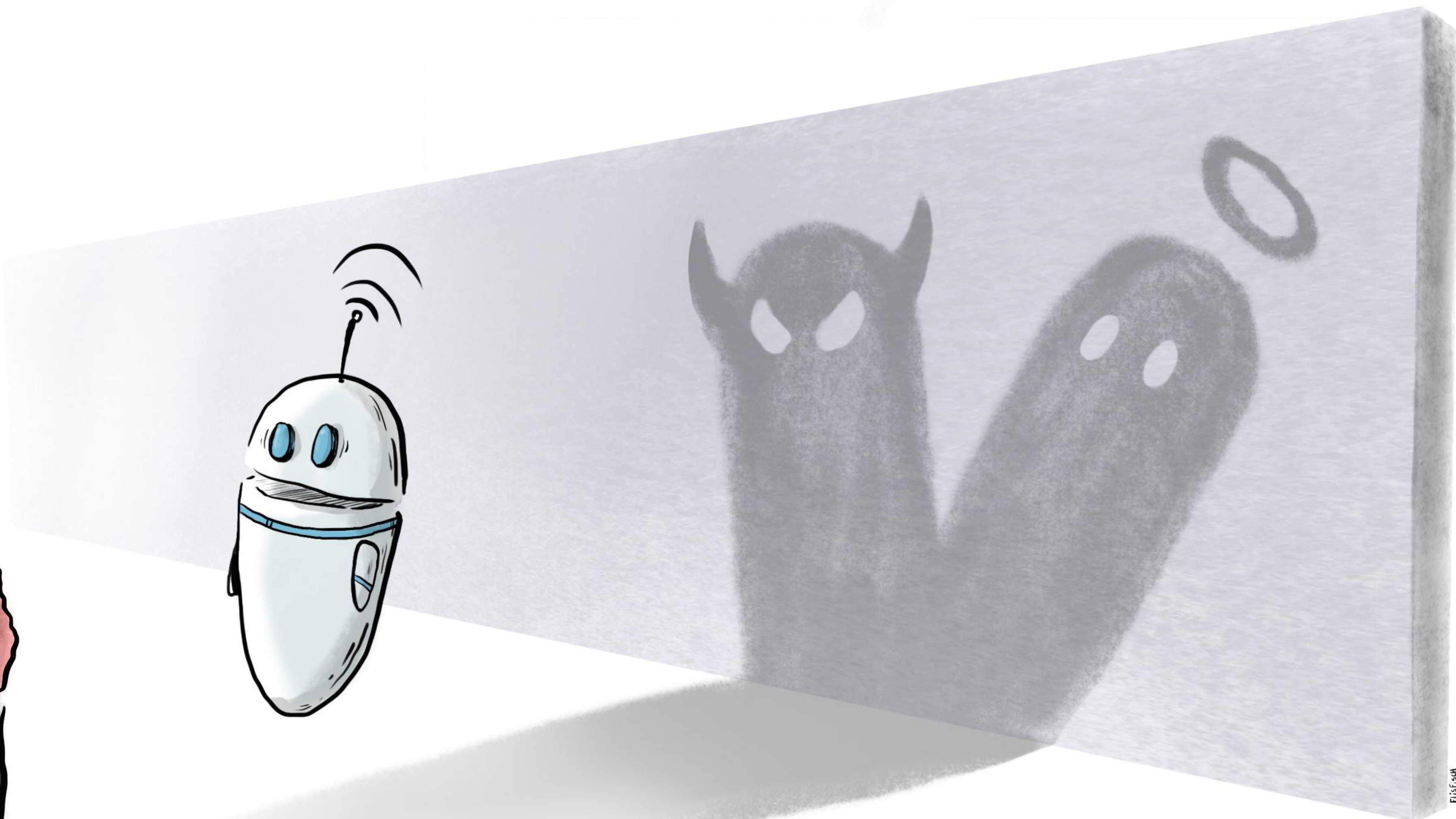
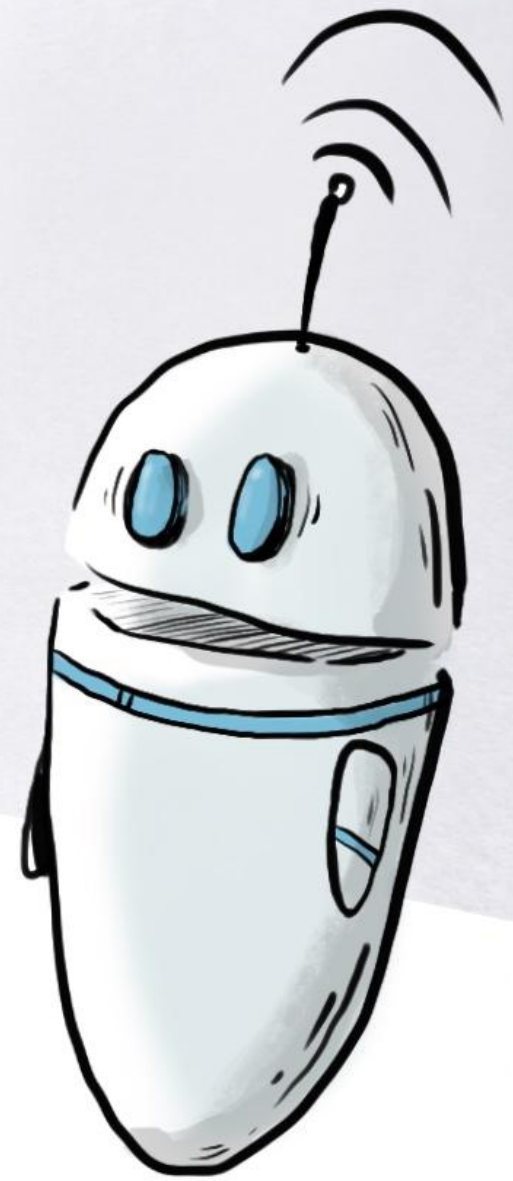
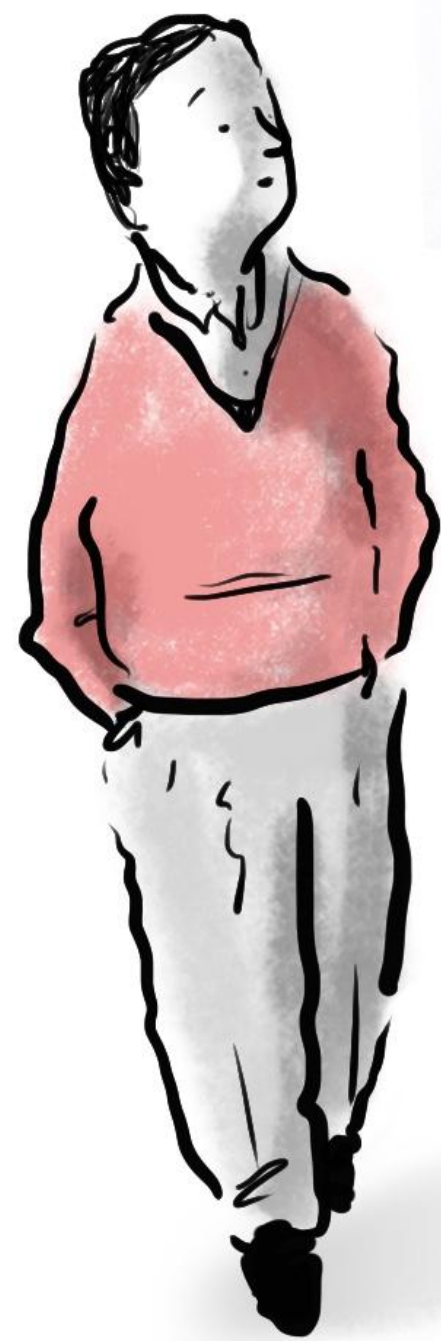




César A. Hidalgo

Director
Center for Collective Learning,
Artificial and Natural Intelligence Institute (ANITI),
University of Toulouse

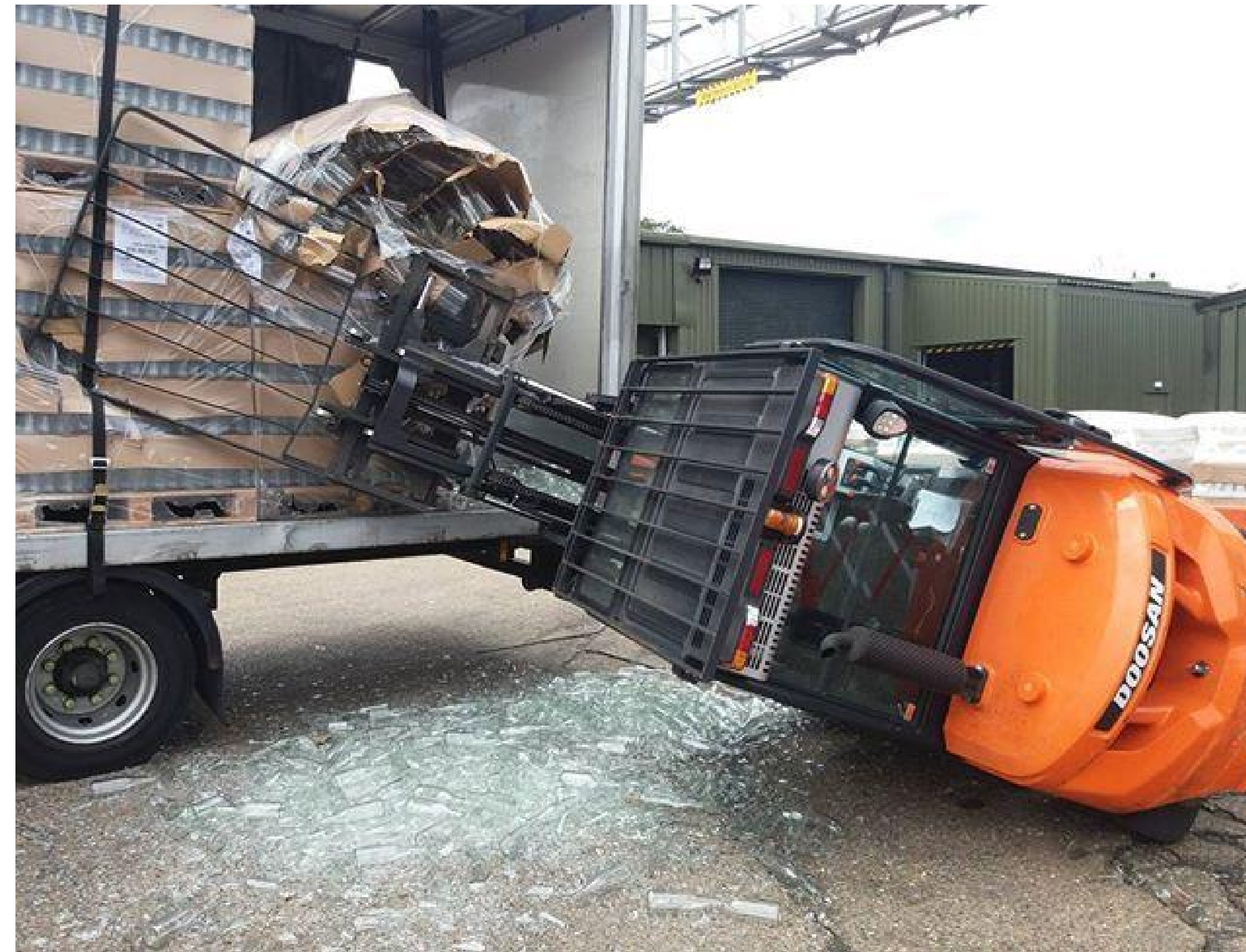
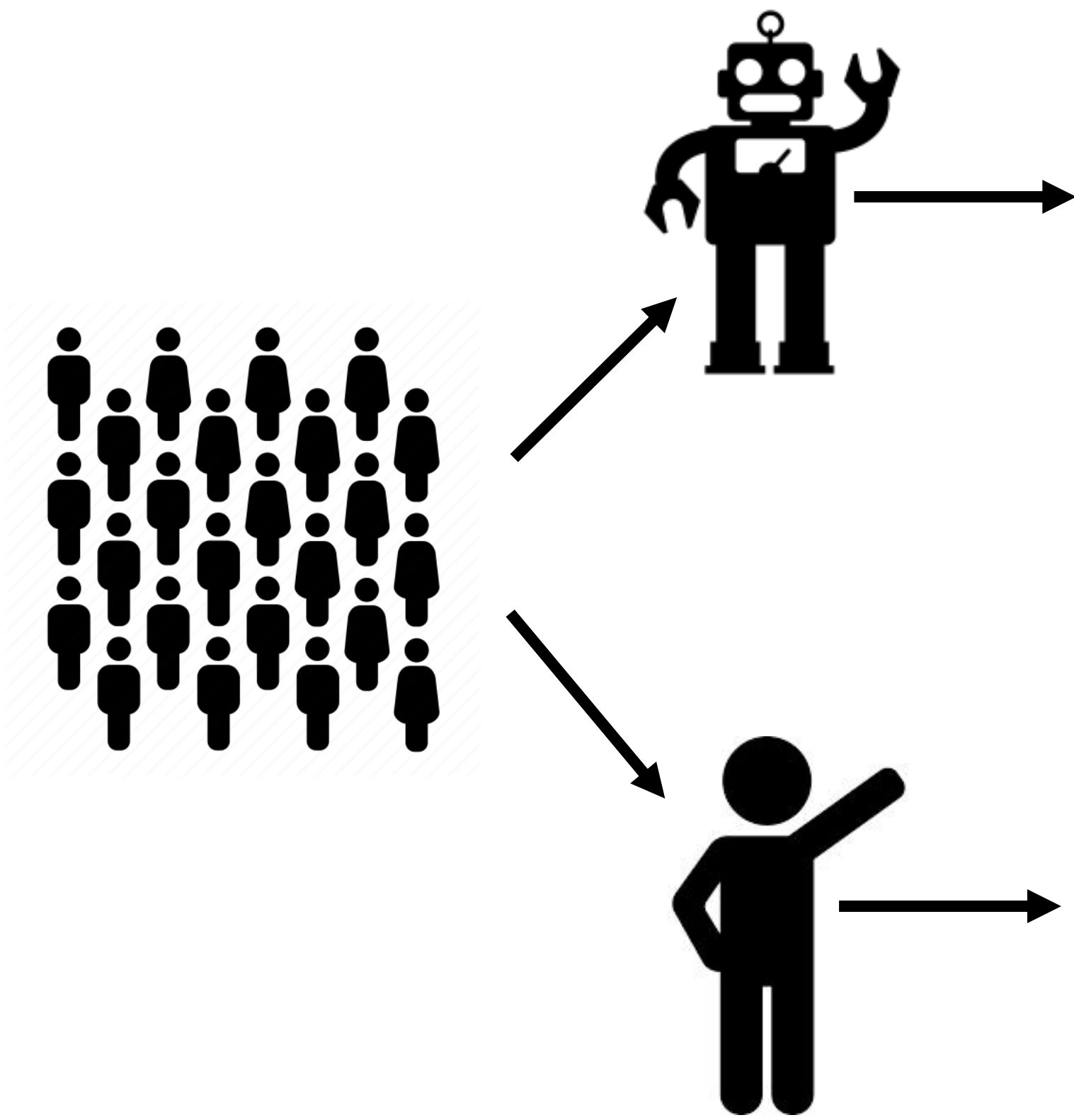
@cesifoti



Randomized Controlled Experiments

Same Mistake

Reaction



$f_h(\dots)$

=?

$f_m(\dots)$

Consider the following scenario

An excavator is digging up a site for a new building. Unbeknownst to the driver, the site contains a grave. The driver does not notice the grave and digs through it. Later, human remains are found.

Would you judge this differently if the driver was a **human** or a **machine**?



People's Reaction to the Scenario

Was the action **harmful**?

Would you **hire** this driver for a similar position?

Was the action **intentional**?

Do you **like** the driver?

How **morally** wrong or right was the driver's action?

Do you agree that the driver should be **promoted** to a position with more responsibilities?

Do you agree that the driver should be replaced with a robot or an algorithm?

[replace different]

Do you agree that the driver should be replaced by another person?

[replace same]

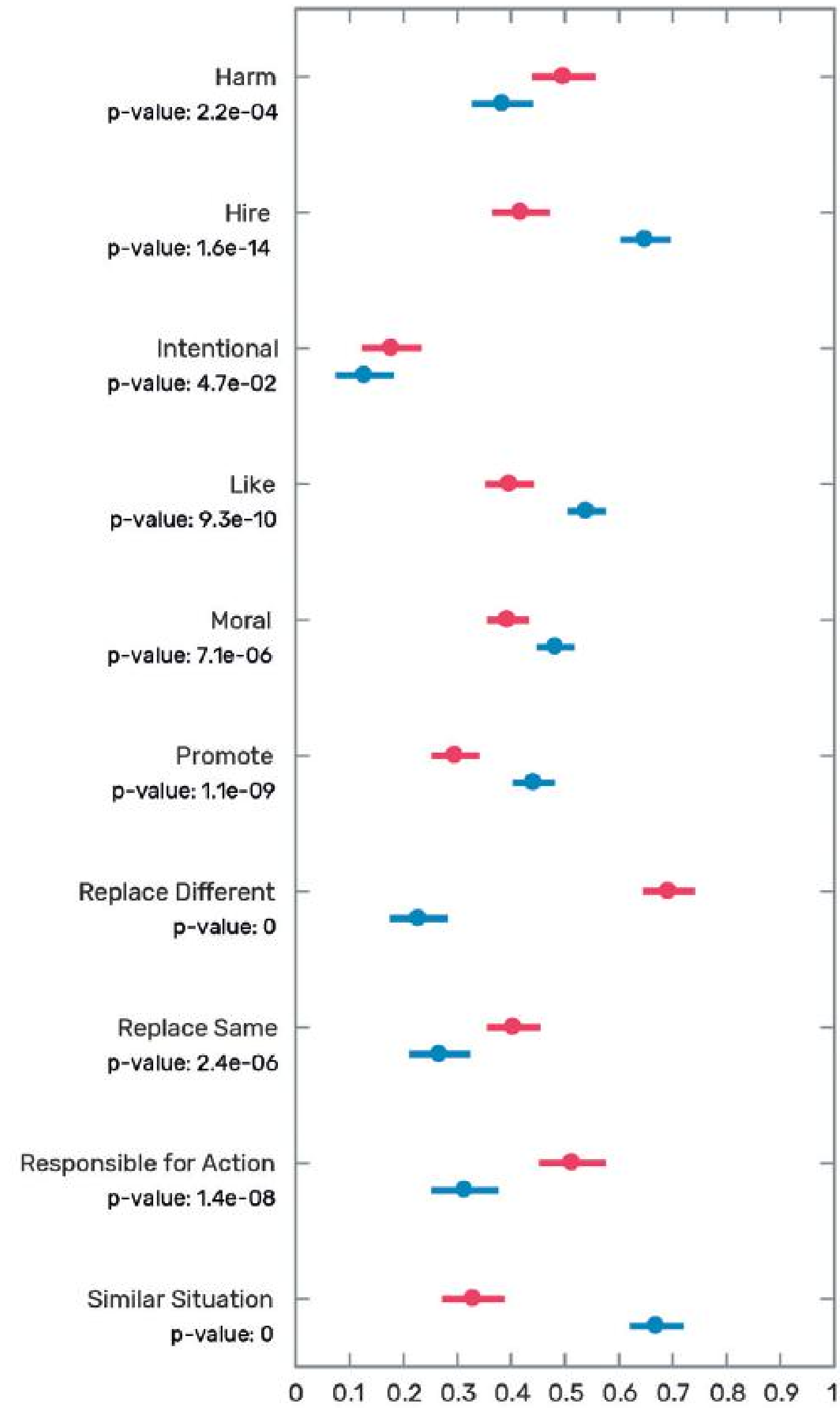
Do you think the driver is **responsible** for unearthing the grave?

If you were in a **similar situation** as the driver, would you have done the same?



