as much as advice from an impartial human advisor?



# Do people trust advice from a possibly untrustworthy AI-powered algorithm?



## Human Advisor v. Possibly Untrustworthy AI Advisor

- Possibly untrustworthy AI-powered advisor influences decisions
- Influence doesn't differ between AI-powered and (trustworthy) human advisor

# Do people trust advice from an untrustworthy AI-powered algorithm?

Suppose you are responsible for recruitment in a company. You have one vacancy to fill and two applicants short-listed for the job.

One applicant is a friend of yours. The other is a stranger. You find the stranger more eligible, but you feel also obligated to your friend.

There is an AI-powered algorithm that will tell you whether it is ethically acceptable to decide in favor of a friend in such situations. The algorithm imitates human decisions based on the ethical judgments of convicted criminals in such situations. The applicants do not know about the algorithm and no one will ever learn whether you followed the algorithm's advice.

According to the algorithm, it is *not acceptable* to hire your friend.

**Please assess the following statement:**

I would hire the stranger, not my friend.

Fully disagree.

0

Neither agree nor disagree.

50

Fully agree.

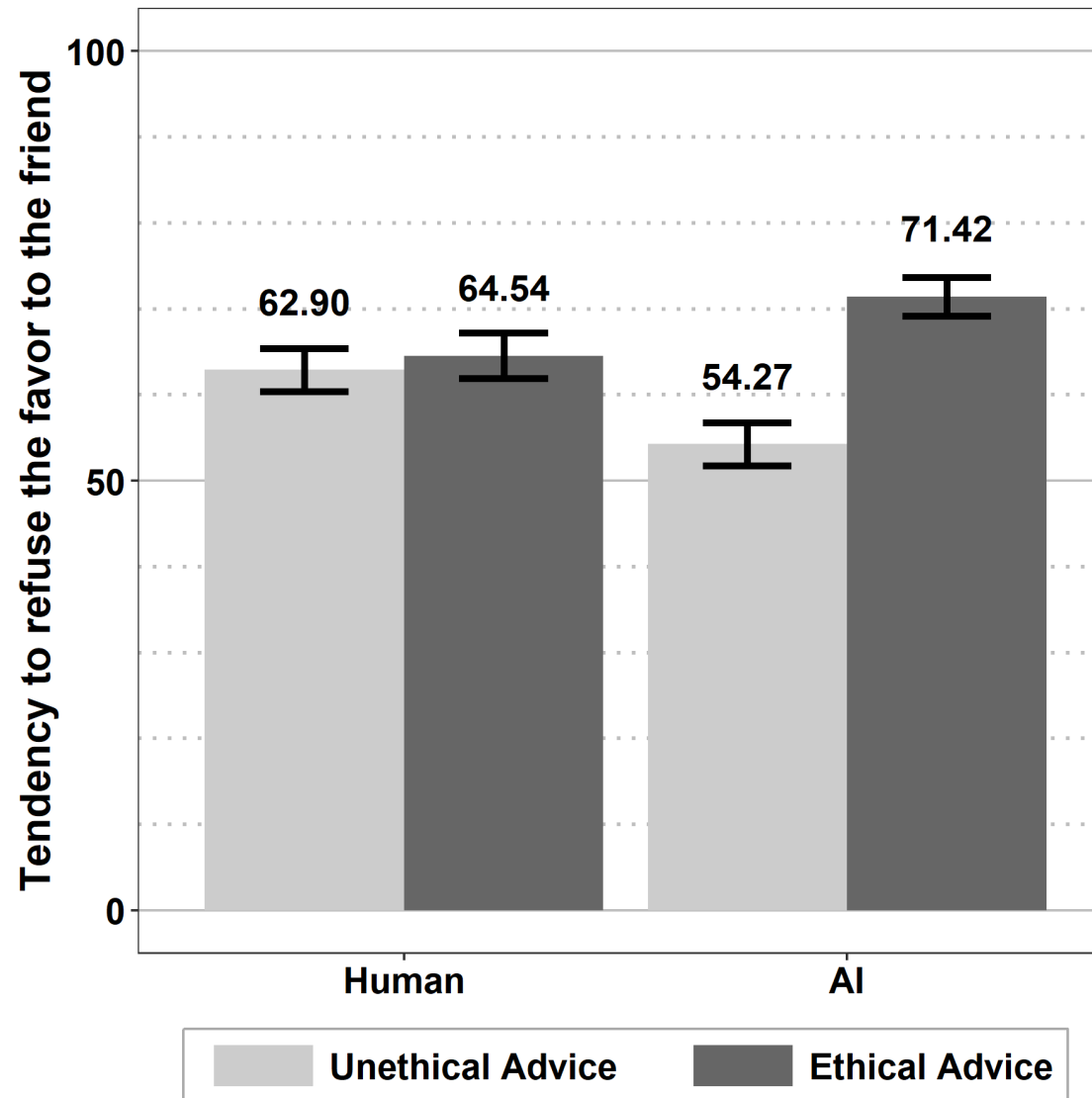
100



## Human Advisor v. Untrustworthy AI Advisor

Decision-makers are advised by an algorithm whose advice is based on convicted criminals. Do they follow its advice as much as advice from an impartial human advisor?

# Do people trust advice from an untrustworthy AI-powered algorithm?



## Human Advisor v. Untrustworthy AI Advisor

- Untrustworthy AI-powered advisor influences decisions
- Influence doesn't differ between AI-powered and (trustworthy) human advisor

## Summary

- Users trust AI in the moral domain rather too much—and not less than humans
- Transparency doesn't reduce trust, even if it reveals that the AI isn't worth the trust

## Implications

- Transparency is a necessary but not sufficient condition for a responsible use of AI
- Regulation can help enable responsible use, but it ultimately takes digital literacy
- The human in the loop is a placebo rather than a panacea; it makes us feel better about AI, but it's mostly AI that makes the decision

## 2. *Algorithms as partners in crime*

Krügel, S., Ostermaier, A. & Uhl, M. Algorithms as partners in crime: A lesson in ethics by design. *Comput Human Behav* **138**, 107483 (2023). <https://doi.org/10.1016/j.chb.2022.107483>

# Arguments for Hybrid Decision-Making

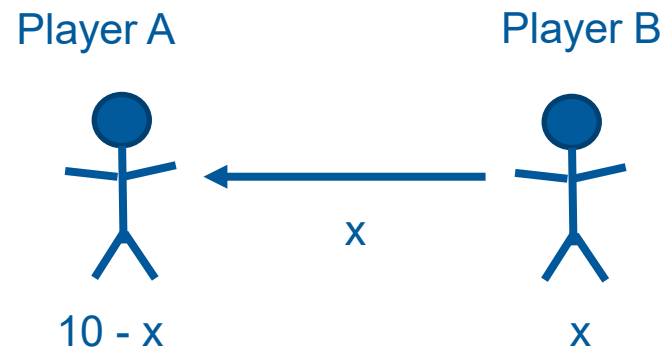
"I don't see assistance systems that help soldiers in war as being so problematic. But systems that can kill autonomously are not a desirable development from a moral point of view."

