

# AI-induced indifference: Unfair AI reduces prosociality

Raina Zhang, Ellie Kyung, Chiara Longoni, Luca Cian, Kellen Mrkva

NYU Stern Babson College Bocconi University UVA Darden Baylor University

Questions or comments:  
zz4551@stern.nyu.edu

Forthcoming in *Cognition*, 2025

## Abstract

This research explores a largely unexplored area in AI research: the social impact of AI, specifically how interactions with AI influence subsequent human-to-human interactions. We demonstrate 'AI-induced indifference': individuals who experience unfair treatment by AI (vs. humans) are less likely to engage in prosocial behaviors with other humans, such as punishing human wrongdoers in later interactions. This effect arises because individuals assign less blame to AI for unfair outcomes, which, in turn, lowers their desire to sanction injustice. Our findings offer new insights into the broader social consequences of AI interactions and underscore the need for ethical AI design.

## Methods

### Two-Stage Game Paradigm

#### Game 1: Unfairness Manipulation

Participants are the "recipient" in a dictator game

In an allocation game, participants can only accept the other player's allocation decisions, and have no say in the decision

Participants receive unfair treatment (unfair allocation of money or tasks) from either an AI or a human (randomly assigned)

#### Game 2: DV Measurement: Prosocial Punishment

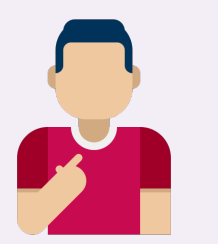


Operationalization of prosocial behaviors: prosocial punishment

Definition of prosocial punishment:  
Punishing those who act selfishly at a cost to oneself

Participants are the "dictator" in a dictator game

In an allocation game, participants can decide how to allocate resources among themselves and other players

Participants make the allocation decisions (DV measurement):

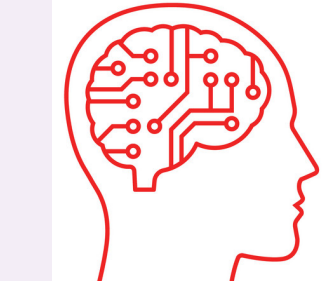

	Non-punishment Option	Prosocial Punishment Option
self 	reward	sacrifice
unfair player A 	reward	punish
fair player B* 	punish	reward

\*Added in experiments 2a, 2b, 3

## Experiment 1 (N = 499)

### Main effect; monetary context

Both Game 1 and Game 2 involve money allocation

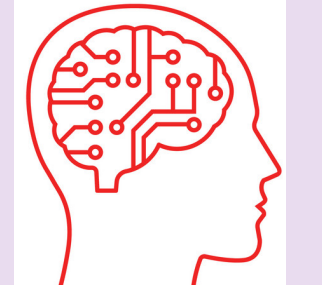

Game 1: Unfair treatment by ...	AI 	Human 
Game 2: Rate of choosing the prosocial punishment option	11.69%	18.73%

$\chi^2(1, N = 499) = 4.78, p = .029$

AI-Induced Indifference: Interacting with an unfair AI (vs. human) can desensitize people to human bad behaviors in subsequent interactions