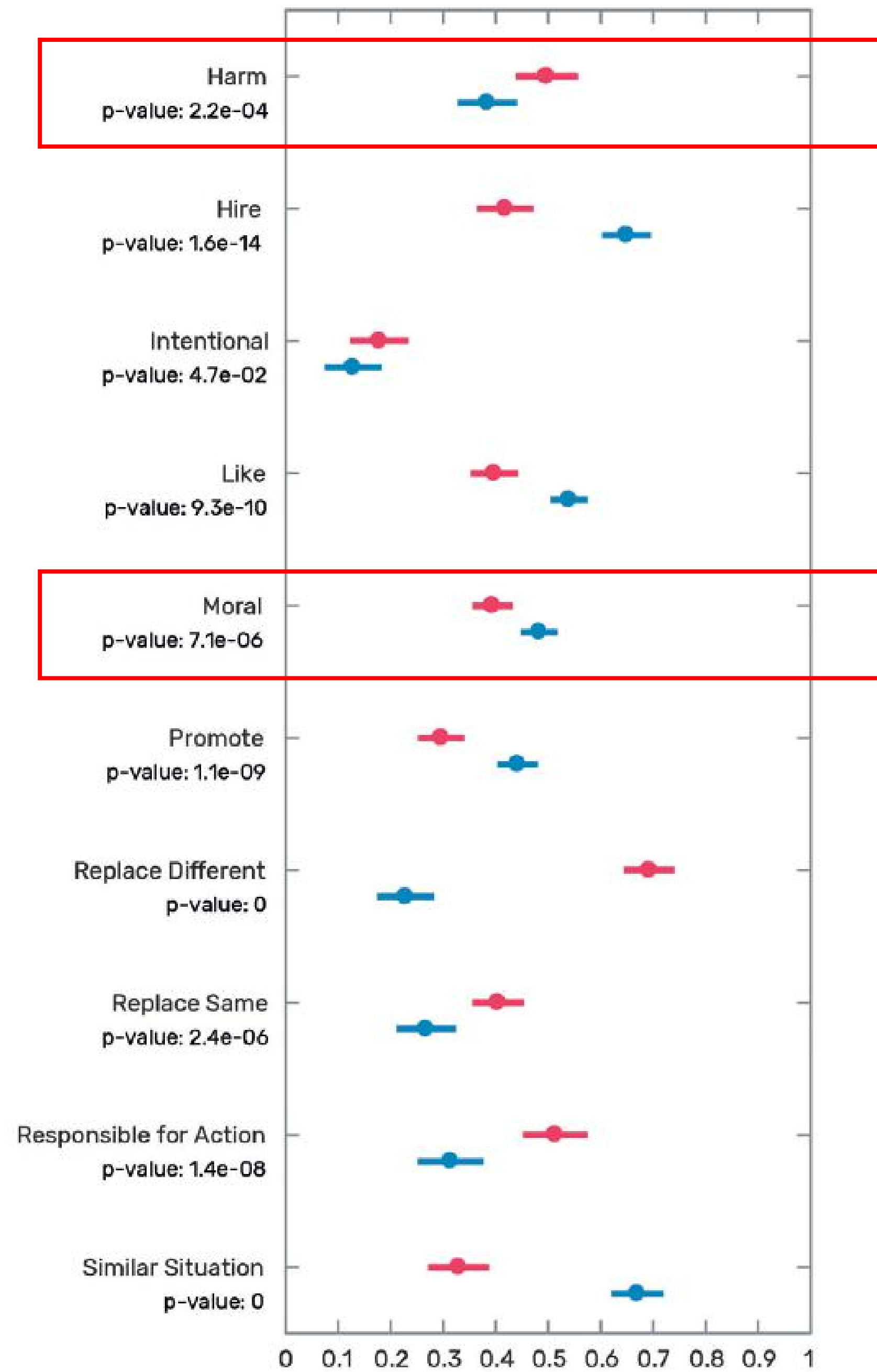
S1

Human
Machine

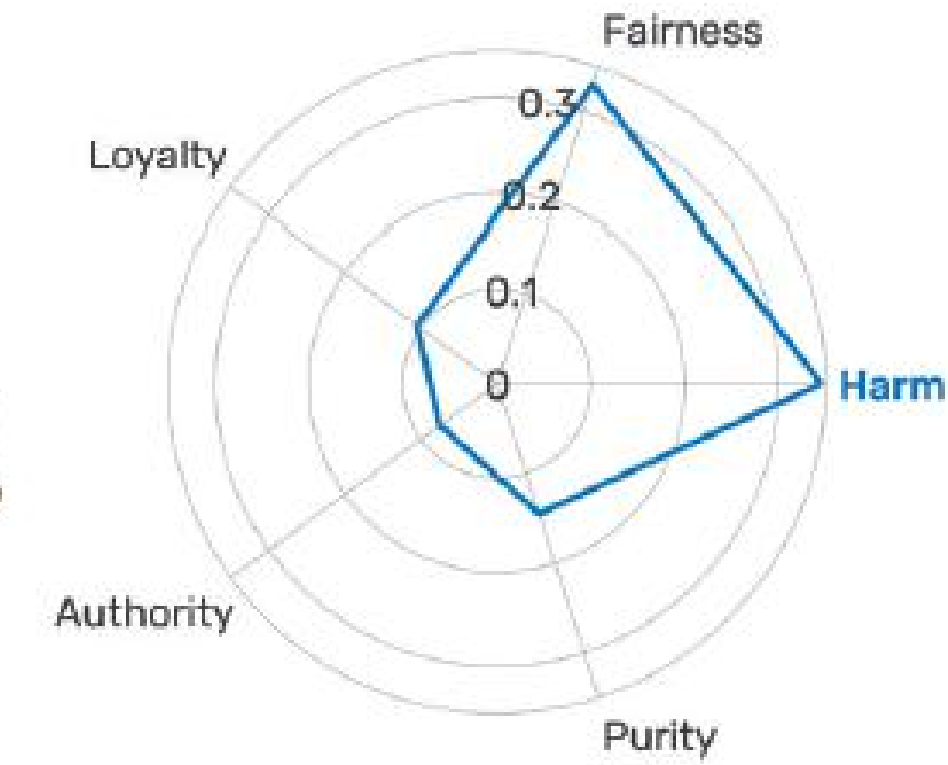
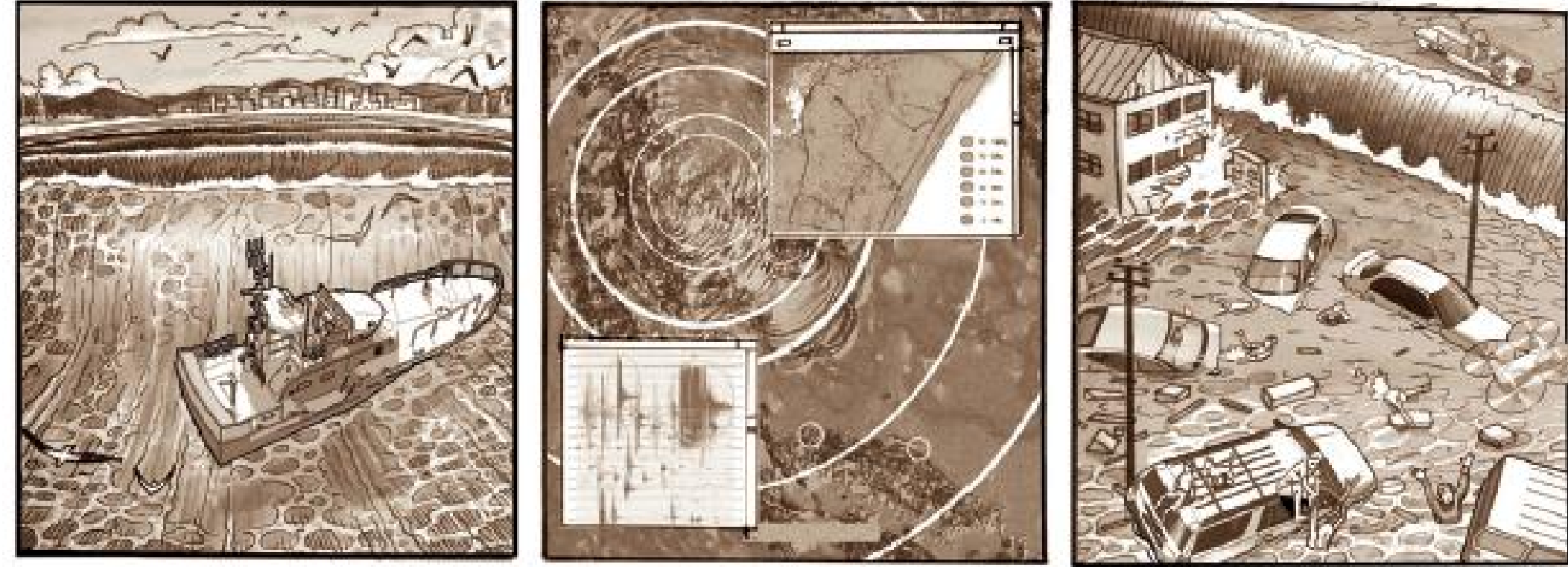


S1

Human
Machine



Consider the following three versions of this moral dilemma:



A large tsunami is approaching a coastal town of 10,000 people, with potentially devastating consequences. The [politician/algorithm] responsible for the safety of the town can decide to evacuate everyone, with a 50 percent chance of success, or save 50 percent of the town, with 100 percent success.

S2

The [politician/algorithm] decides to save everyone, but the rescue effort fails. The town is devastated, and a large number of people die.

S3

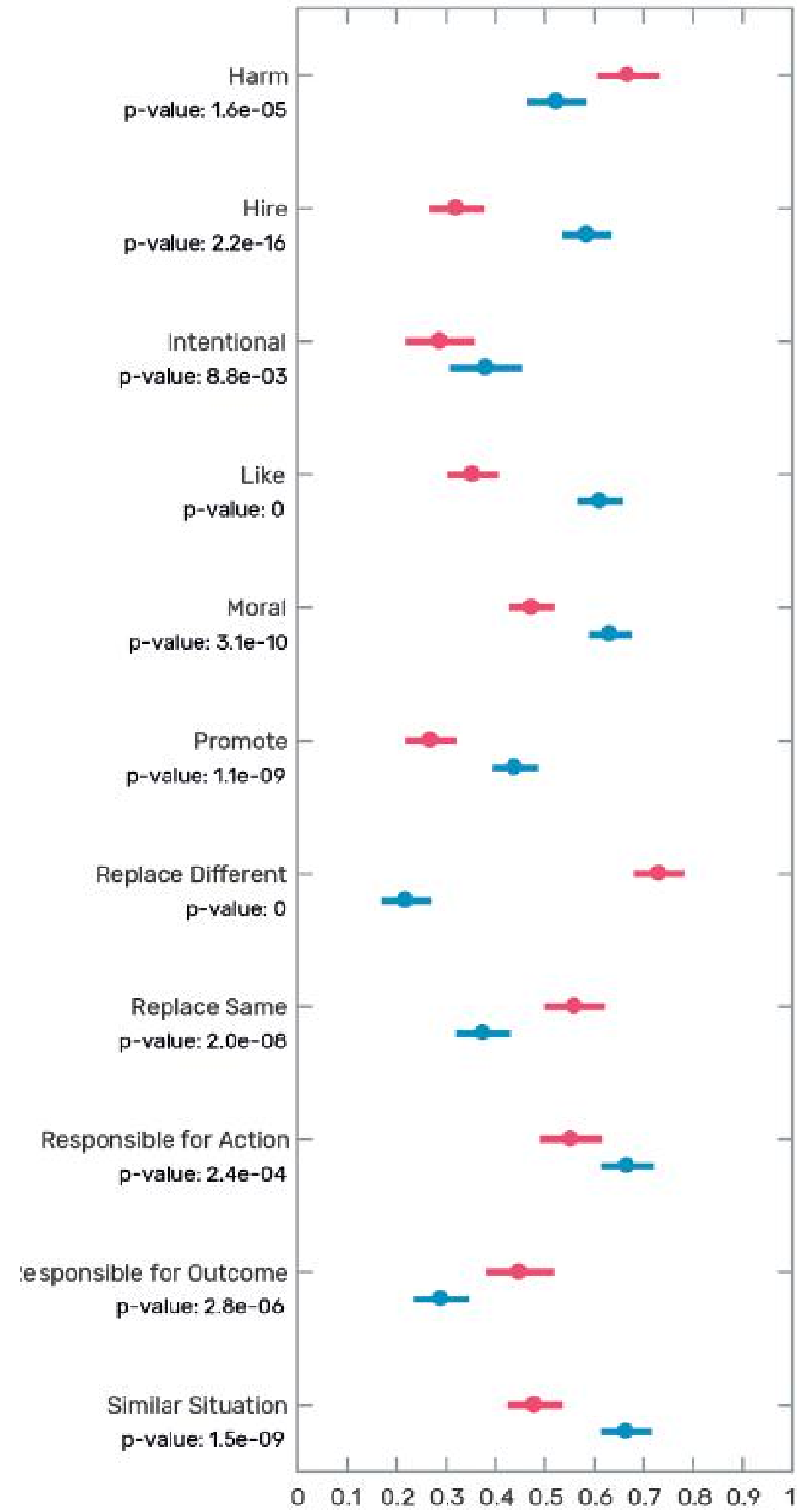
The [politician/algorithm] decides to save everyone, and the rescue effort succeeds. Everyone is saved.

S4

The [politician/algorithm] decides to save 50 percent of the town.