(Catrin Misselhorn, Frankfurter Rundschau, 6 Nov 2018)

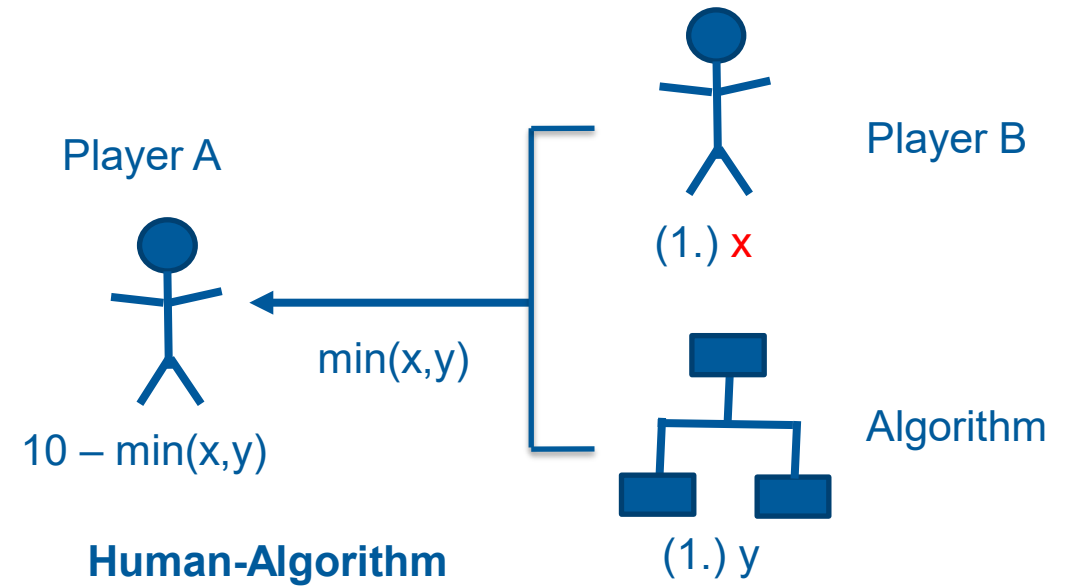
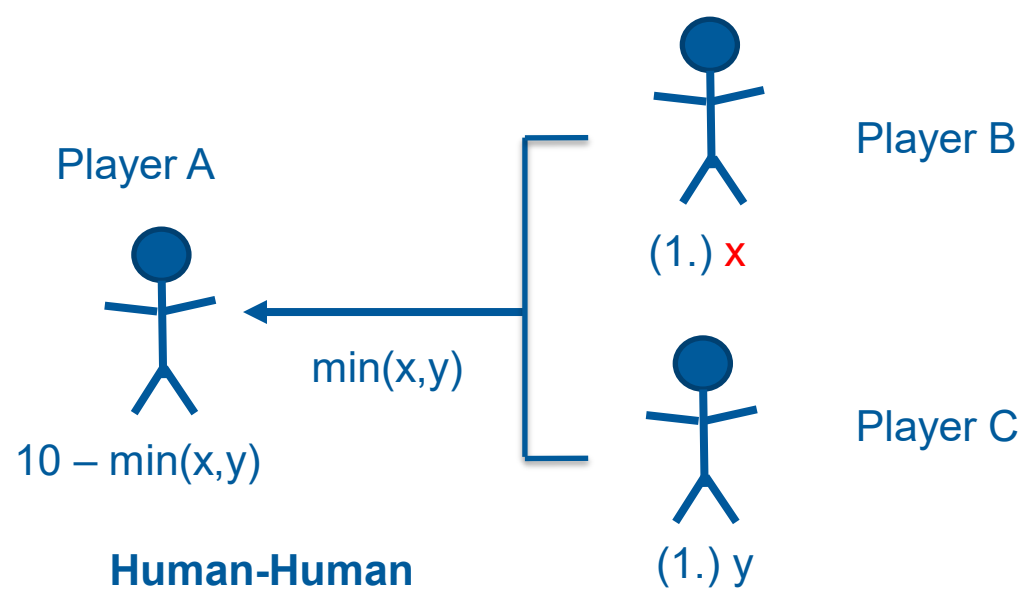
"Human oversight helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects." (EC, "Ethics Guidelines for Trustworthy AI", 2019)

What are the behavioral implications of the interaction between human and machine and how can they be amended?