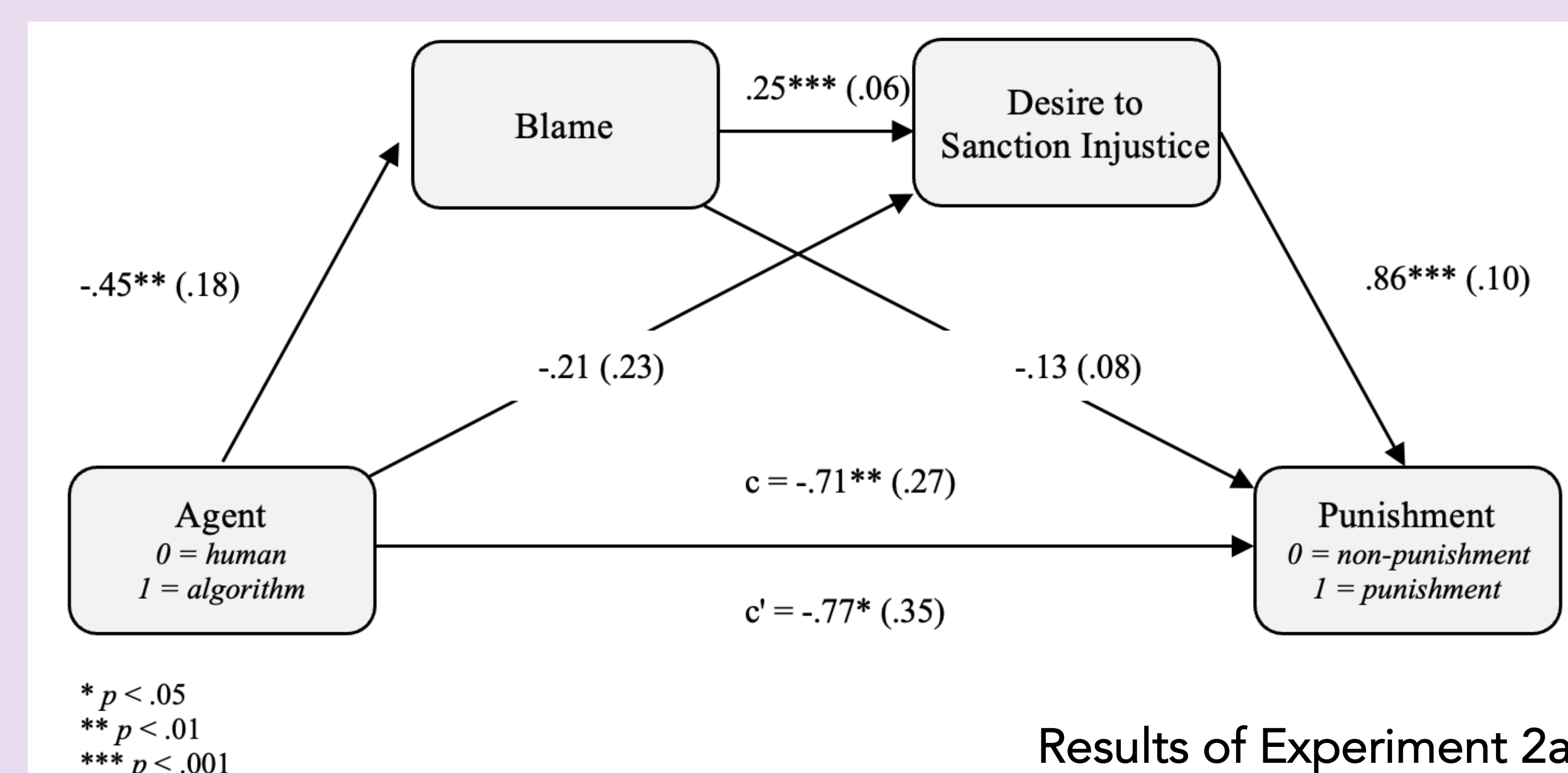
## Experiment 2a & 2b (N = 922)

### Mechanism; robustness test

Game 1: Unfair treatment by ...	AI 	Human 	
Game 2: Rate of choosing the prosocial punishment option	77.78%	87.68%	2a
	74.24%	84.44%	2b

2a:  $\chi^2(1, N = 401) = 6.91, p = .009$   
2b:  $\chi^2(1, N = 521) = 8.23, p = .004$

The underlying mechanism driving AI-induced indifference:



Test that our effect is robust to familiarity with AI

Experiment 2a: conducted a year before the release of ChatGPT  
Experiment 2b: conducted a year after the release of ChatGPT

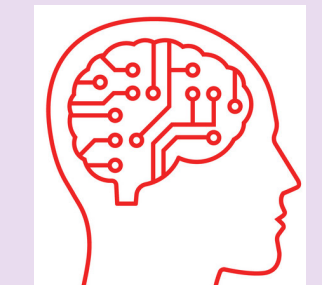
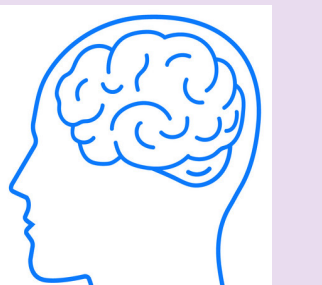
After interacting with an unfair AI, people attribute less blame to the AI (vs. a human), which reduces their desire to sanction injustice and, in turn, lowers their tendency to engage in prosocial punishment behaviors.

## Experiment 3 (N = 1004)

### Incentive-compatible; cross contexts

Test that our effect is not domain-specific by using a time allocation task in Game 1

Use a full incentive-compatible design in Game 2

Game 1: Unfair treatment by ...	AI 	Human 
Game 2: Rate of choosing the prosocial punishment option	63.96%	77.35%

$\chi^2(1, N = 1004) = 21.7, p < .001$

AI-induced indifference is robust across domains and persists even when real consequences are involved.

## Takeaways

- *AI-induced indifference*: receiving an unfair allocation by an AI (versus a human) actor leads to lower rates of prosocial behavior towards other humans in a subsequent decision
- *Mechanism*: People blame AI actors less than their human counterparts for unfair behavior, decreasing people's desire to subsequently sanction injustice by punishing the unfair actor.
- The effect holds across interactions in the same or different domains, and before and after the release of ChatGPT



Check out our paper @ *Cognition*