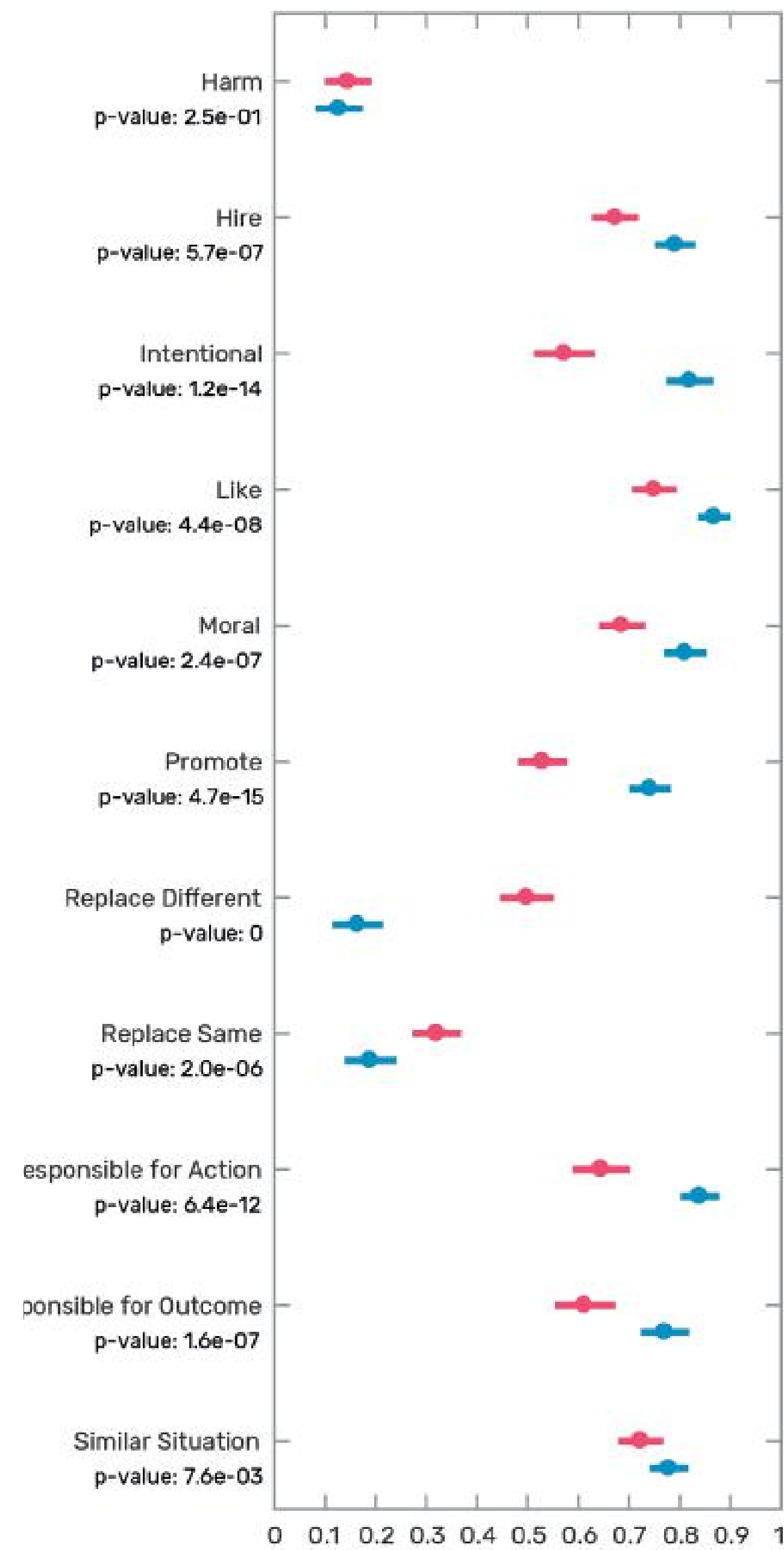
Try to save everyone & fail

S2



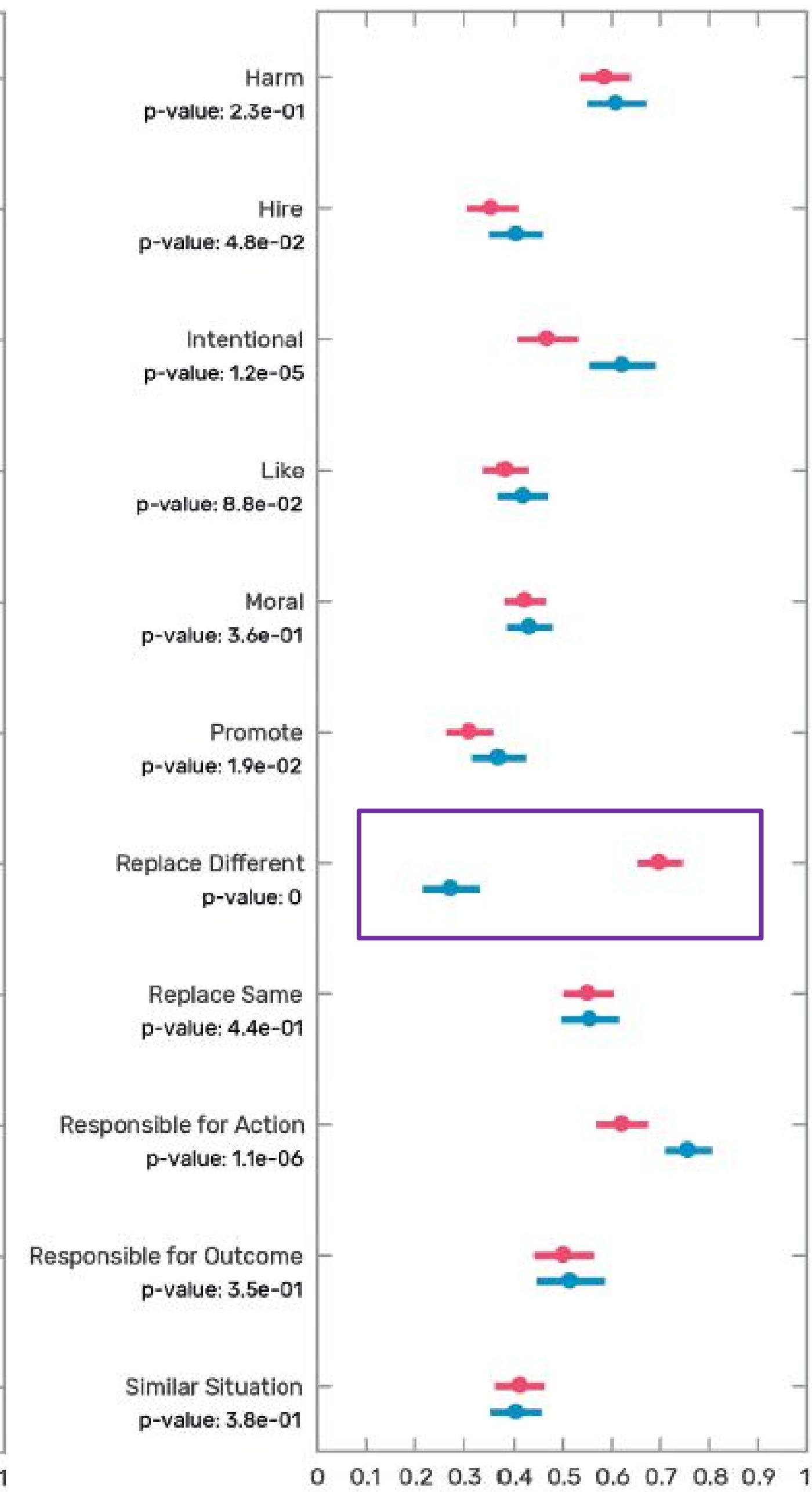
Try to save everyone & succeed

S3



Take Compromise

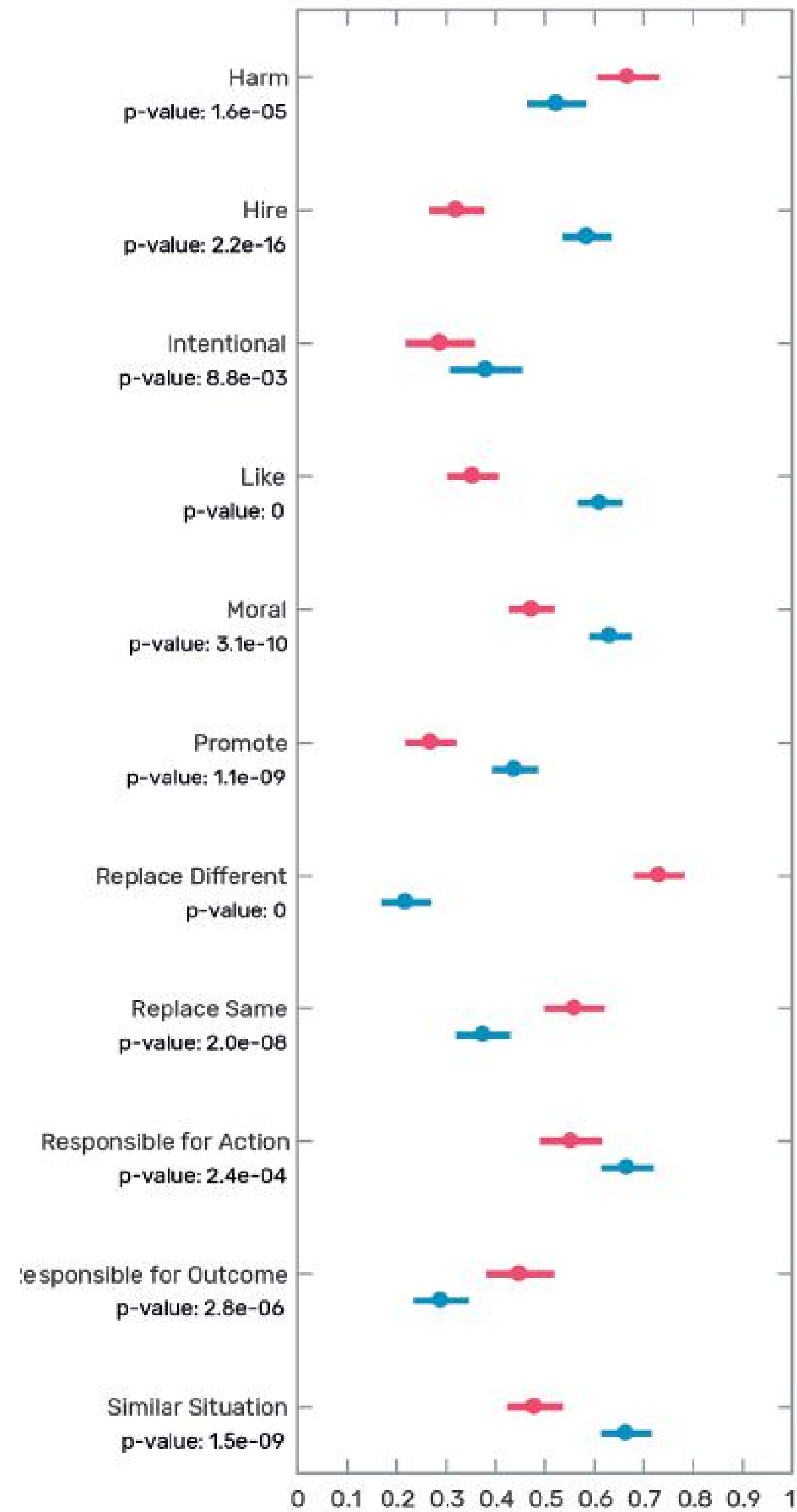
S4



Human
Machine

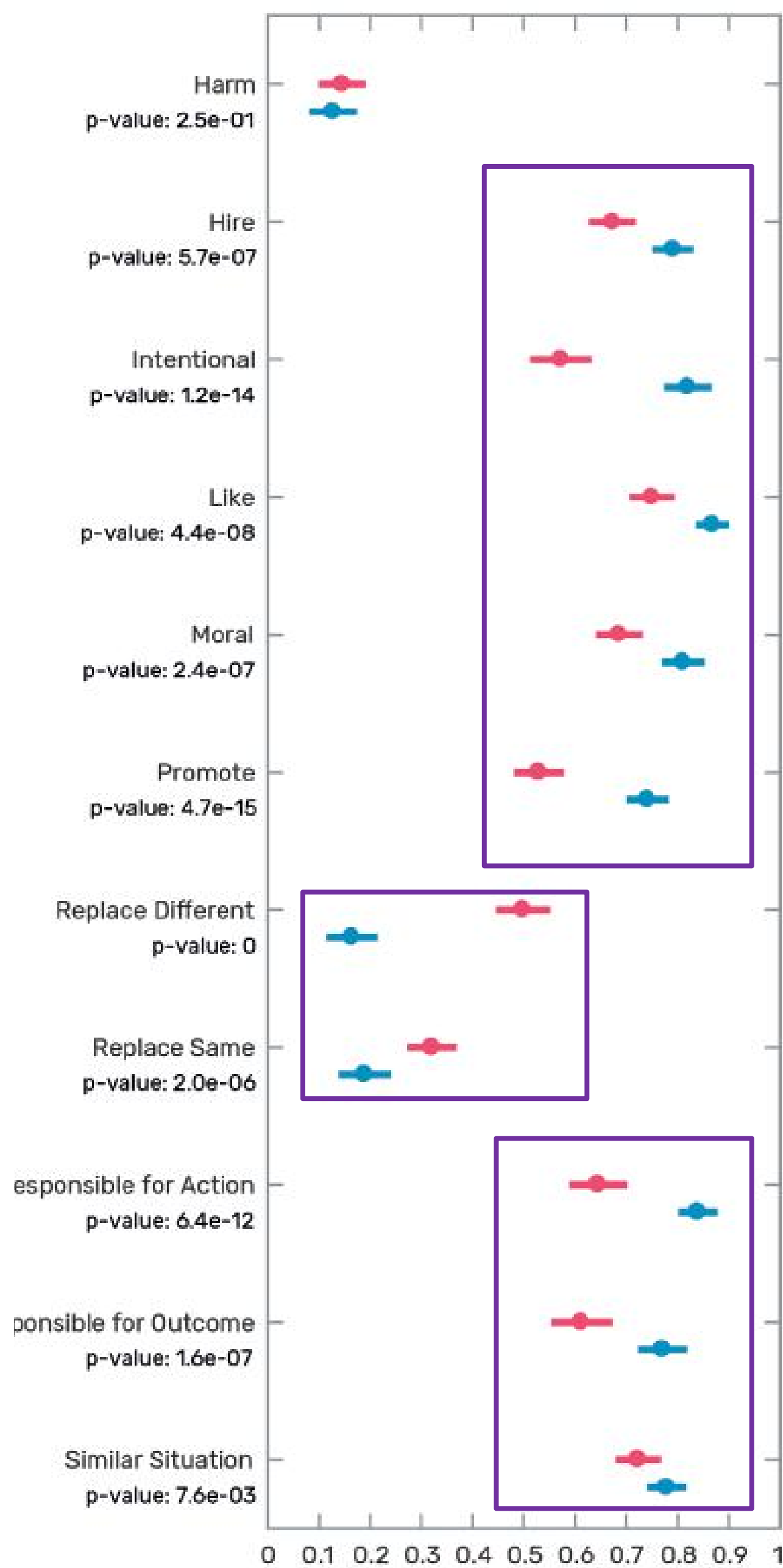
Try to save everyone & fail

S2



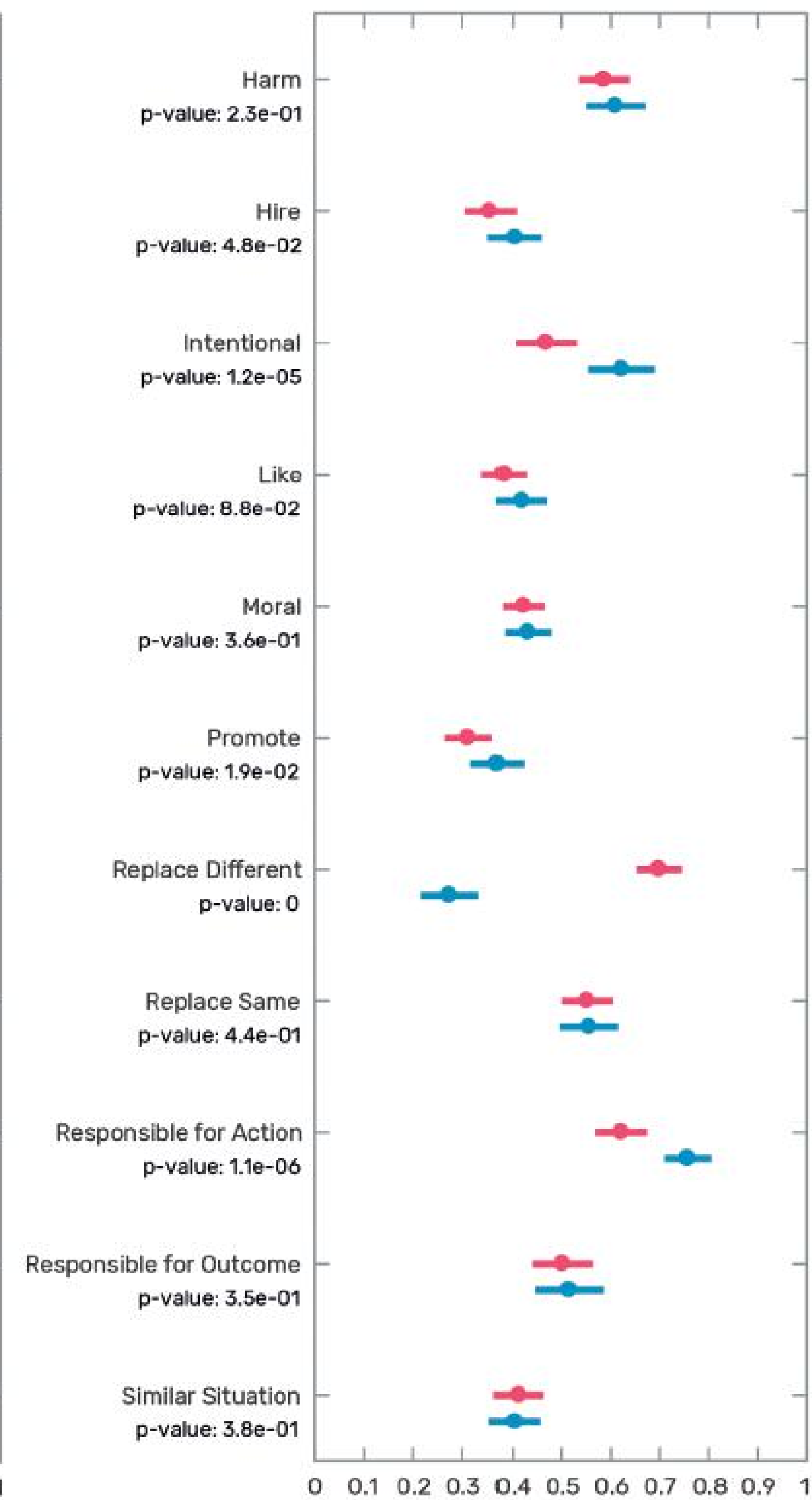
Try to save everyone & succeed

S3



Take Compromise

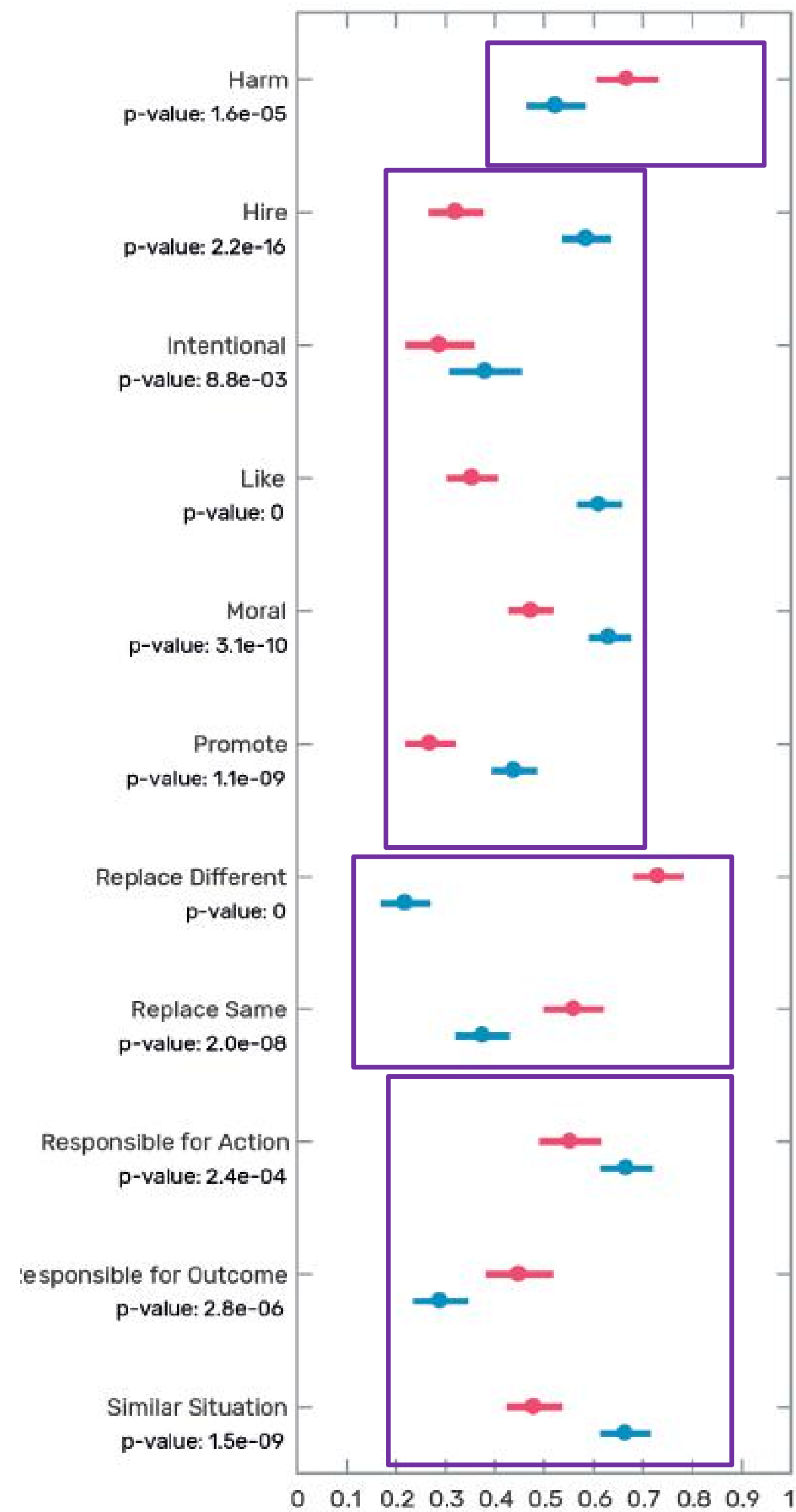
S4



Human
Machine

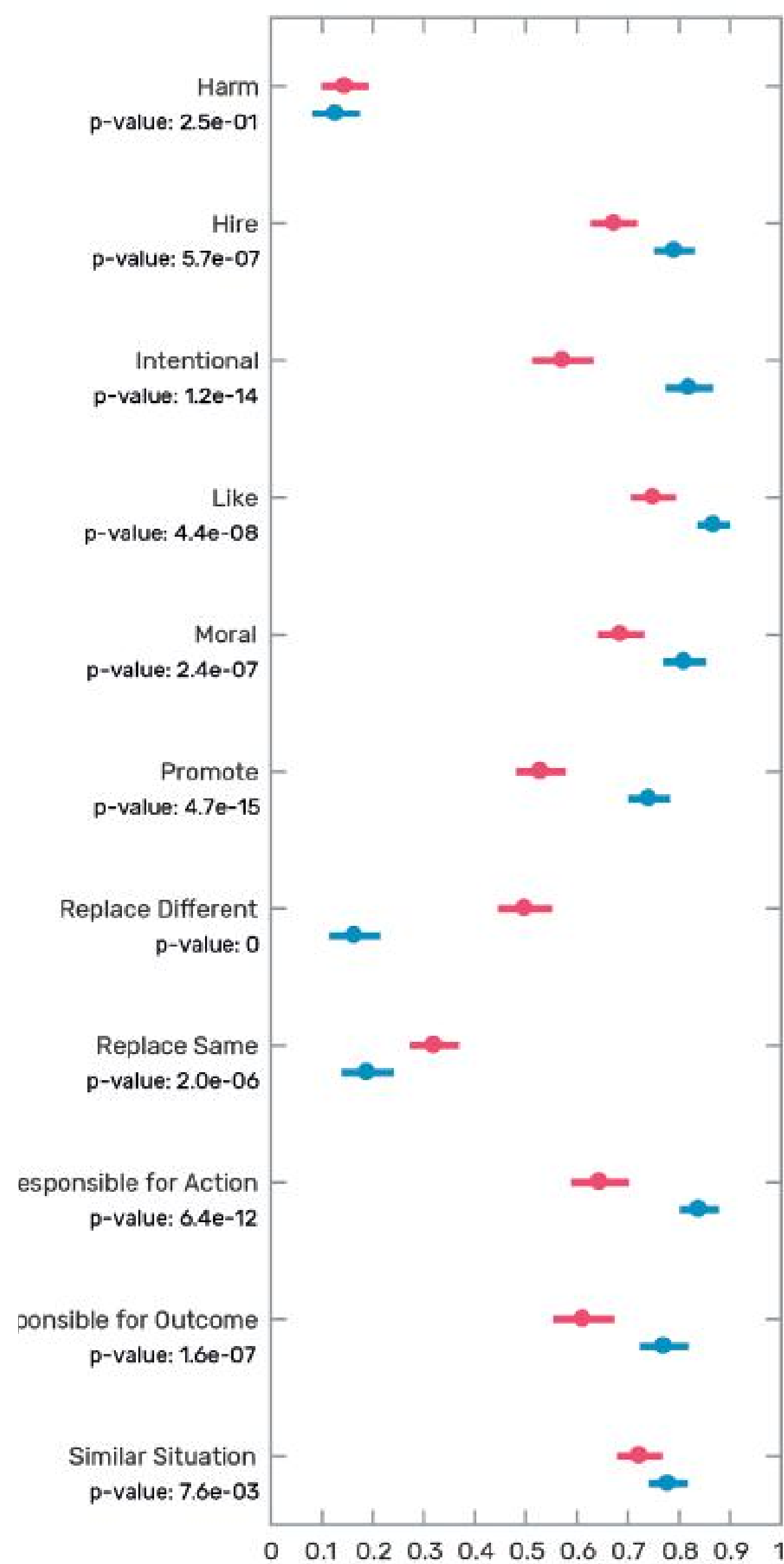
Try to save everyone & fail

S2



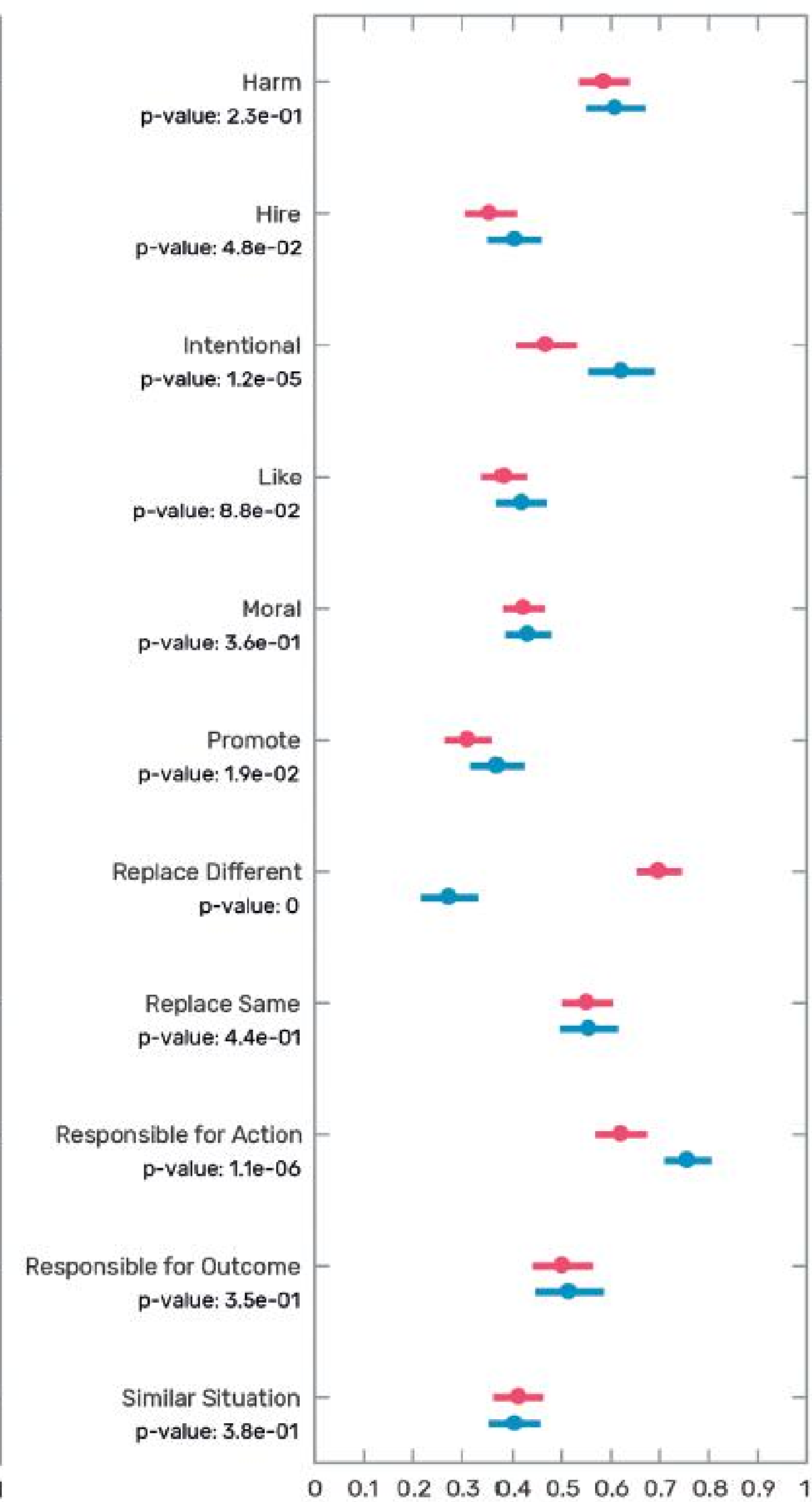