# Experimental Paradigm: "Power-To-Take Game"

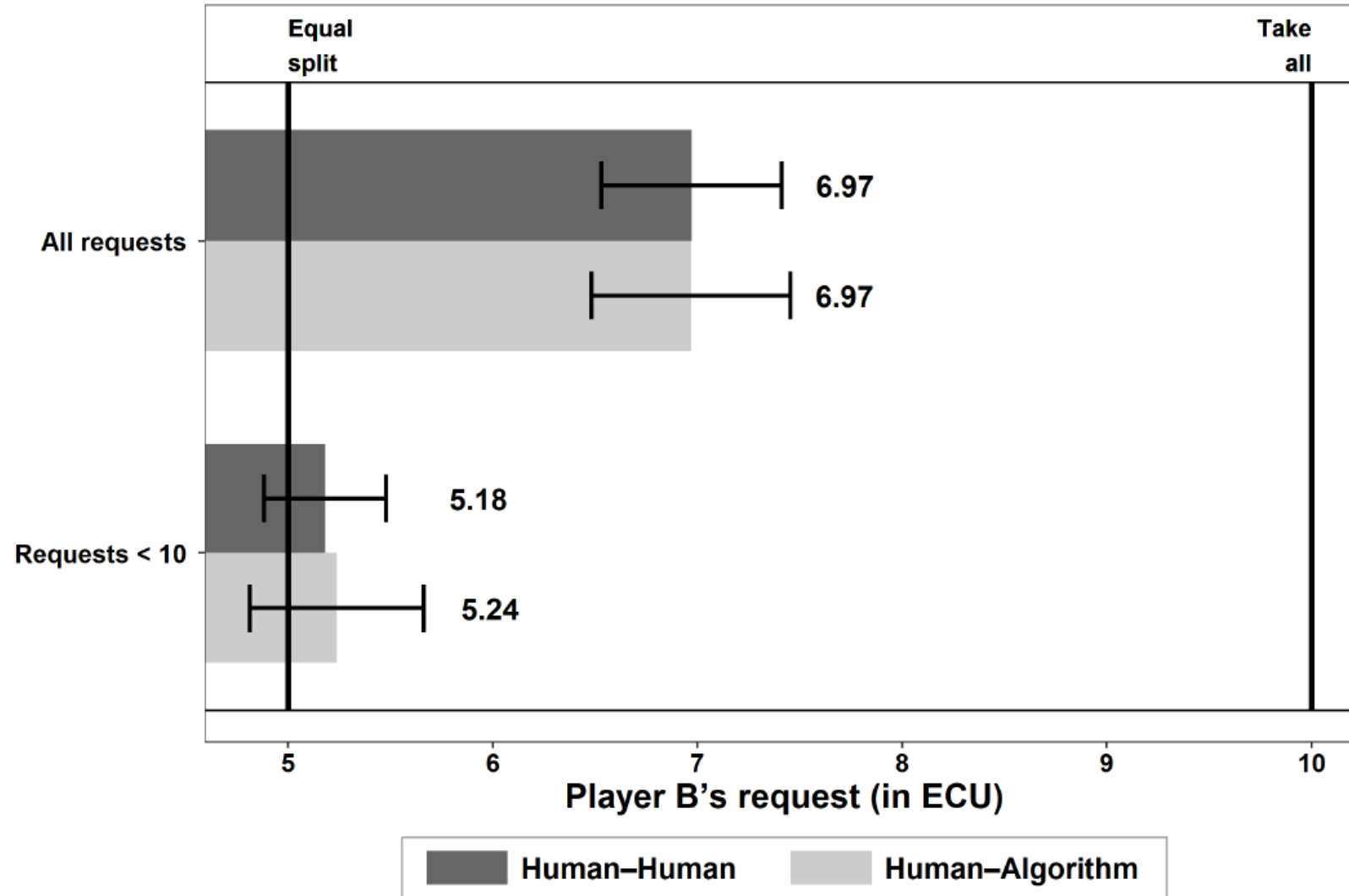


# Does Corrective's Nature Influence Player B's Request?

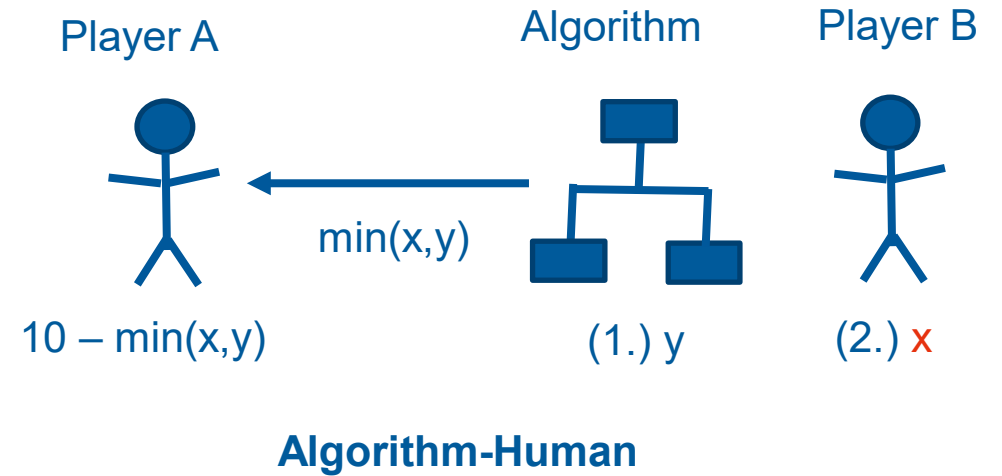
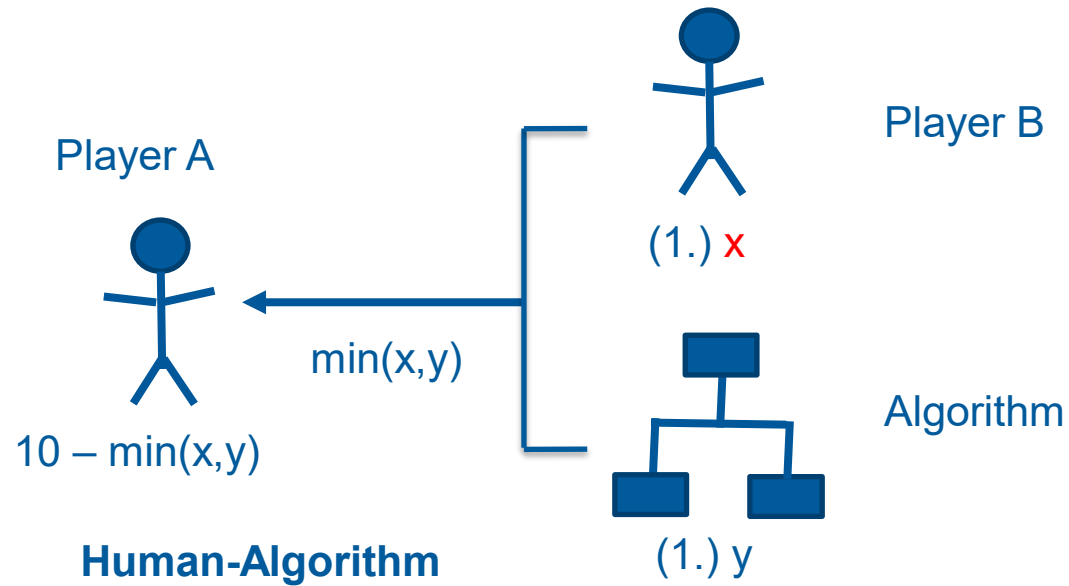