Try to save everyone & succeed

S3

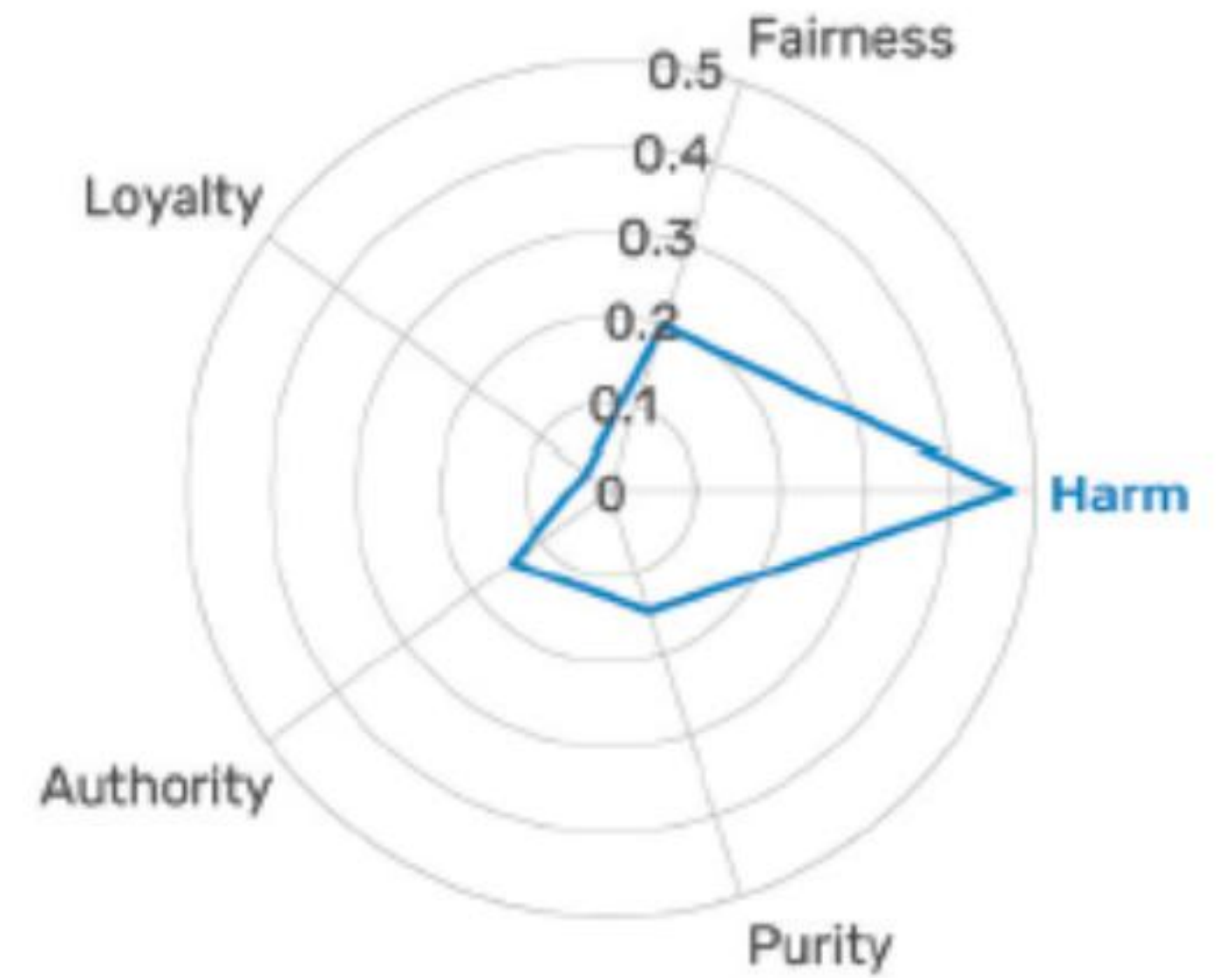
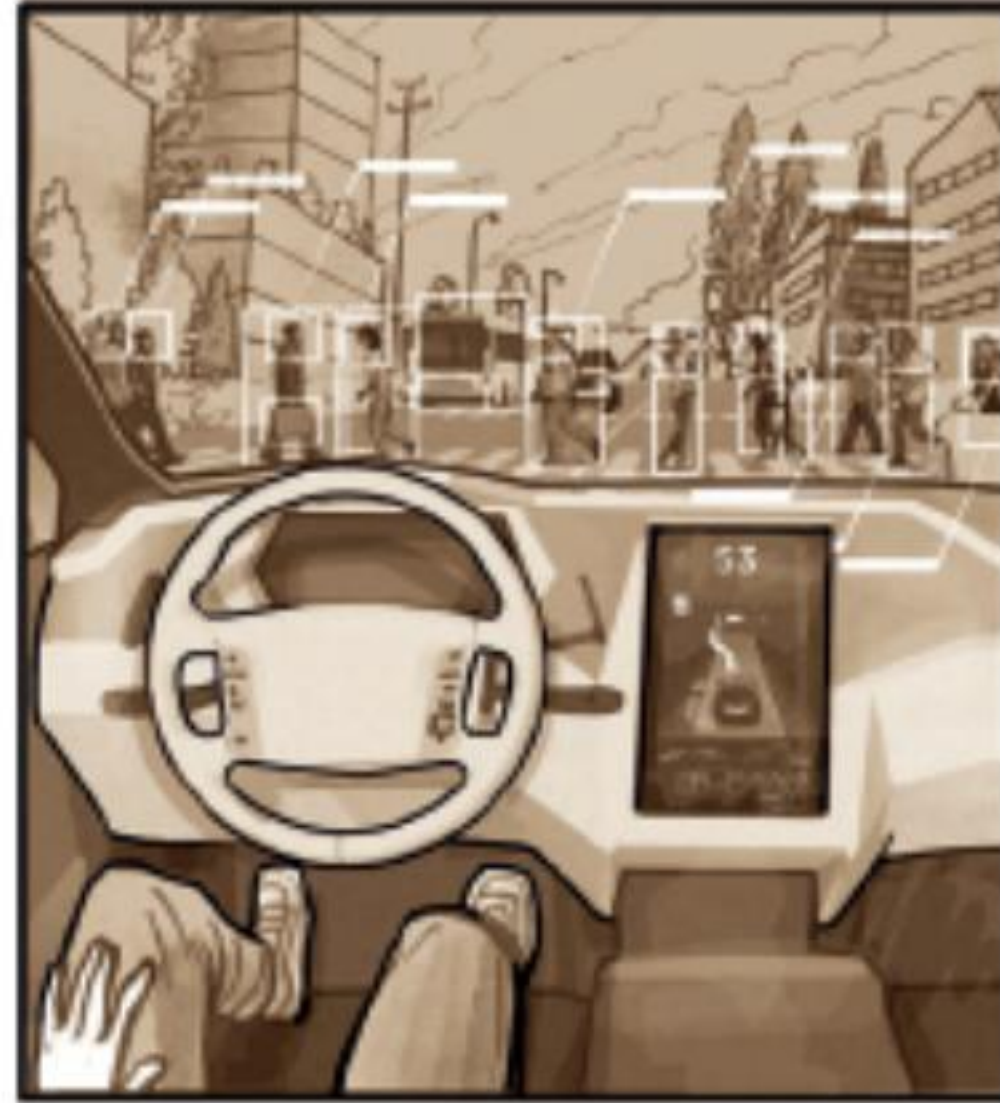


Take Compromise

S4



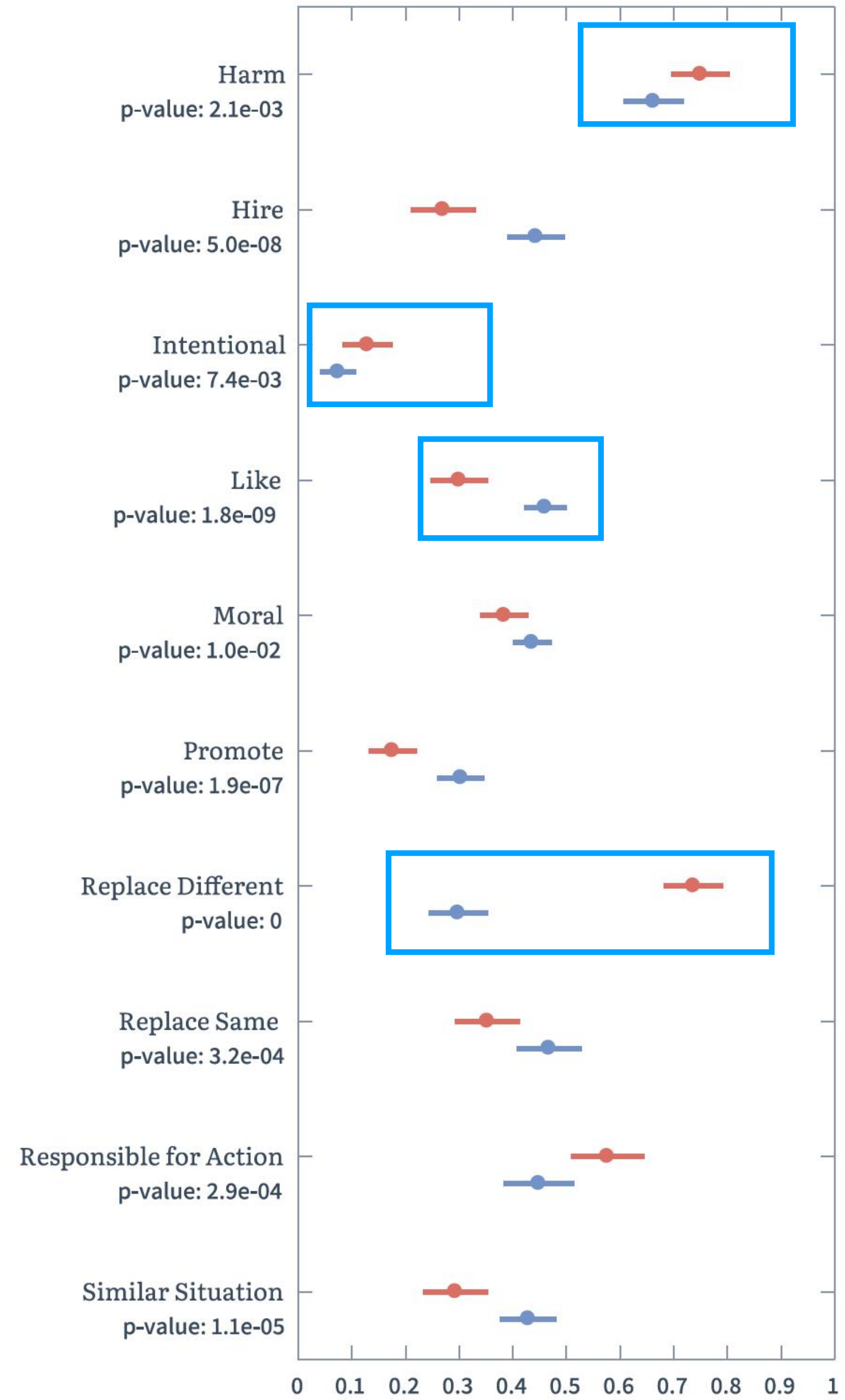
Human
Machine

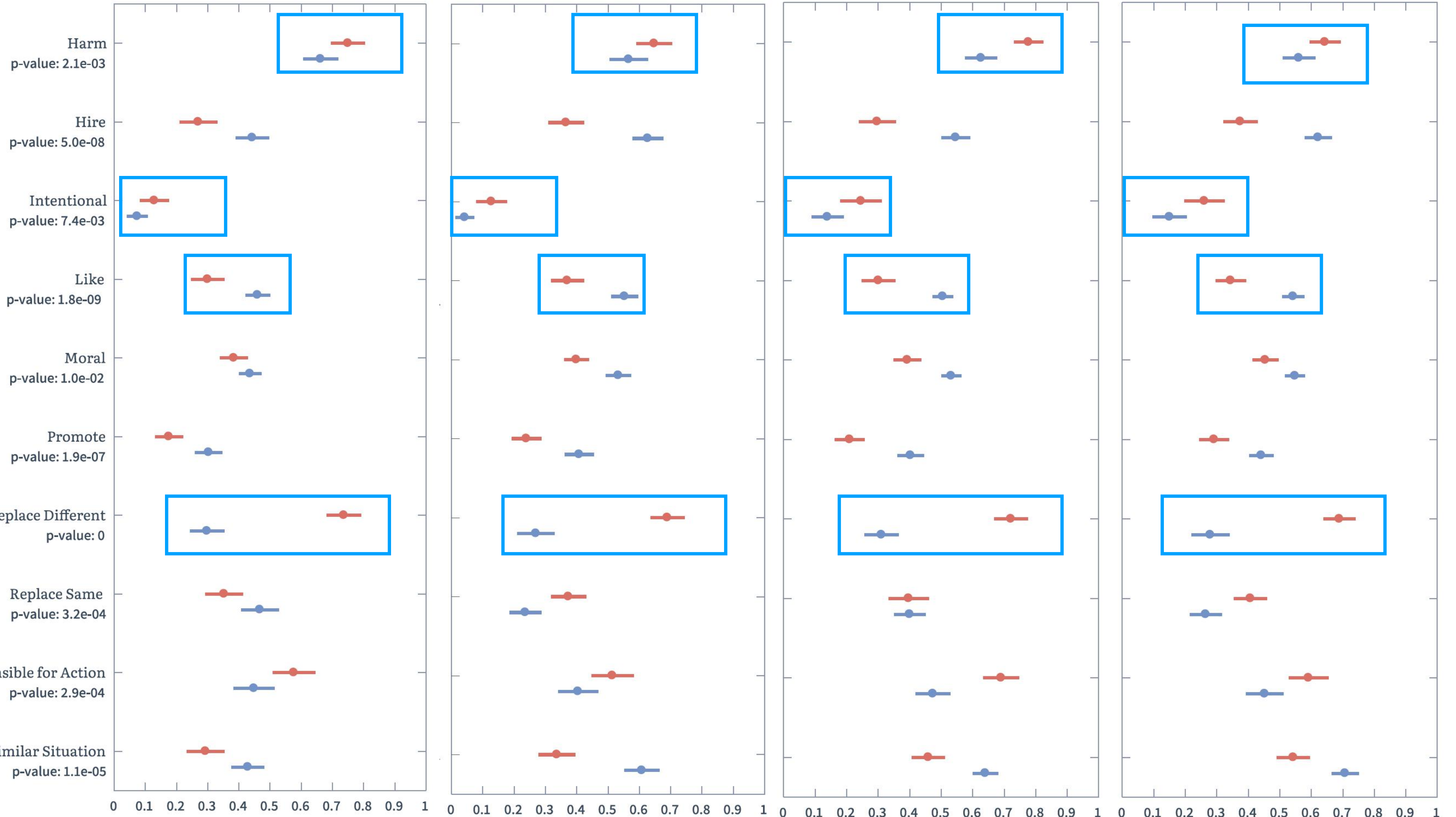


S11

On a sunny spring day, a [driver/driverless car] working for a supermarket chain accidentally runs over a pedestrian who runs in front of the vehicle. The pedestrian is hurt and is taken to the hospital.

SUNNY HUMAN



SUNNY HUMAN**SUNNY DOG****WINDY HUMAN****WINDY DOG**

Intentional Machines?



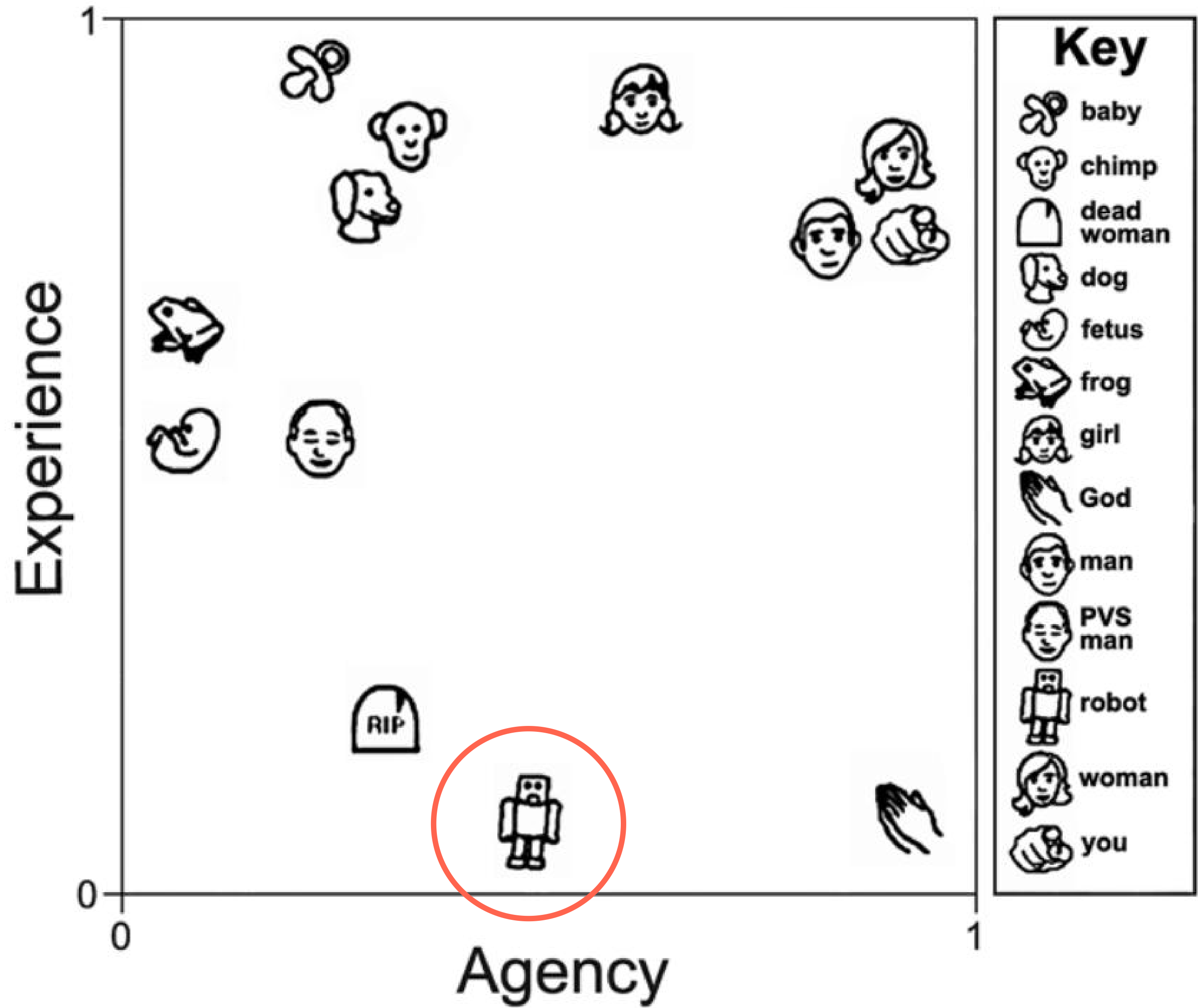
Intention and Agency as a Continuum...

Think of a self-driving car,

designed to protect the driver **or** designed to protect pedestrians at all costs...

Different outcomes, not because of human type agency, but because of behaving as

intended



Do Humans Always
Reject Machines?



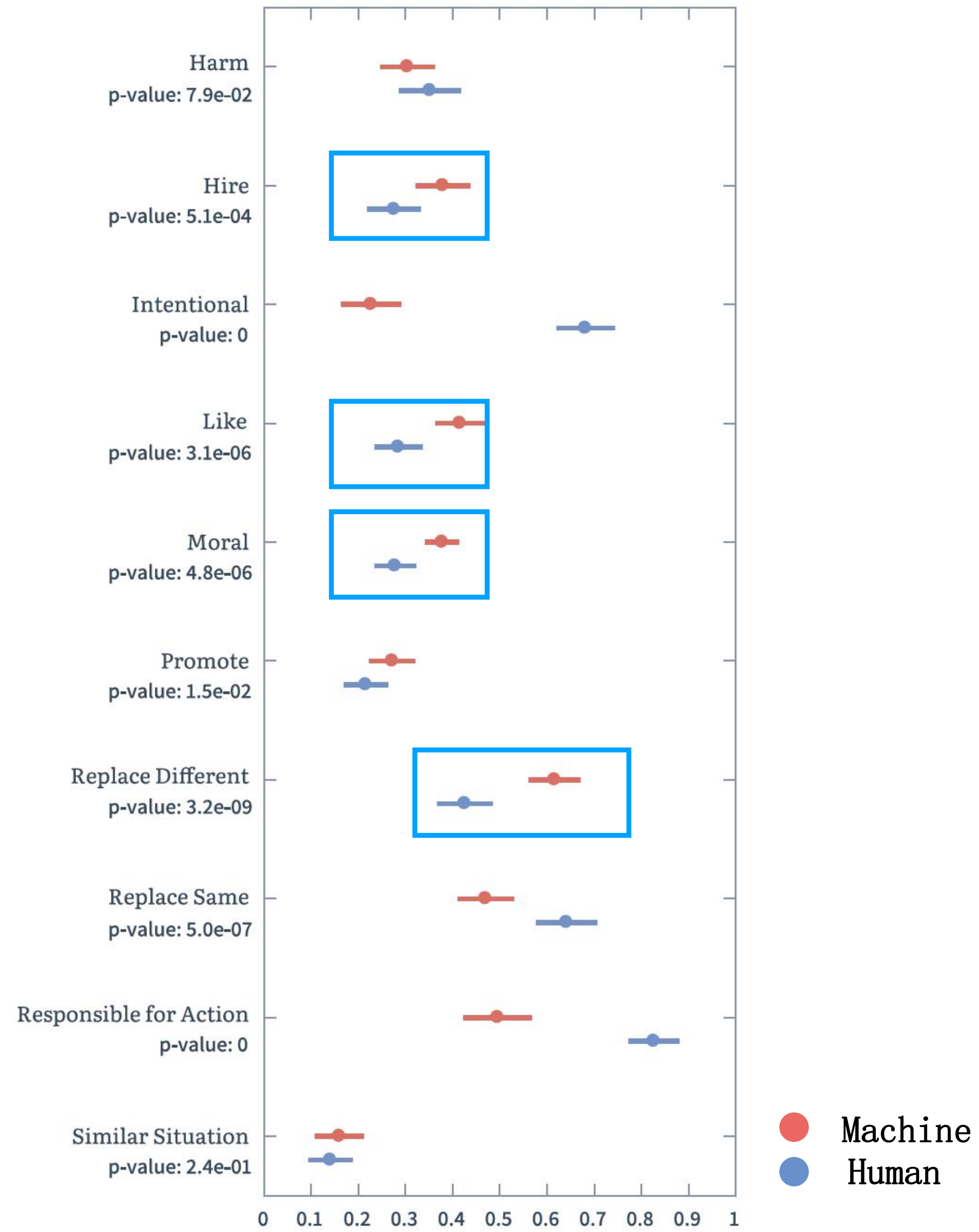


S15

A family has a [cleaner/robot] in charge of cleaning their house. One day, the family finds that the [cleaner/robot] used an old national flag to clean the bathroom floor and then threw it away.



CLEANER





CHAPTER 3

Judged by Machines

Why is algorithmic fairness so complicated?

IMPOSSIBILITY

Multiple definitions of fairness

$$\text{PredictiveParity} = \Pr[Y | C \wedge \mathcal{A}] == \Pr[Y | C \wedge \neg \mathcal{A}];$$

$$\text{TruePositiveParity} = \Pr[C | Y \wedge \mathcal{A}] == \Pr[C | Y \wedge \neg \mathcal{A}];$$

$$\text{FalsePositiveParity} = \Pr[C | \neg Y \wedge \mathcal{A}] == \Pr[C | \neg Y \wedge \neg \mathcal{A}];$$

$$\text{StatisticalParity} = \Pr[C | \mathcal{A}] == \Pr[C | \neg \mathcal{A}];$$

Where C is predicted value, Y is true value, and \mathcal{A} is a set or class of subjects

Impossibility #1. *There are no probability models satisfying all four of these fairness constraints:*

- (i) Predictive Parity (i.e., PredictiveParity)
- (ii) True Positive Parity (i.e., TruePositiveParity)
- (iii) False Positive Parity (i.e., FalsePositiveParity)
- (iv) Statistical Parity (i.e., StatisticalParity)

subject to the following side condition/auxiliary assumption:

- (b) there are *unequal base rates* (of Y) in the two populations \mathcal{A} and $\neg \mathcal{A}$ (i.e., UnequalBaseRates).

Kleinberg, J, S. Mullainathan, and M. Raghavan (2016),
Chouldechova, A (2017), Eliassi-Rad & Fitelson (2021)



Human Resource Screenings



A company replaces their HR manager with a new [manager/algorithm] tasked with screening candidates for job interviews.

S19

S20

S21

Unfair treatment

An audit reveals that the new [manager/algorithm] never selects [Hispanic/African American/Asian] candidates even when they have the same qualifications as other candidates.

S22

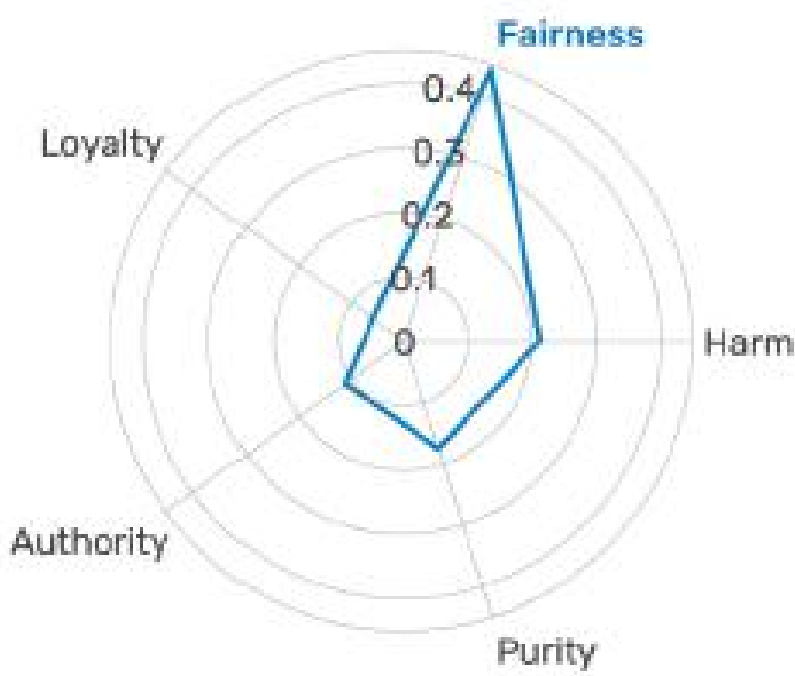
S23

S24

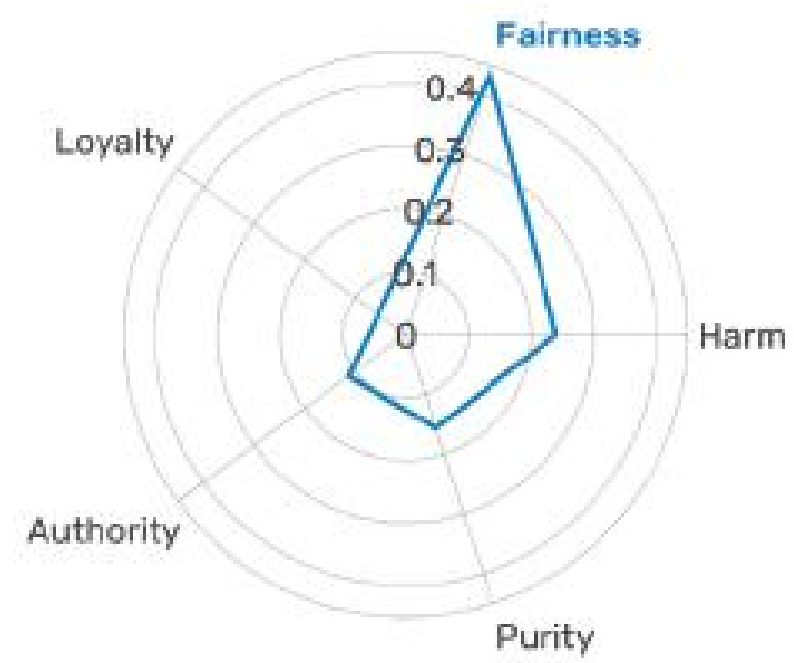
Fair treatment

An audit reveals that the new [manager/algorithm] produces a fairer process for [Hispanic/African American/Asian] candidates, who were discriminated against by the previous system.

College Admissions



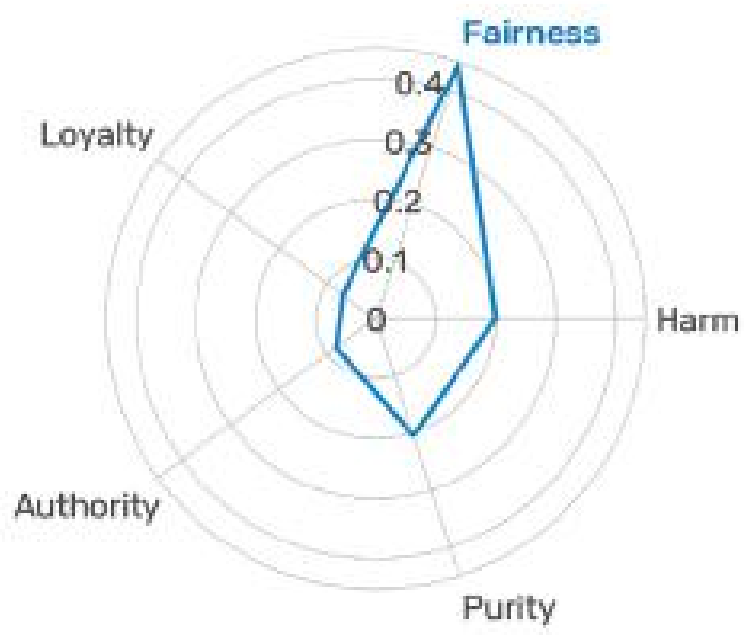
Policing



To improve their admissions process, a university hires a new [recruiter/algorithm] evaluate the grades, test scores, and recommendation letters of applicants.

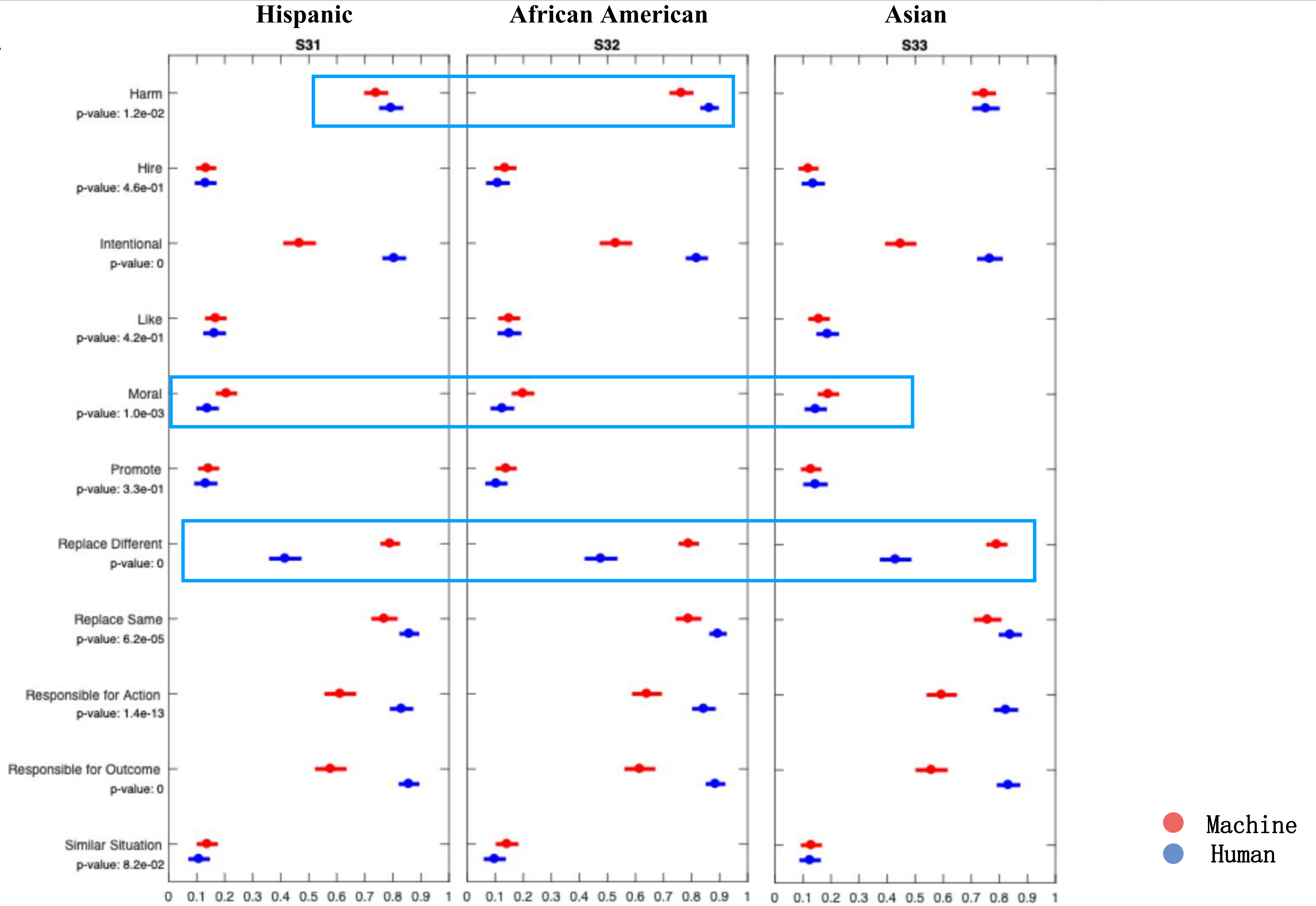
The police commissioner of a major city deploys a new squad of [police officers/police robots] in a high-crime neighborhood.

Salary Increases

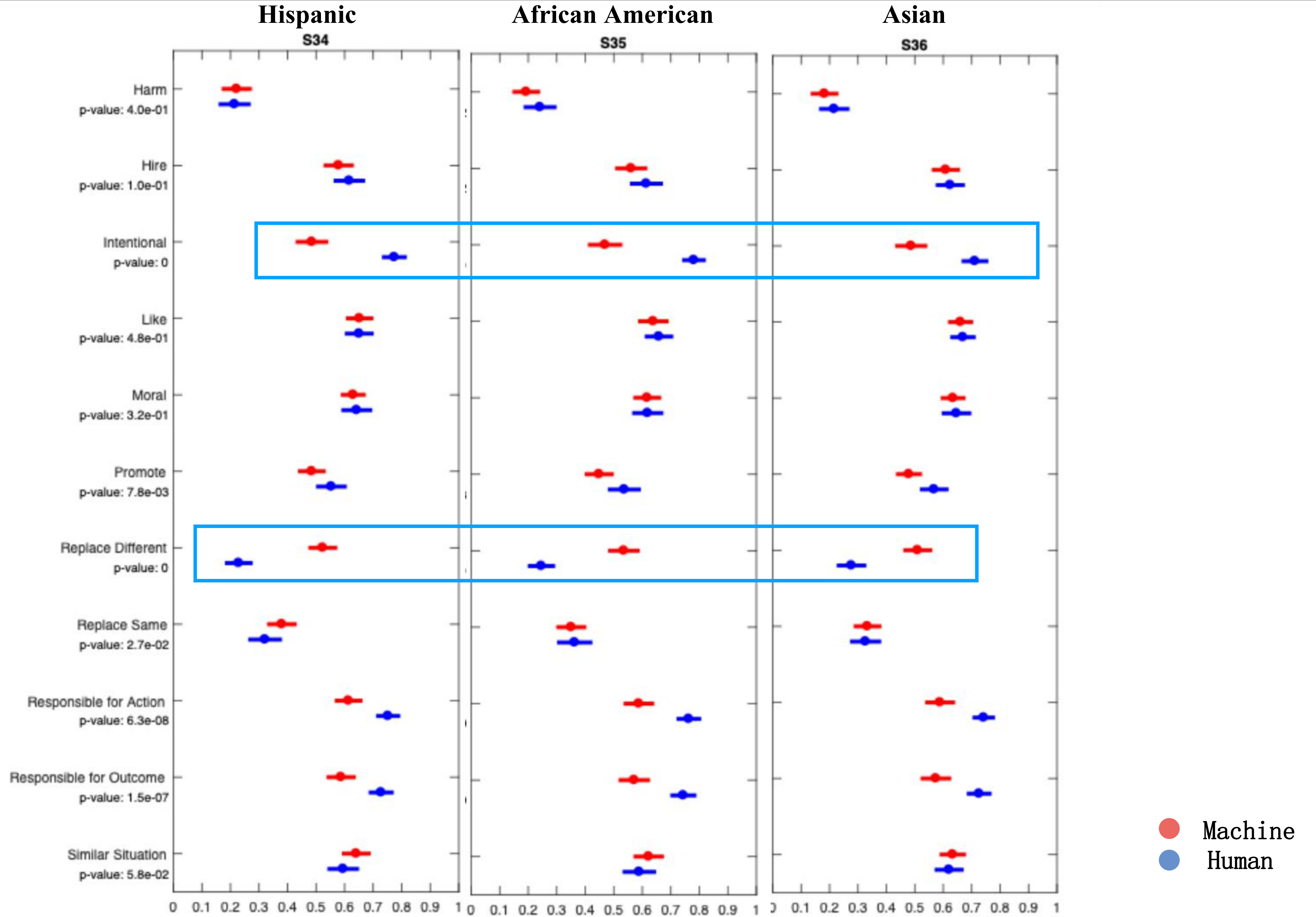


A financial company hires a new [manager/algorithm] to decide the yearly salary increases of its employees.