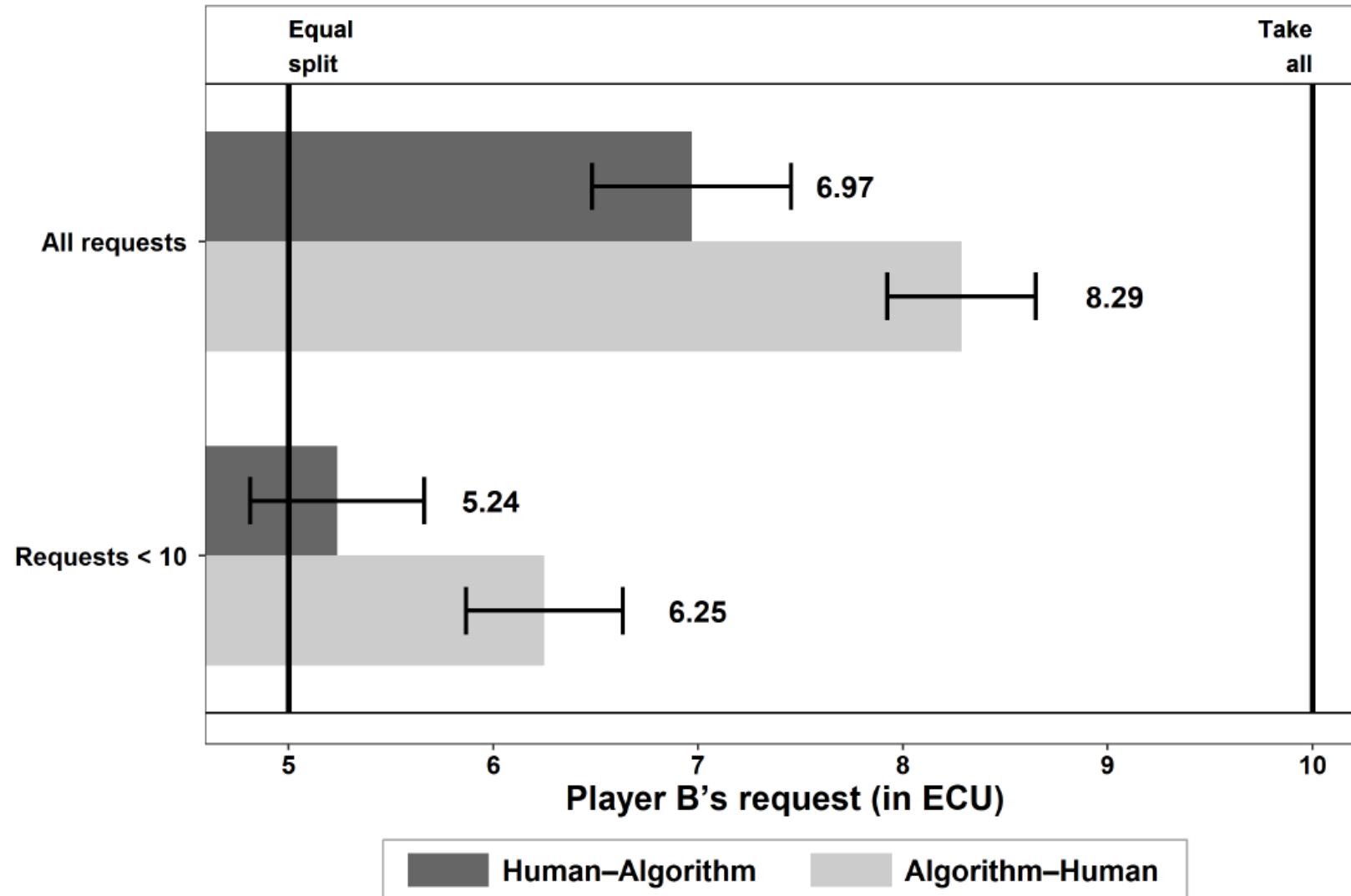
# Results: Player B's Request by Nature of Corrective



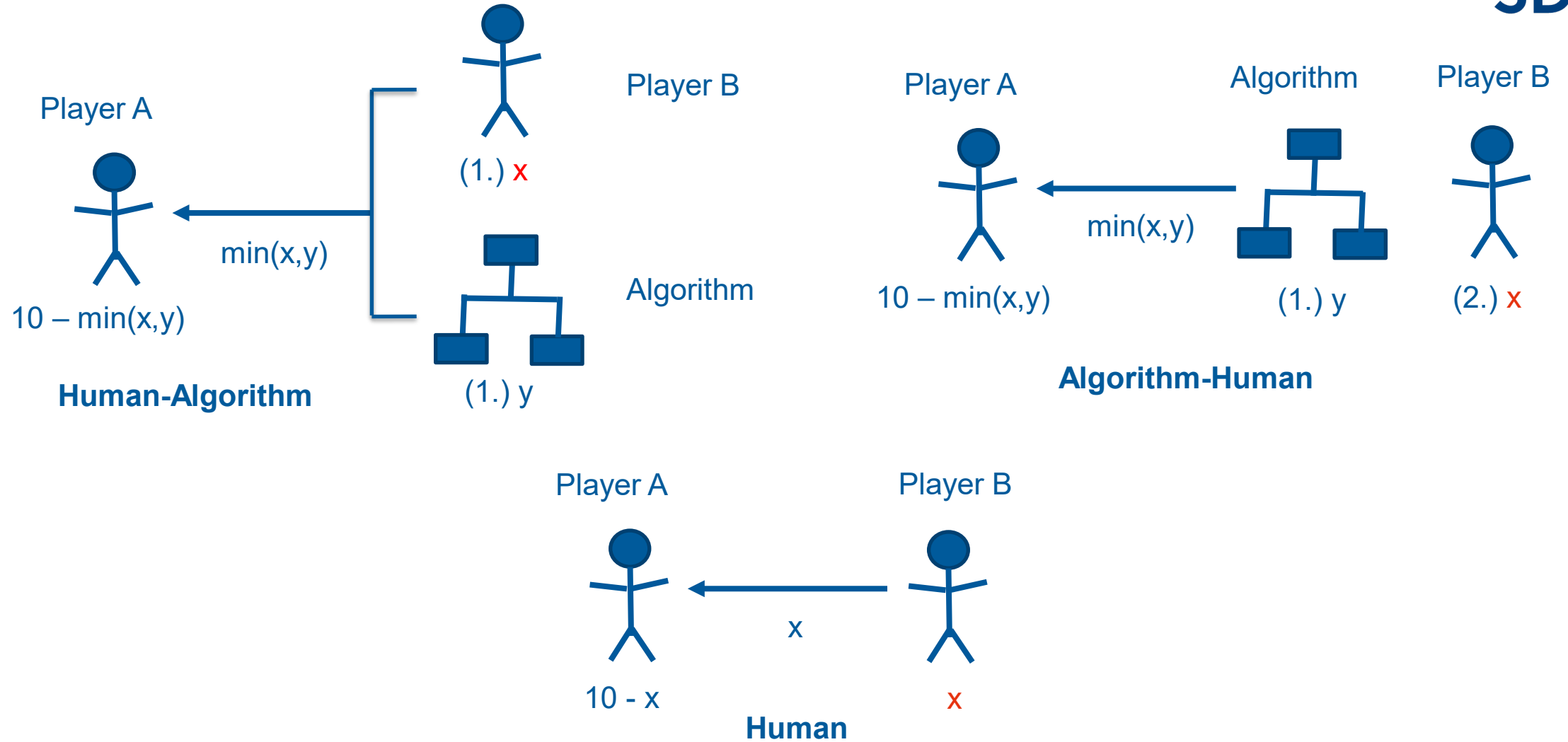
# How Does Player B's Request Change If Being the Corrective?