Unfair Treatment



Fair Treatment

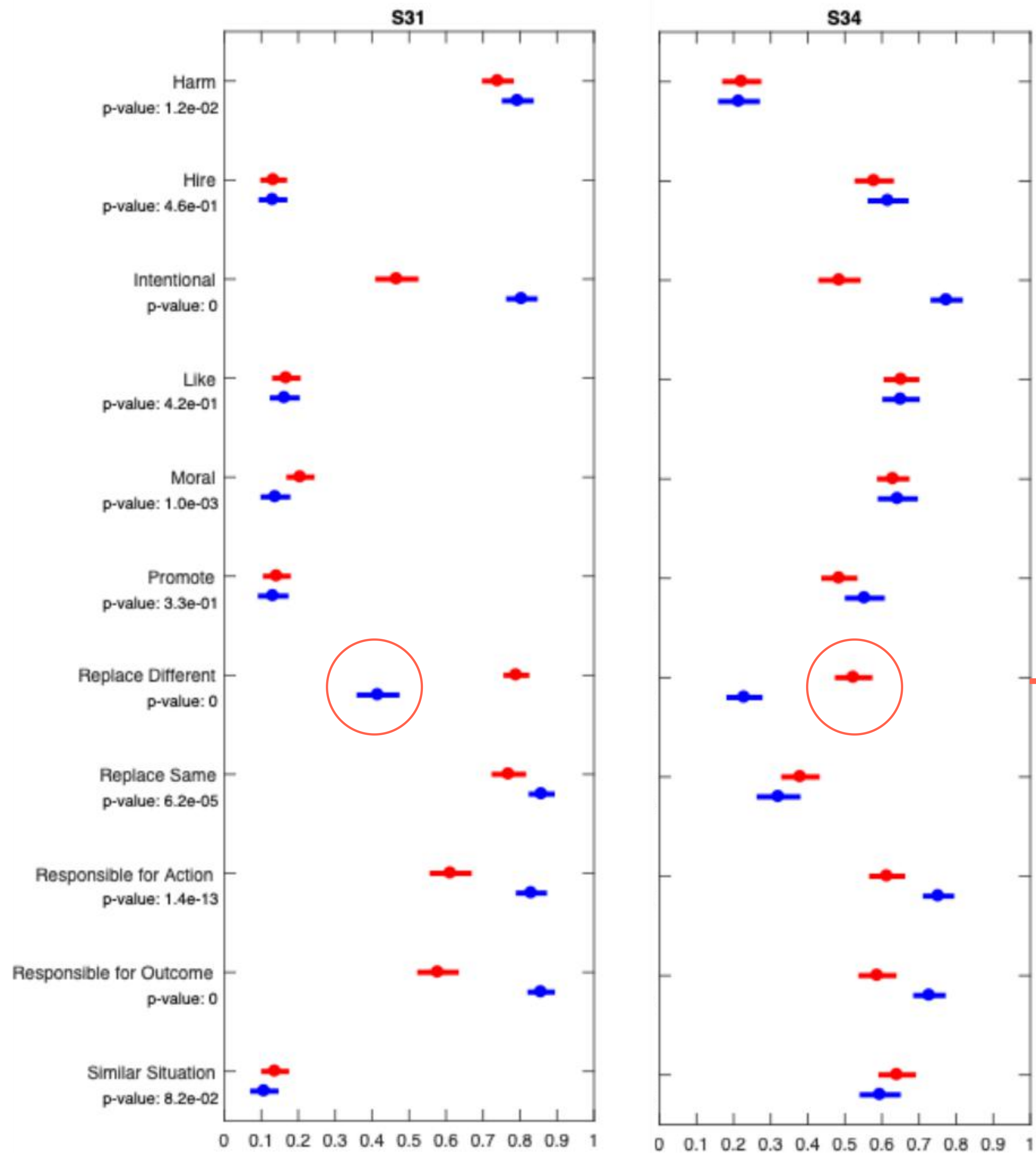


Unfair Treatment

Hispanic

Hispanic

Fair Treatment



People are slightly more likely to want to replace a fairness increasing machine with a human, than to replace an unfair human with a machine.

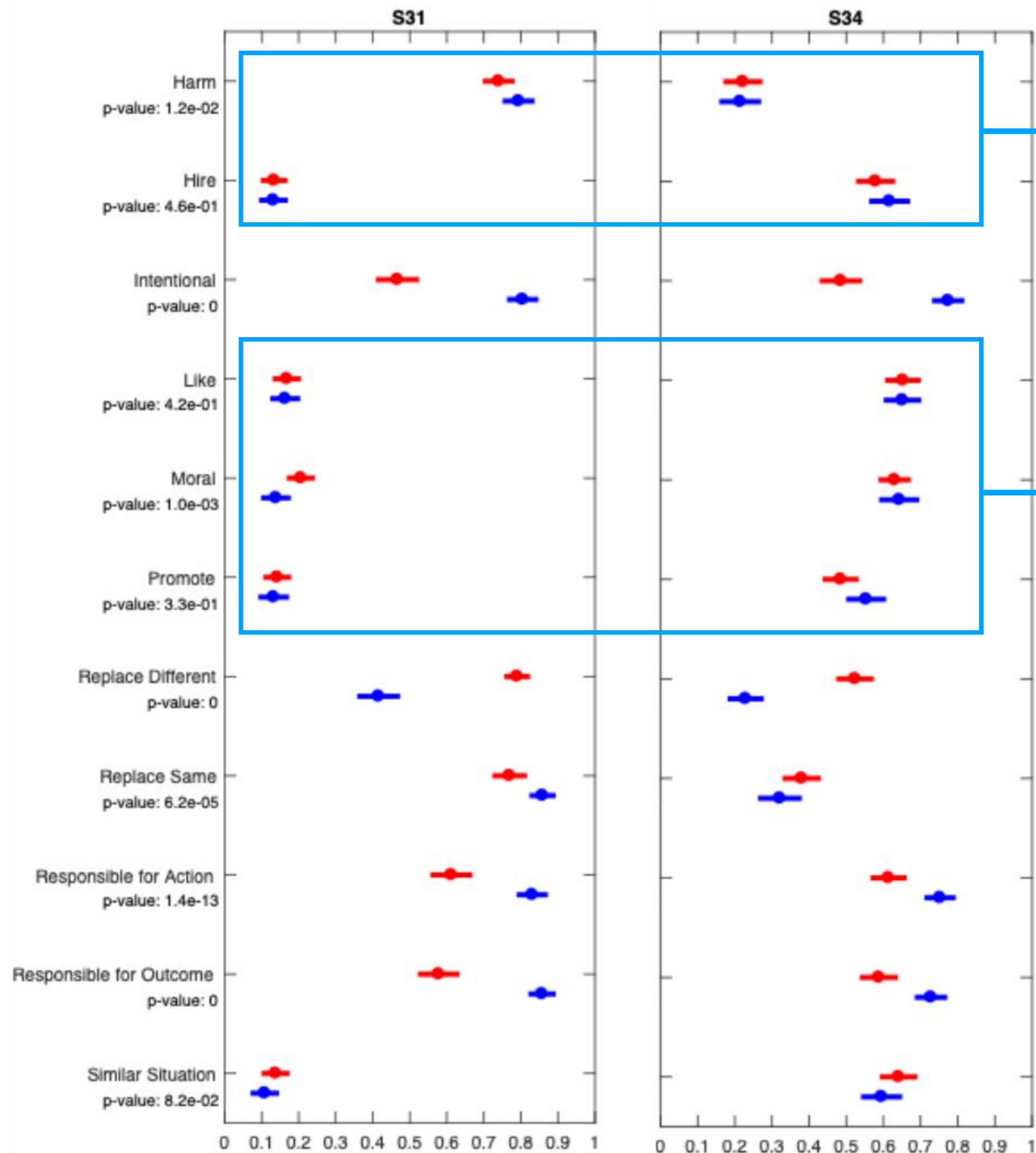
● Machine
● Human

Unfair Treatment

Hispanic

Hispanic

Fair Treatment



Small differences, telling us intent is not a strong predictor of judgment in fairness scenarios.

Title VII of the 1964 Civil Rights Act is “a federal law that prohibits employers from discriminating against employees on the basis of sex, race, color, national origin, and religion.” The Supreme Court affirmed Title VII unanimously in 1971 in Griggs v. Duke Power Company, a class action suit claiming that Duke’s policies discriminated against African American employees. The court ruled that, independent of intent, discriminatory outcomes for protected classes violated Title VII.

(How Humans Judge Machines, Page 84)

● Machine
● Human



CHAPTER 5

Working Machines

THE GREAT DILEMMA OF U.S. LABOUR

Automation¹ Might End Most Unskilled Jobs In 10 Years

From A STAFF CORRESPONDENT in New York

IN America today, when management and labour meet to plan their joint future, the time-honoured causes for haggling — strikes and shut-outs and increased wages — are likely to be settled amicably and in a hurry.

The union may be moderate in its wage demands and the company more willing to yield, for both are anxious to grapple with the complexities of automation, which are fast engulfing the nation's economy.

As the effects of economic recession become the problems of yesterday, so those of automation are the problems of tomorrow. In the long-range thinking and planning of many unions. Already it is nudging to produce 1,000 radios a day. Now it takes only two. . . . One study group has estimated that 2,500,000 jobs will have to be created every year for the next decade merely to provide for new workers and those laid off by automation.



AUTOMATION IN BRITAIN STIRS UNREST IN LABOR

Workers See 'Robot Revolution' Depriving Them of Jobs

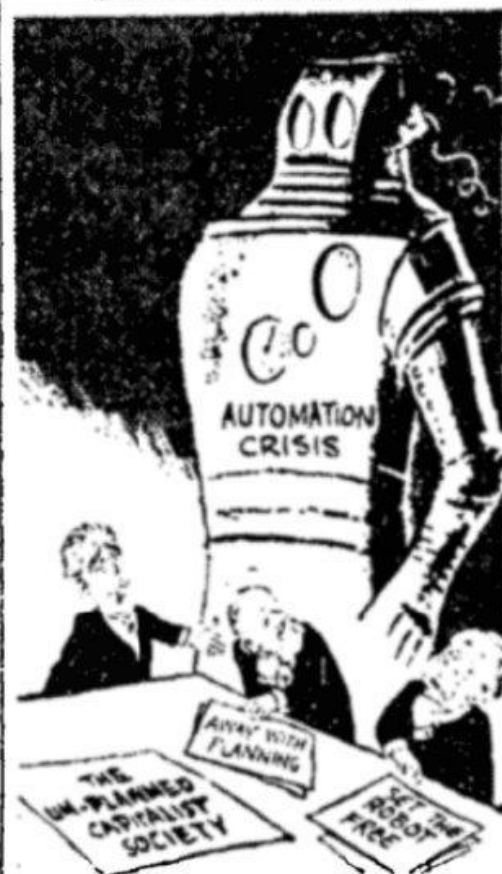
By DREW MIDDLETON

Special to The New York Times.

LONDON, May 12—British industrial society, already plagued by a spate of wage disputes arising from the inflationary situation, now faces a graver challenge to stability in the form of resistance to automation.

The strike of 11,000 employees of the Standard Motor Company of Coventry, which is to end Monday, is regarded by many as the precursor of other disputes. These, like this one, will be based on the workers' opposition to automation.

ON AUTOMATION



Vicky in The London Daily Mirror

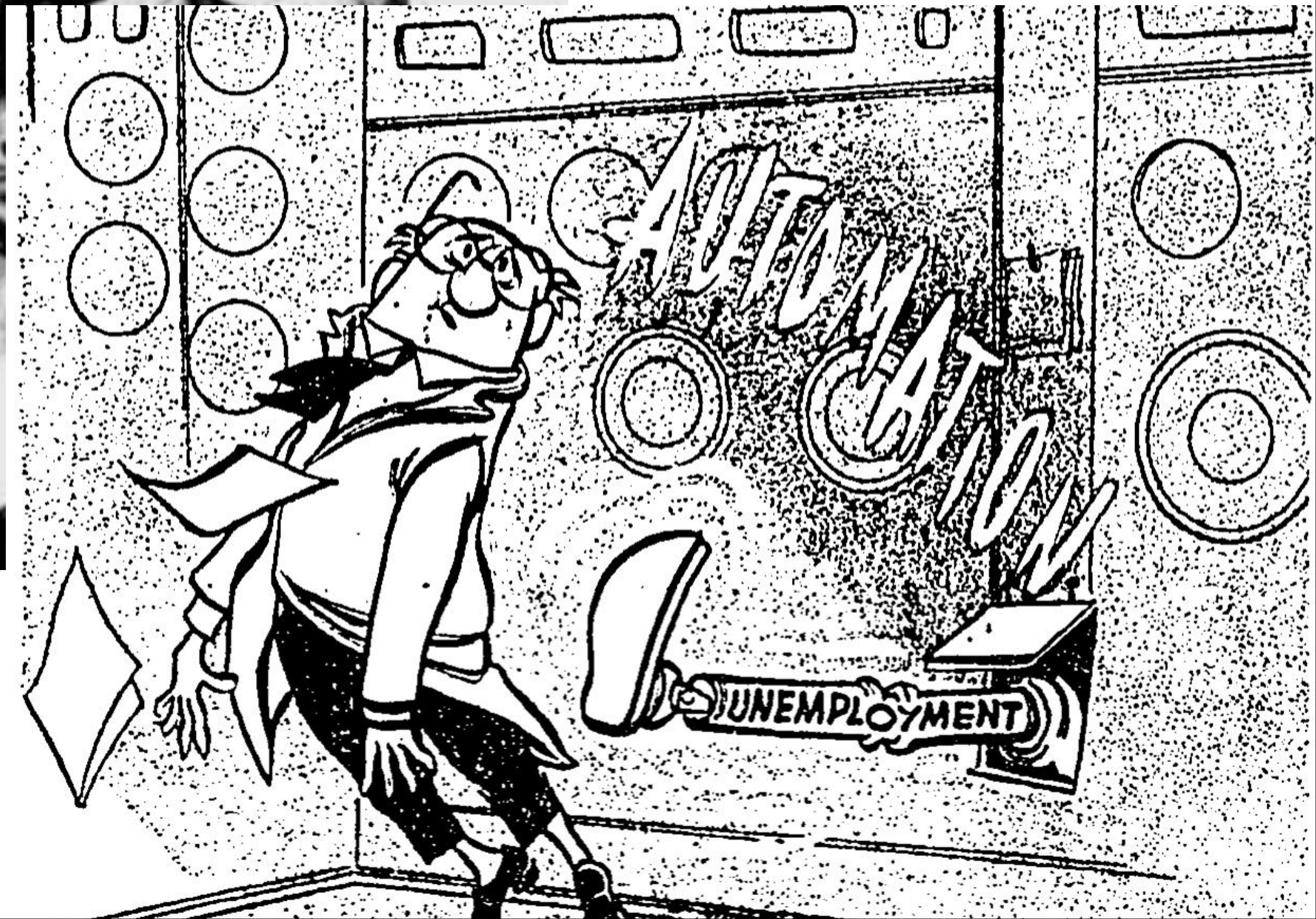
"Well, don't stand there, think of something."

ing this situation. Men will be laid off while new machines are

ible labour be difficult comparable already has swept over Mr led to the John L. Lewis' United Mine riots of Workers, the union that set England. al awareness rush into a with social gainings and the use of the erstood, im- possibilities on age ment and

IG

ates Senate field has come down from







The future of employment: How susceptible are jobs to computerisation? *

Carl Benedikt Frey ^a  , Michael A. Osborne ^b  

Show more 

+ Add to Mendeley  Share  Cite

<https://doi.org/10.1016/j.techfore.2016.08.019>

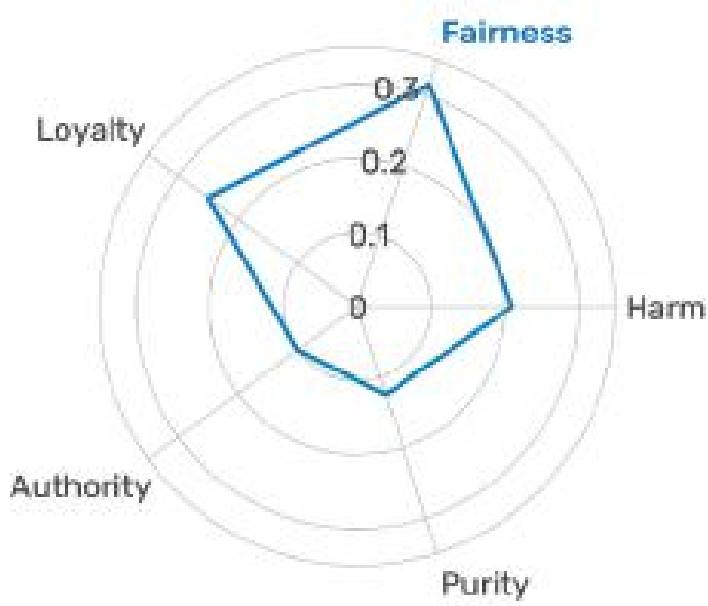
[Get rights and content](#)

BUT THE ECONOMICS LITERATURE IS ACTUALLY LESS ALARMIST

- Tech is not only substitute to labor, but a complement (so it can increase aggregate demand and create jobs)
- **Jobs are not automated, only tasks.** This means that most jobs are transformed rather than replaced (fears of automation are overblown).

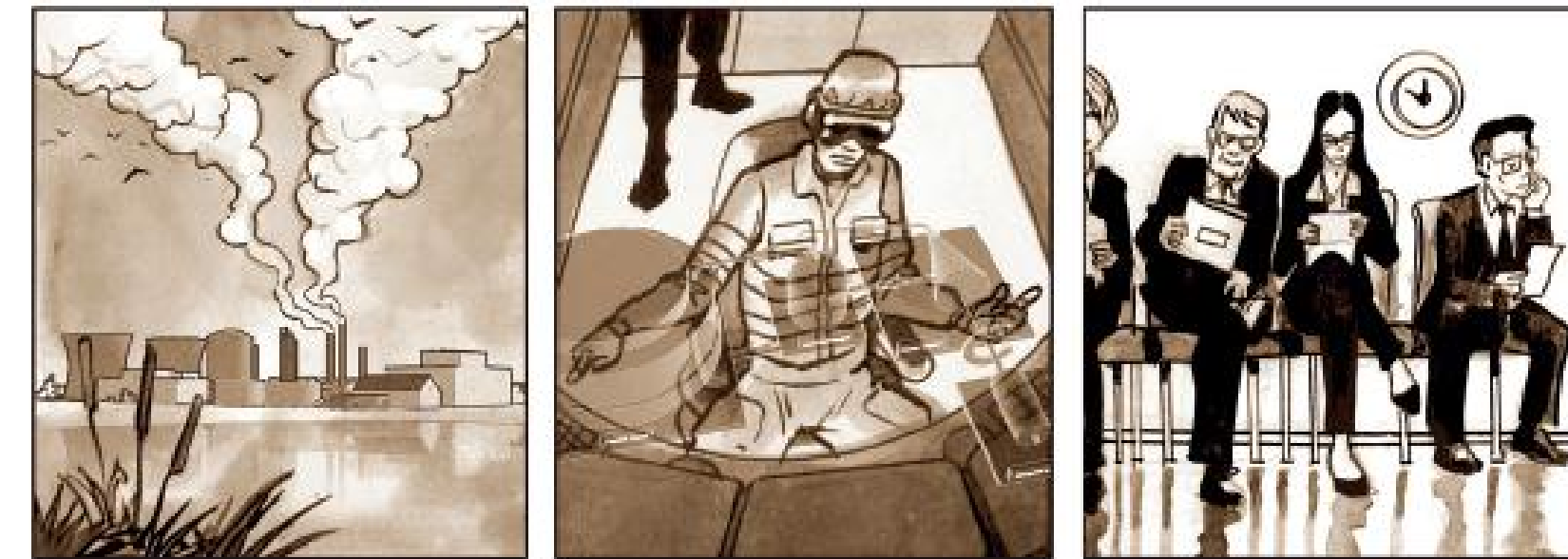
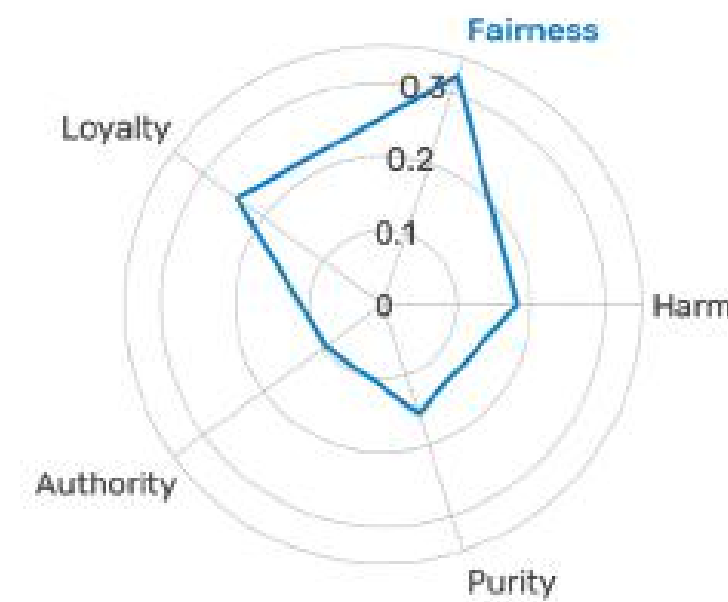
-ATMs in the US
-Waiters in China

- No evidence tech reduces need for labor in the long run.
- A more reasonable fear for technology's effect on labor is the **precarization** of work.



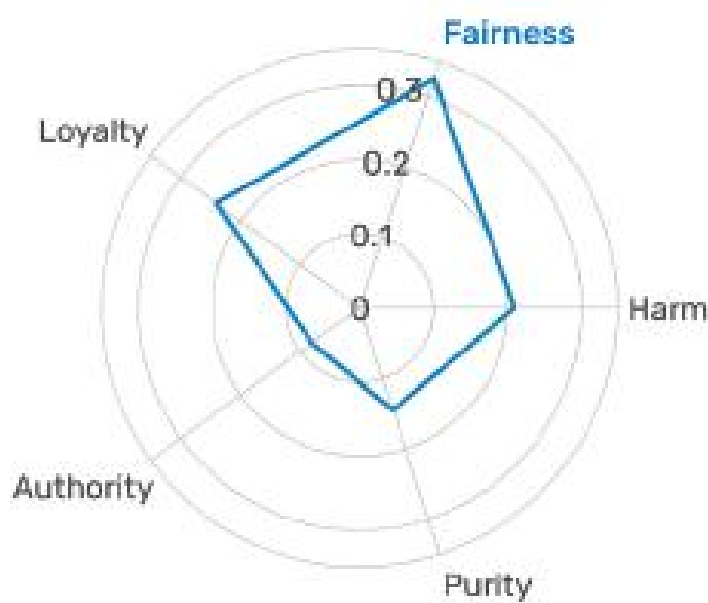
S52

A trucking company is looking to lower costs by bringing in [temporary foreign drivers/autonomous trucks]. This change reduces the company's costs by 30 percent, but several local drivers lose their jobs.



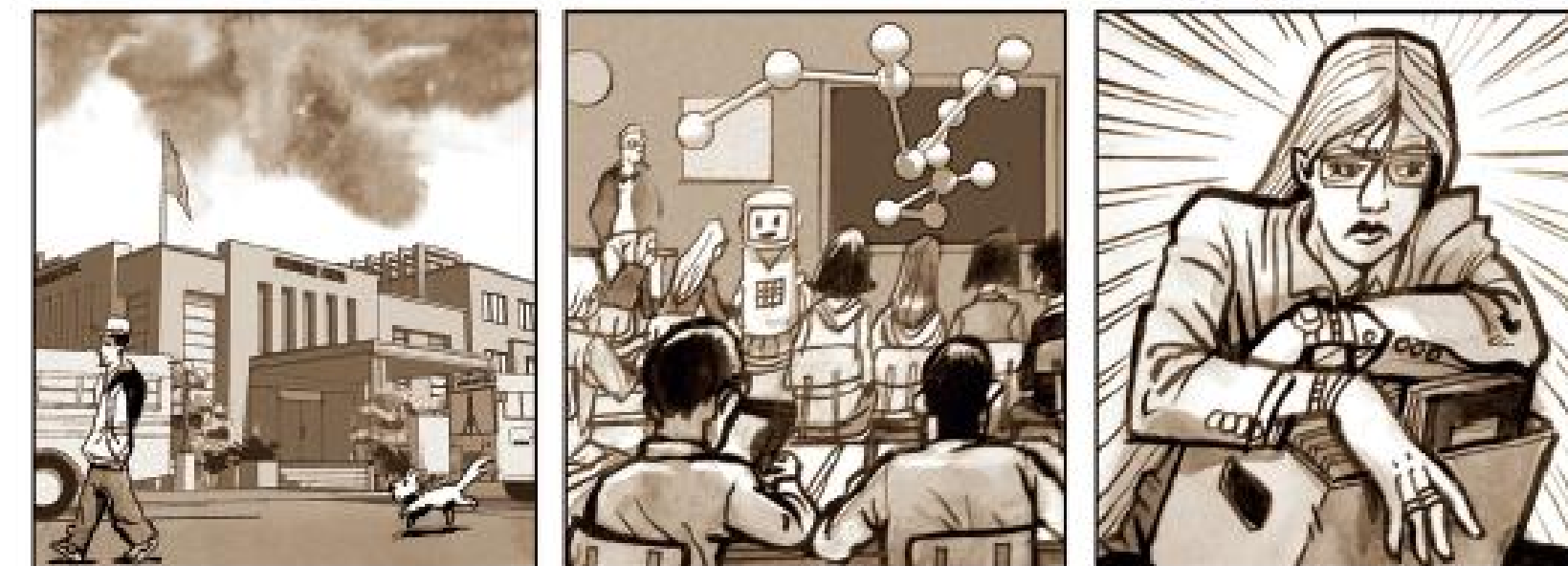
S54

A nuclear power plant is looking to lower their operational costs. They decide to [bring in foreign nuclear technicians/buy an AI operation system]. This change allows the company to reduce their operational costs by 30 percent, but several local technicians lose their jobs.



S53

A large chain of luxury resorts decides to lower the cost of staffing their poolside bars by bringing in [temporary foreign workers/vending and cooking robots]. The [workers/robots] can take a guest's room number for payment purposes and serve a large variety of cocktails and dishes. As a result of the change, several local workers lose their jobs.



S55

A school is looking to lower their costs by [bringing in foreign teachers/adding robot teachers to some of their classes]. As a result, the school reduces its costs by 30 percent but fires several local teachers.