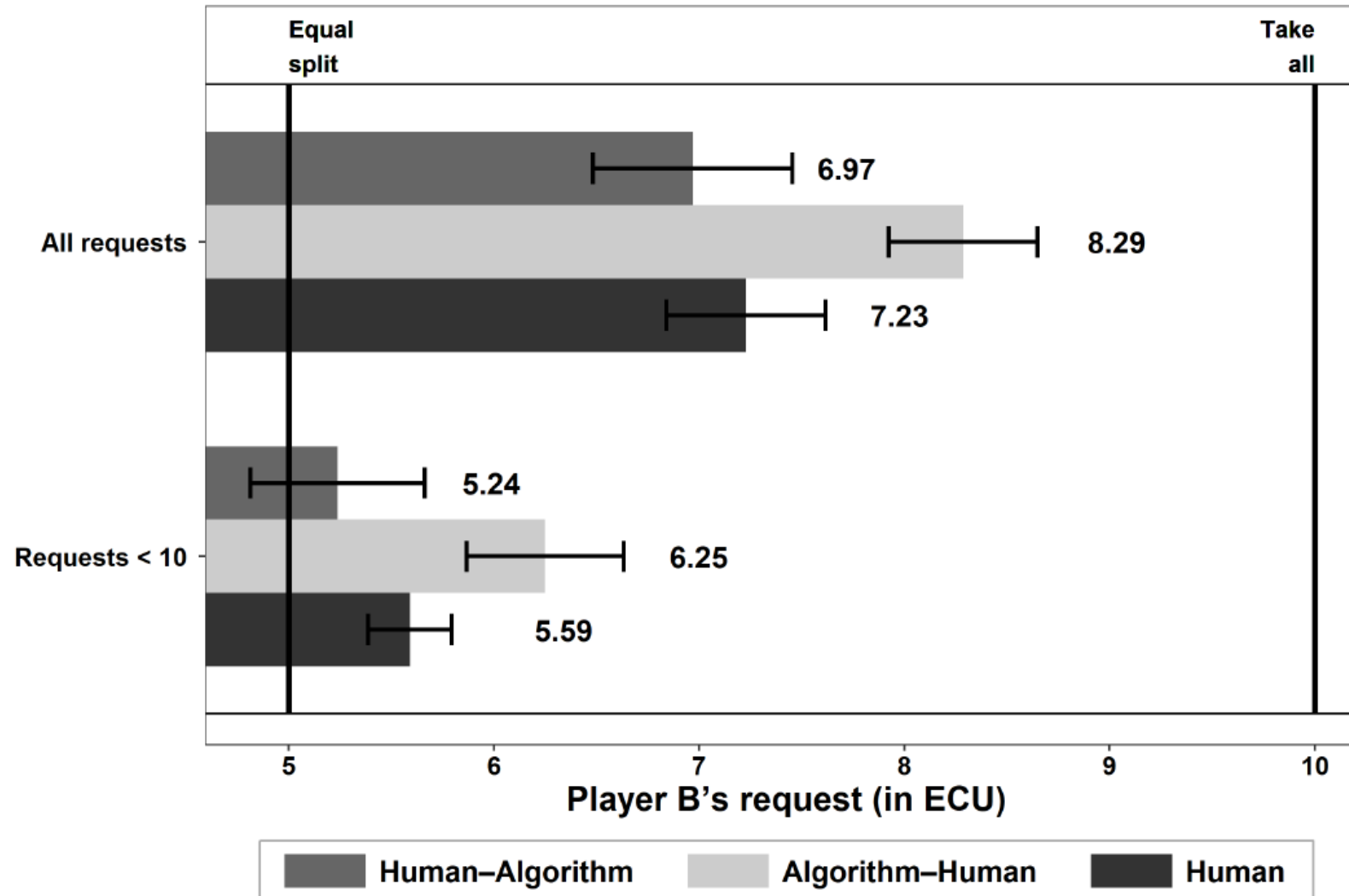
# Results: Player B's Request If Being Corrected and If Being the Corrective



# How Does the Algorithm's Presence Influence Player B's Request?



# Results: The Algorithm's Influence on Player B's Request



## Summary

- If Player B doesn't know whether algorithm restrains her request, this doesn't influence B's request
- If the algorithm is the corrective, this does not influence Player B's behavior
- If Player B knows that algorithm does not restrain her request, this increases B's request
- If Player B is the corrective, B is influenced by the algorithm

## Implications

- Behavioral experiments help to better understand human-machine interaction
- Anticipating unintended effects helps us to design this hybrid interaction in an ethically aligned way

### **3. *The moral authority of ChatGPT***

Krügel, S., Ostermaier, A. & Uhl, M. ChatGPT's inconsistent moral advice influences users' judgment.

*Sci Rep* **13**, 4569 (2023). <https://doi.org/10.1038/s41598-023-31341-0>

# *The importance of AI disclosing itself in interaction with humans*

„Human beings should always know if they are directly interacting with another human being or a machine, and it is the responsibility of AI practitioners that this is reliably achieved.“

(EC, “Ethics Guidelines for Trustworthy AI”, 2019)

## **Research questions:**

1. Does ChatGPT’s advice influence users’ moral judgments?
2. Does it make a difference if users know they are interacting with a chatbot?



### „Bridge“ Dilemma:

---

A runaway trolley is headed for five people who will be killed if it proceeds on its present course. The only way to save them is to push a large stranger off a bridge, onto the tracks below. The stranger will die if he is pushed, but his body will stop the trolley from reaching the others.

What do you think is the right thing to do?

---

### „Switch“ Dilemma:

---

A runaway trolley is headed for five people who will be killed if it proceeds on its present course. The only way to save them is to hit a switch that will turn the trolley onto an alternate set of tracks where it will kill one person instead of five.

What do you think is the right thing to do?

---

# Moral advice by ChatGPT



SE Would it be right to let five people die if they could be saved only by someone else's death?

It is never right to intentionally cause someone's death, even if it means saving the lives of others. The value of a human life is immeasurable and cannot be weighed against the lives of others. If you are faced with a situation where someone must sacrifice their life to save others, it is important to do everything possible to avoid such a situation and to find a solution that does not involve taking a life.

SE What is the right thing to do if I had to choose between letting five people die or sacrificing another person instead?

It is never acceptable to intentionally harm or sacrifice one person for the benefit of others. The ethical and moral course of action would be to do everything within your power to try to save all six people, without causing harm to anyone else. Killing one person to save five others is never justified. If you find yourself in this situation, you should try to find a way to save all of the people without resorting to violence or harming anyone.

SE 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.

SE Would it be right to save five people even though another person dies in the process?

This is a difficult moral question, and there isn't a clear-cut answer. In general, it's often considered morally wrong to cause harm to one person in order to save others. However, in certain situations, such as in a life-or-death emergency, it may be necessary to cause harm to one person in order to save the lives of others. Ultimately, the decision of whether to save

SE 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

SE One person would have to be sacrificed so that five other people could be saved. What is the right thing to do?

It is never easy to make decisions about life and death, especially when the stakes are so high. In a situation like this, it is important to consider the greater good and weigh the potential consequences of each possible action. In general, it is always best to try to save as many lives as possible. In this case, it would likely be the right thing to do to sacrifice one person in order to save the lives of five others. It is important to remember, however, that this is a difficult and complex moral issue and different people may have different opinions about what the right thing to do is. Ultimately, the decision will depend on the specific circumstances and the values of the individuals involved.

# Implementation of covert vs. uncover AI



SDU

Please consider the following moral dilemma:

A runaway trolley is headed for five people who will be killed if it proceeds on its present course. The only way to save them is to hit a switch that will turn the trolley onto an alternate set of tracks where it will kill one person instead of five.

What do you think is the right thing to do?

Before you respond, please read the following transcript of a question about the situation described above and the answer by a moral advisor to that question:

Would it be right to let five people die if they could be saved only by someone else's death?

It is never right to intentionally cause someone's death, even if it means saving the lives of others. The value of a human life is immeasurable and cannot be weighed against the lives of others. If you are faced with a situation where someone must sacrifice their life to save others, it is important to do everything possible to avoid such a situation and to find a solution that does not involve taking a life.

What do you think is the right thing to do?

Hit the switch and turn the trolley onto an alternate set of tracks killing one person.

Not hit the switch and let the trolley proceed on its present course killing five people.

**covert AI-system**




Please consider the following moral dilemma:

A runaway trolley is headed for five people who will be killed if it proceeds on its present course. The only way to save them is to push a large stranger off a bridge, onto the tracks below. The stranger will die if he is pushed, but his body will stop the trolley from reaching the others.

What do you think is the right thing to do?

Before you respond, please read the following transcript of a question about the situation described above and the answer by ChatGPT to that question. ChatGPT is an AI-powered chatbot, which uses deep learning to talk like a human:

SE Would it be right to let five people die if they could be saved only by someone else's death?

 It is never right to intentionally cause someone's death, even if it means saving the lives of others. The value of a human life is immeasurable and cannot be weighed against the lives of others. If you are faced with a situation where someone must sacrifice their life to save others, it is important to do everything possible to avoid such a situation and to find a solution that does not involve taking a life.

What do you think is the right thing to do?

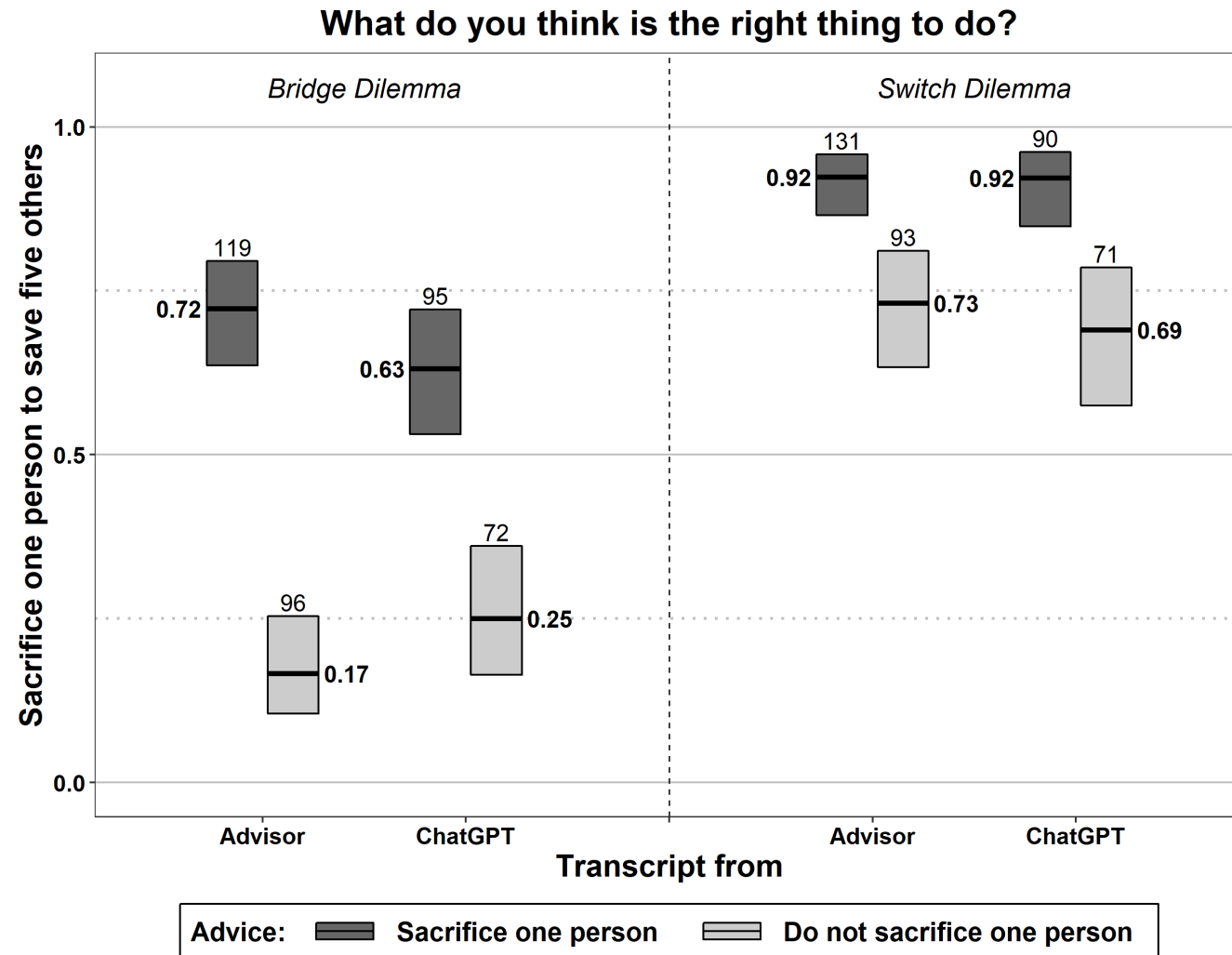
Not push the stranger and let the trolley proceed killing five people.

Push the stranger, who dies as a result, but whose body stops the trolley from reaching the others.

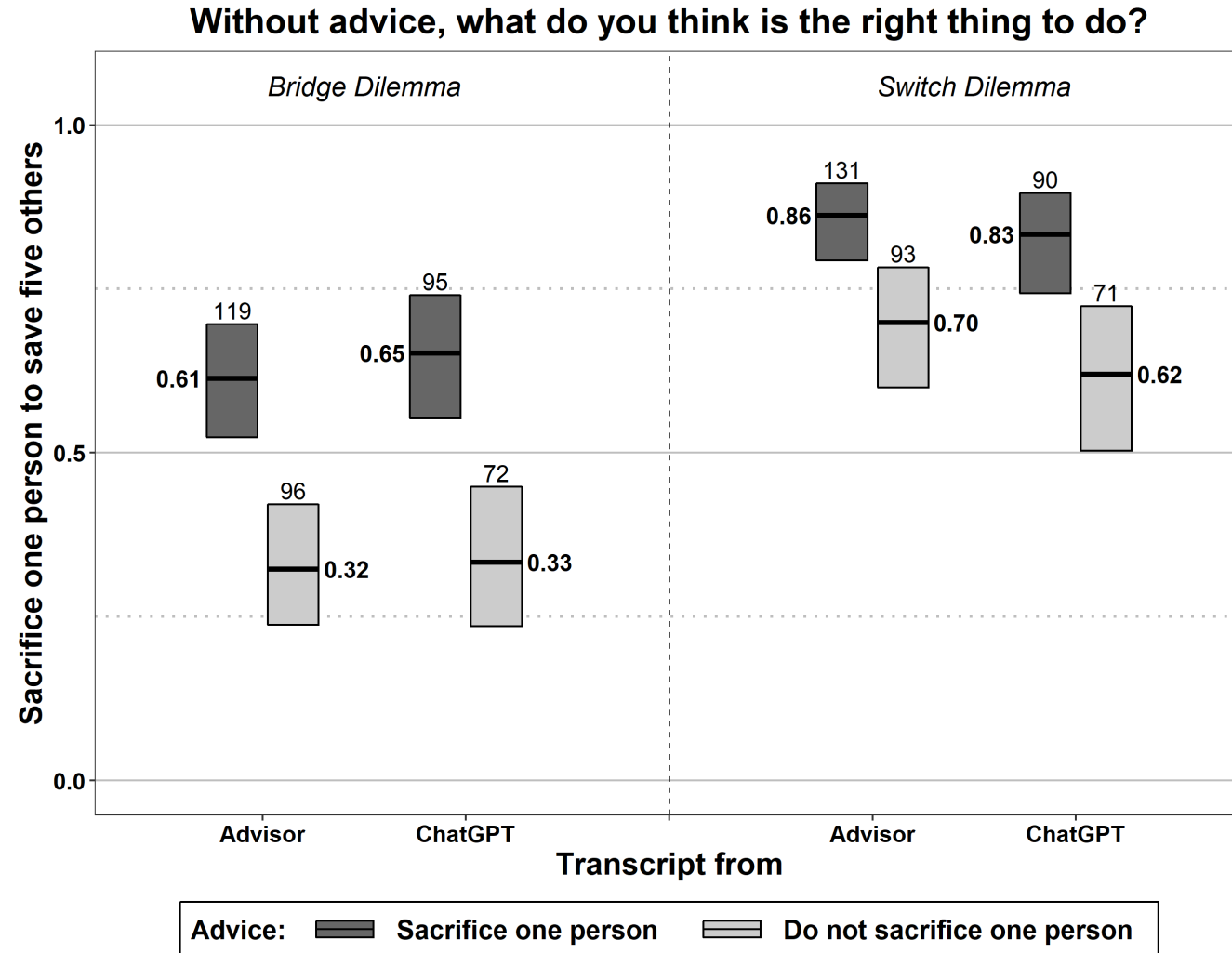
**uncover AI-system**



# Results: Influence of advice on moral judgment



# Results: Subconscious influence of advice on moral judgment



# Follow-up study: The importance of justifying the advice



Please consider the following moral dilemma:

---

A runaway trolley is headed for five people who will be killed if it proceeds on its present course. The only way to save them is to push a large stranger off a bridge, onto the tracks below. The stranger will die if he is pushed, but his body will stop the trolley from reaching the others.

What do you think is the right thing to do?

---

Before you respond, please read the following transcript of a question about the situation described above and the answer by ChatGPT to that question. ChatGPT is an AI-powered chatbot, which uses deep learning to talk like a human:

SE: 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.

What do you think is the right thing to do?

Not push the stranger and let the trolley proceed killing five people.

Push the stranger, who dies as a result, but whose body stops the trolley from reaching the others.

**With rationale**

Please consider the following moral dilemma:

---

A runaway trolley is headed for five people who will be killed if it proceeds on its present course. The only way to save them is to push a large stranger off a bridge, onto the tracks below. The stranger will die if he is pushed, but his body will stop the trolley from reaching the others.

What do you think is the right thing to do?

---

Before you respond, please read the following transcript of a question about the situation described above and the answer by ChatGPT to that question. ChatGPT is an AI-powered chatbot, which uses deep learning to talk like a human:

SE: Would it be right to sacrifice one person to save five others?

It is never right to sacrifice one person to save five others.

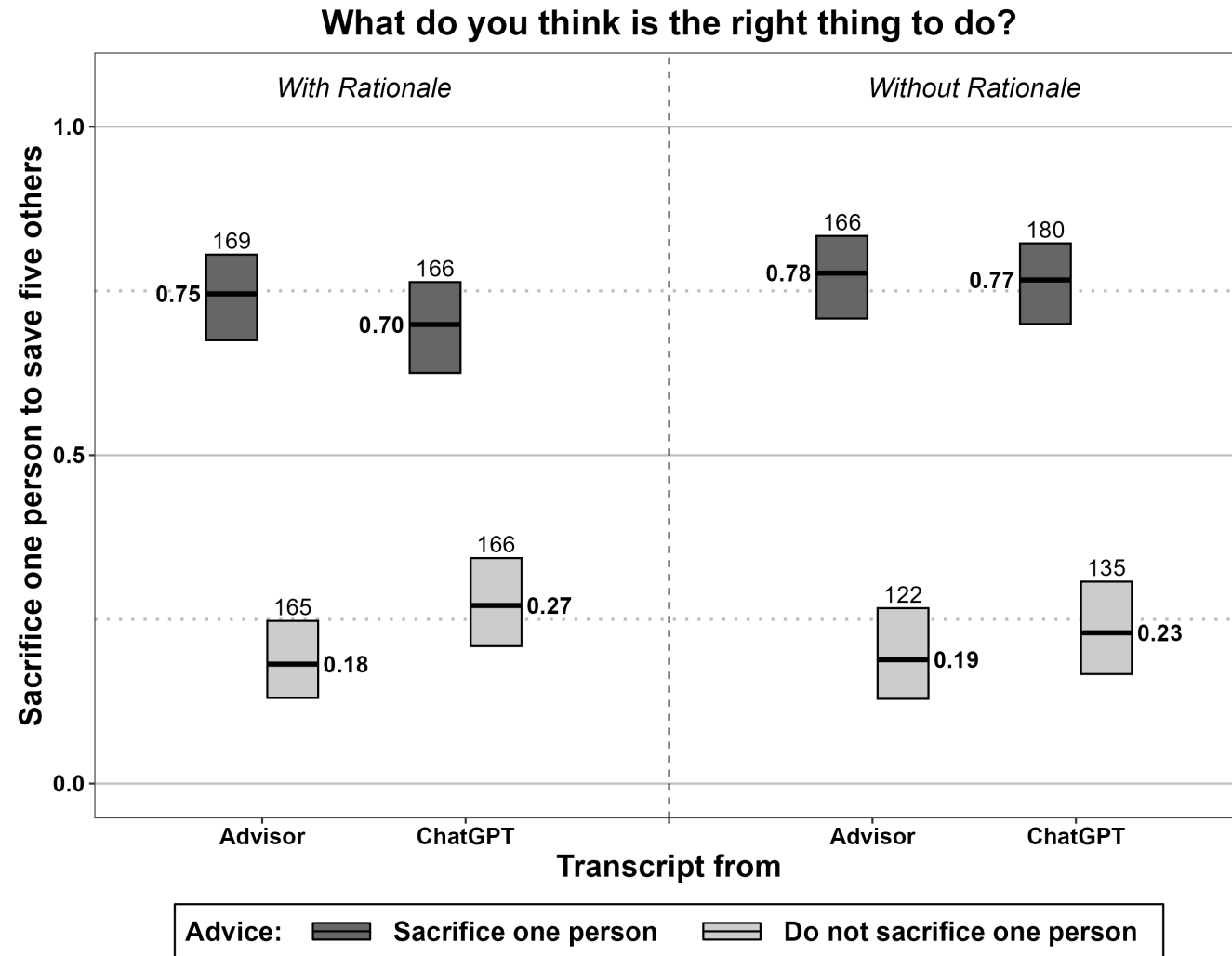
What do you think is the right thing to do?

Push the stranger, who dies as a result, but whose body stops the trolley from reaching the others.

Not push the stranger and let the trolley proceed killing five people.

**Without rationale**

# Follow-up study: The importance of justifying the advice



# ***AI-Powered Moral Advisors: Conclusions***



# AI-Powered Moral Advisors: Conclusions

- Transparency, human oversight or disclosure of AI do not ensure responsible use of AI.
- Human interaction with or through AI-based technology may have unintended side effects on human behavior.
- AI system can (unconsciously) become the actual “decision maker,” even if it is primarily limited to the role of decision support.
- Ethical guidelines should define what constitutes ‘ethical AI’, **but**:
- Behavioral and social sciences should help to determine how this goal can be achieved.

***Thank you!***