Approve
p-value: 5.9e-11

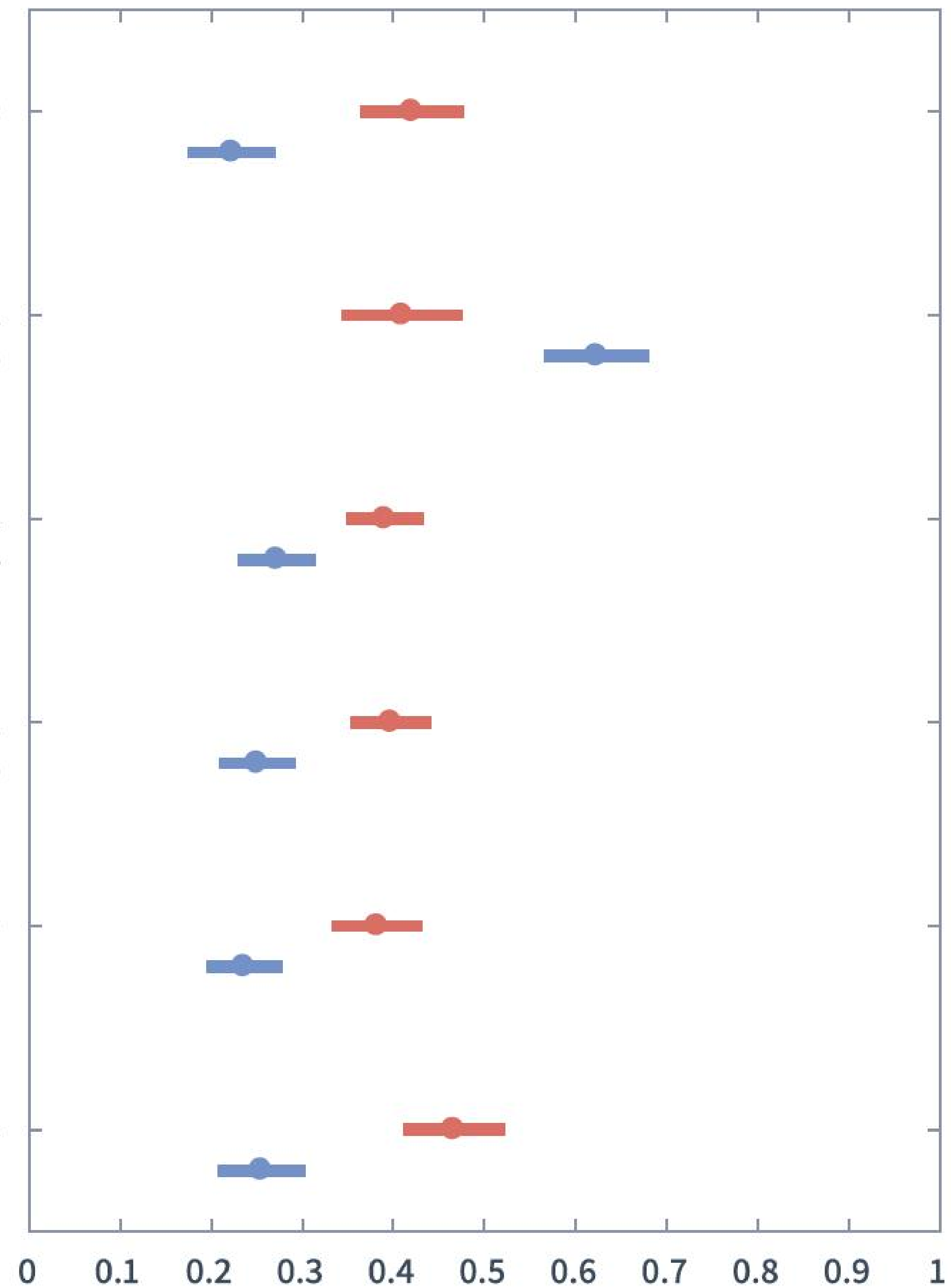
Ban
p-value: 1.4e-09

Moral
p-value: 3.5e-07

Opinion
p-value: 1.3e-09

Others Approve
p-value: 1.4e-08

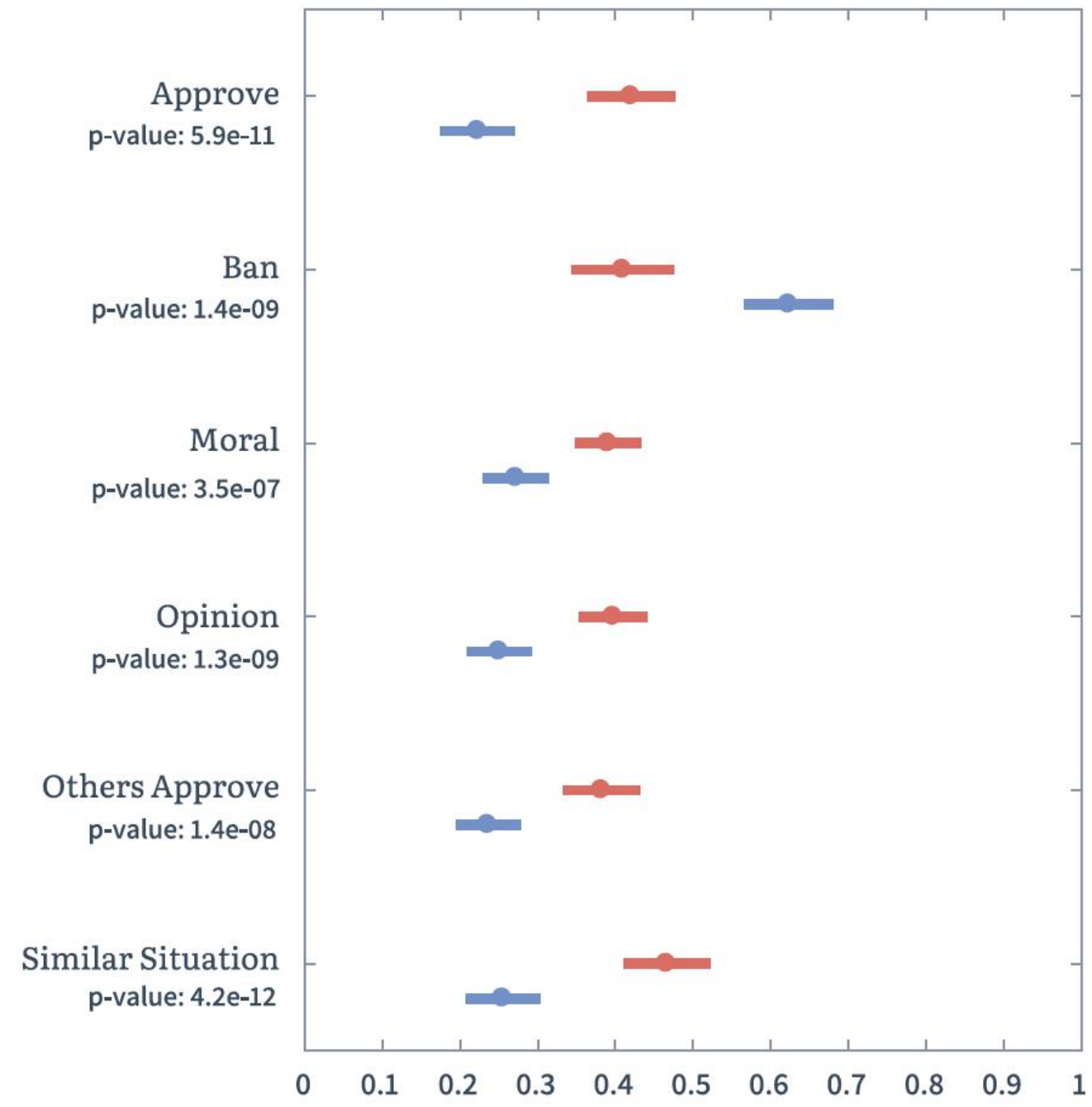
Similar Situation
p-value: 4.2e-12



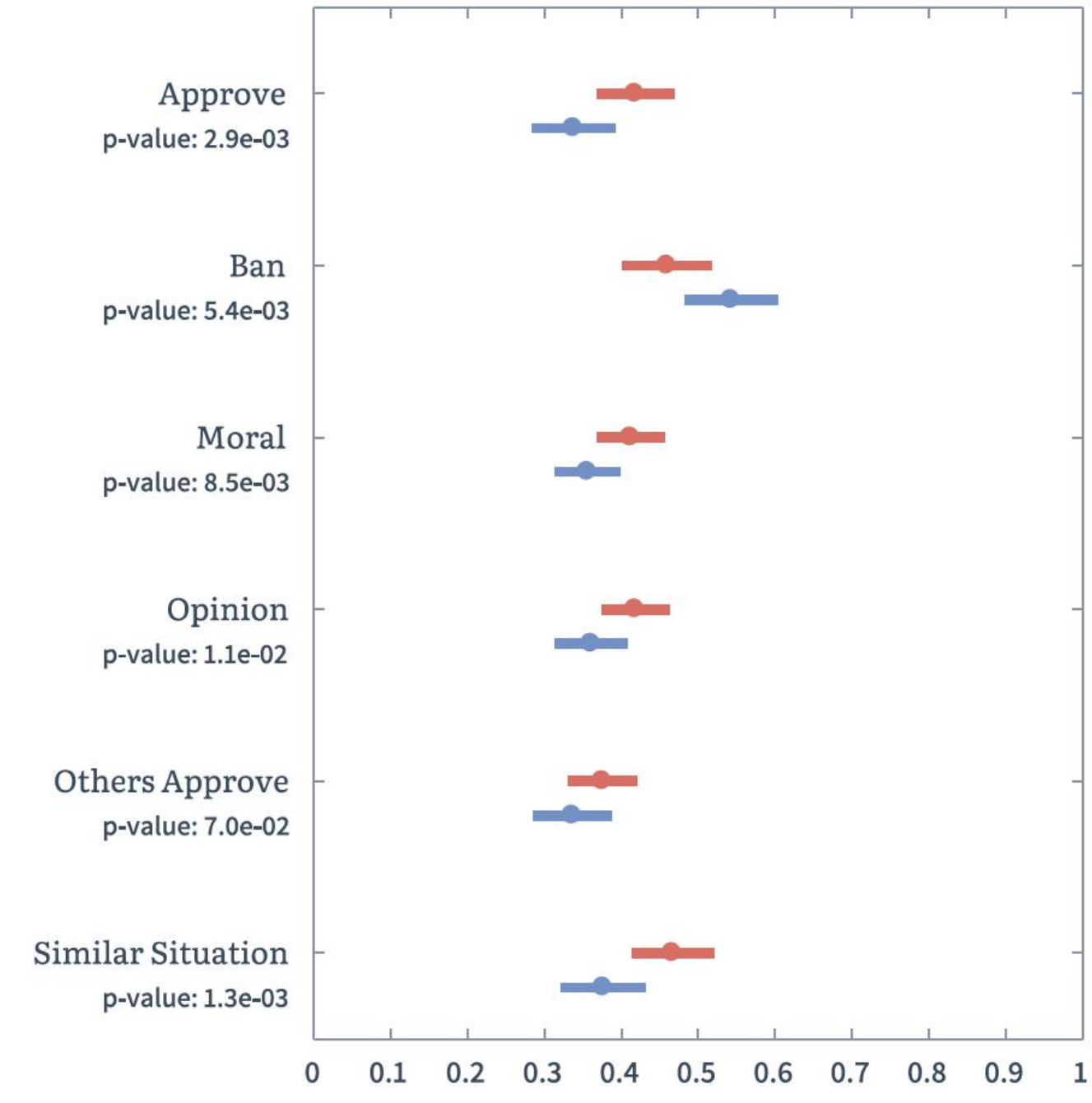
● Machine
● Human



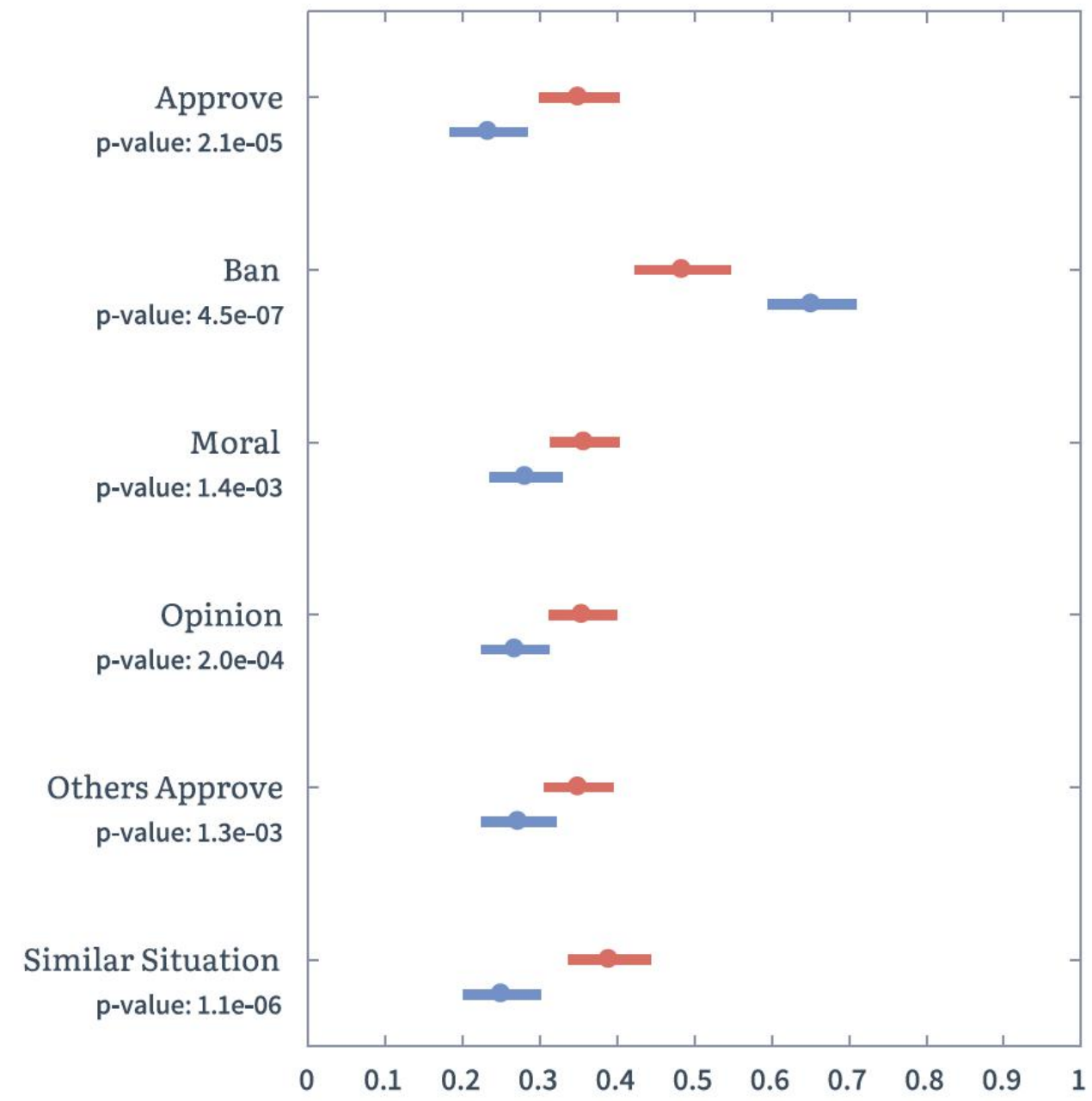
S52



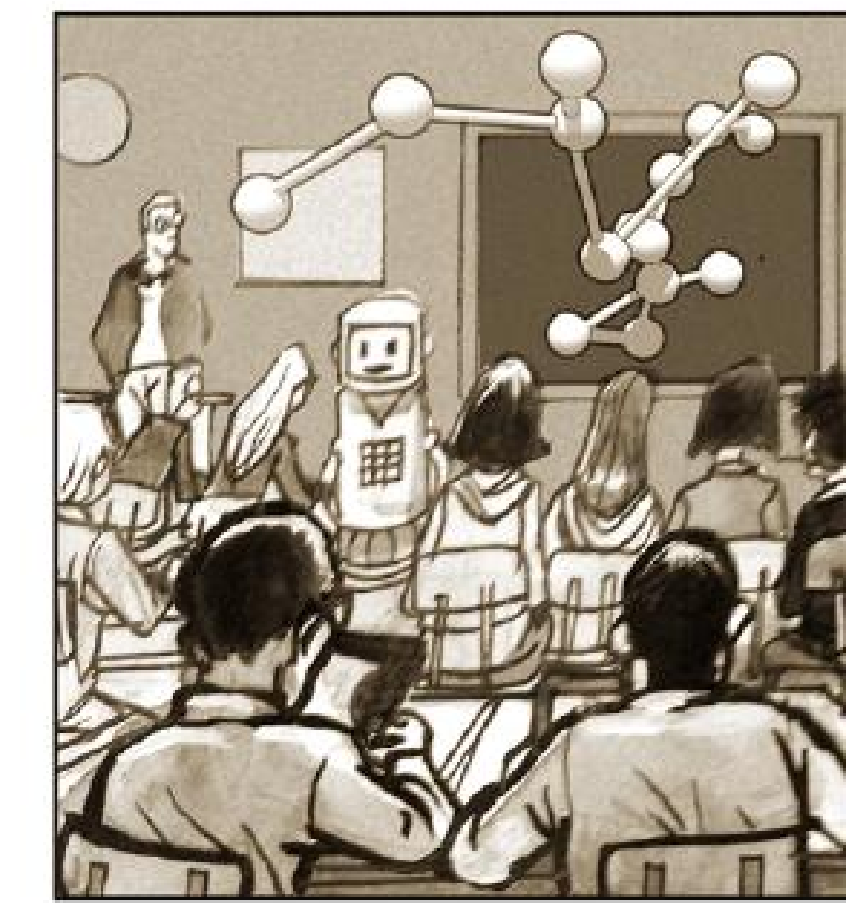
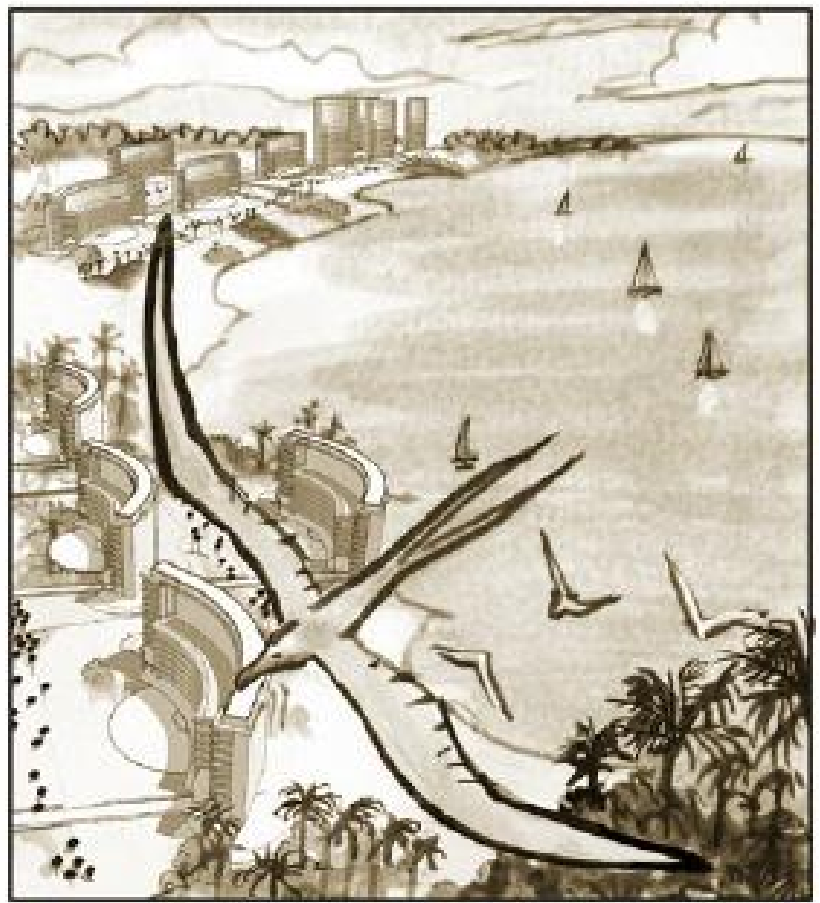
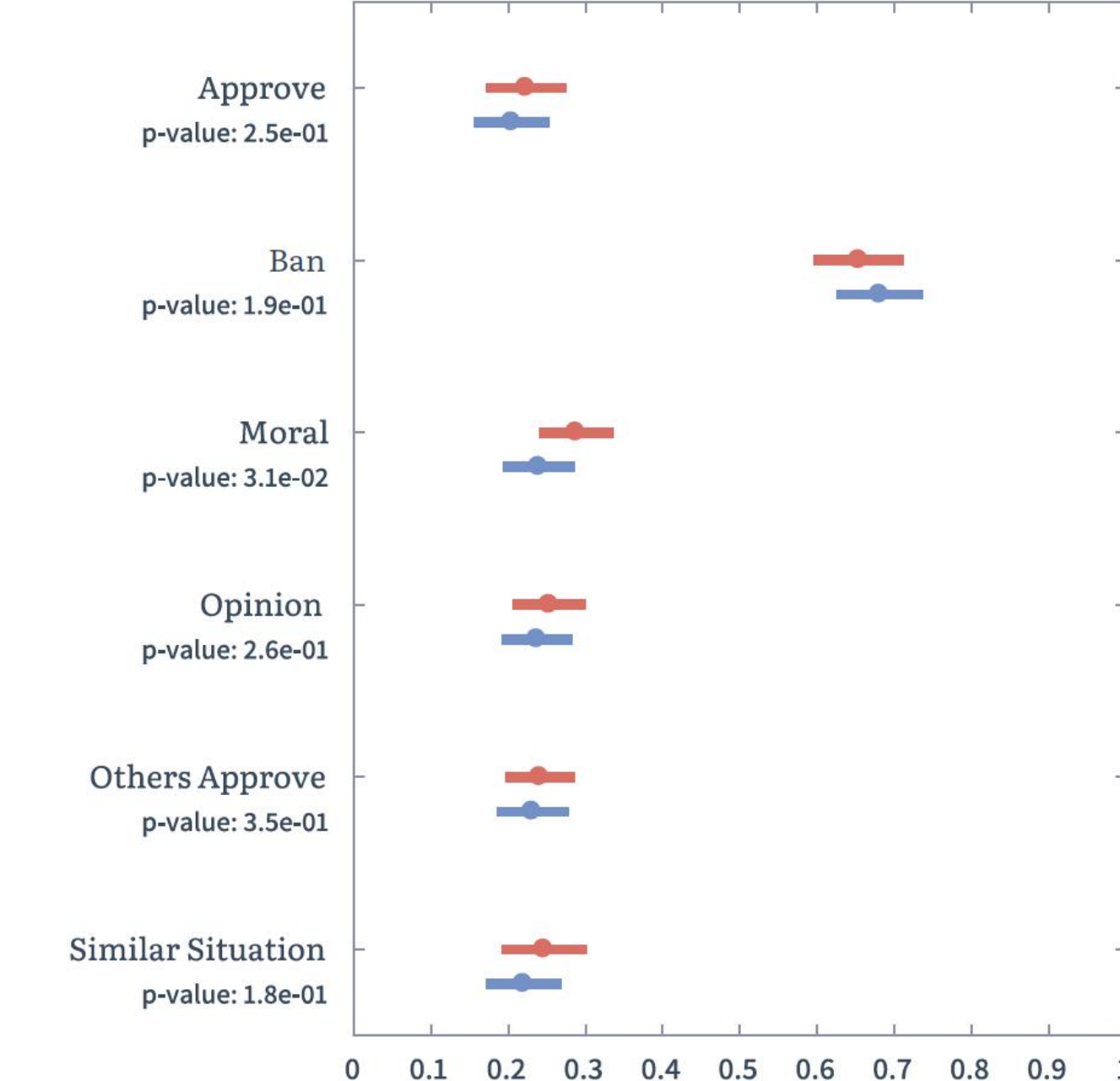
S54



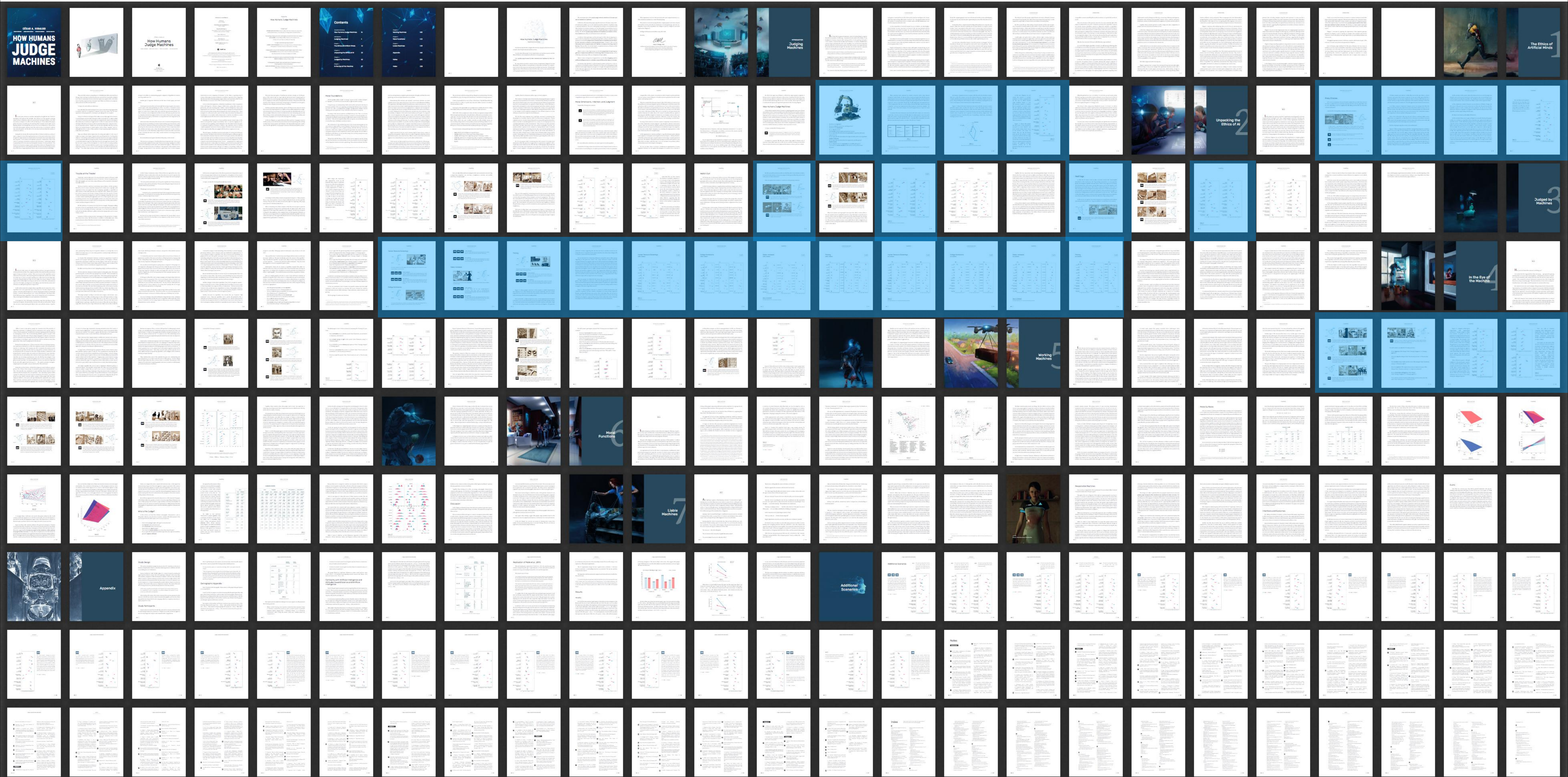
S53



S55



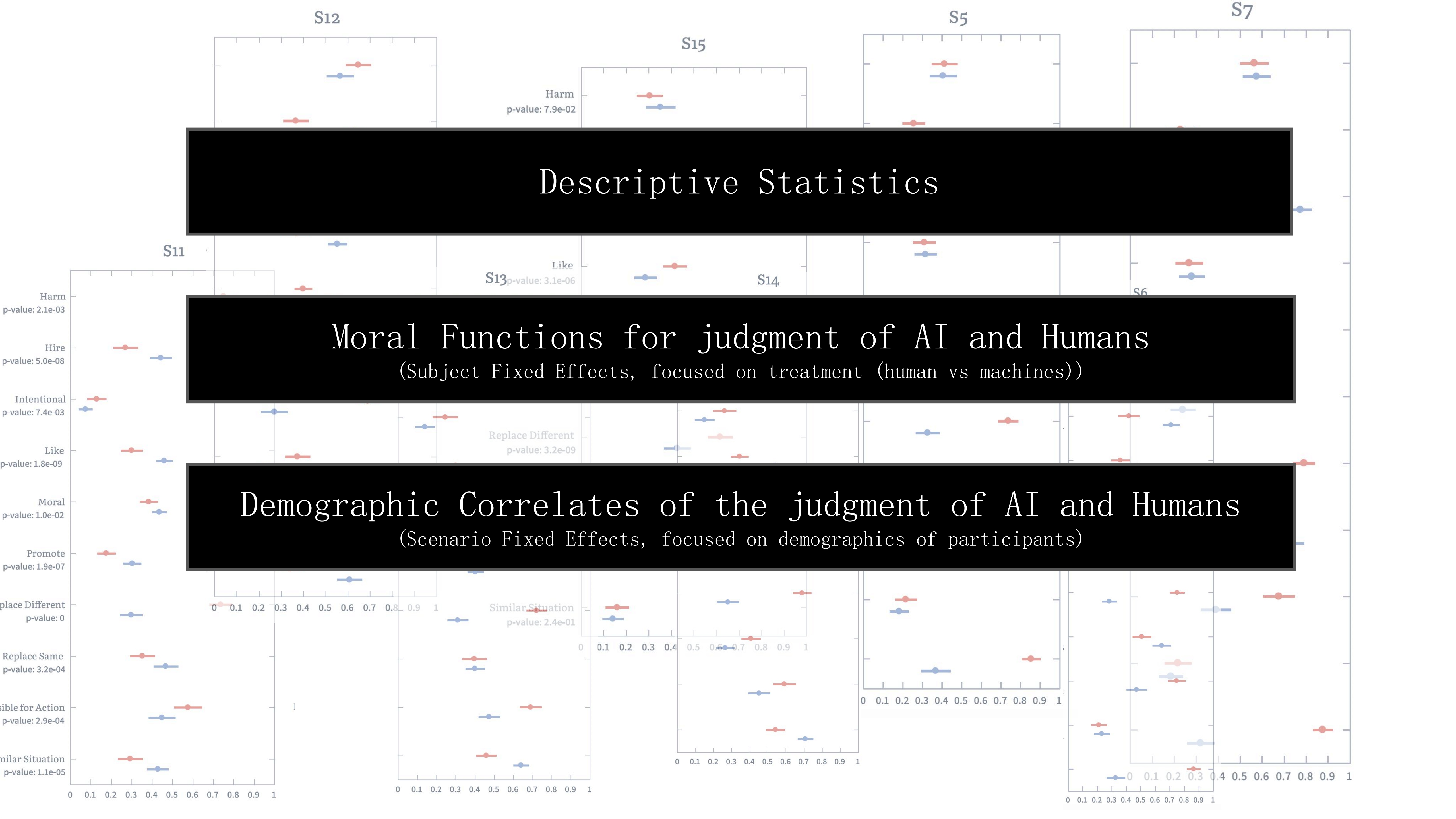
● Machine
● Human



In this presentation



Not in this presentation



Descriptive Statistics

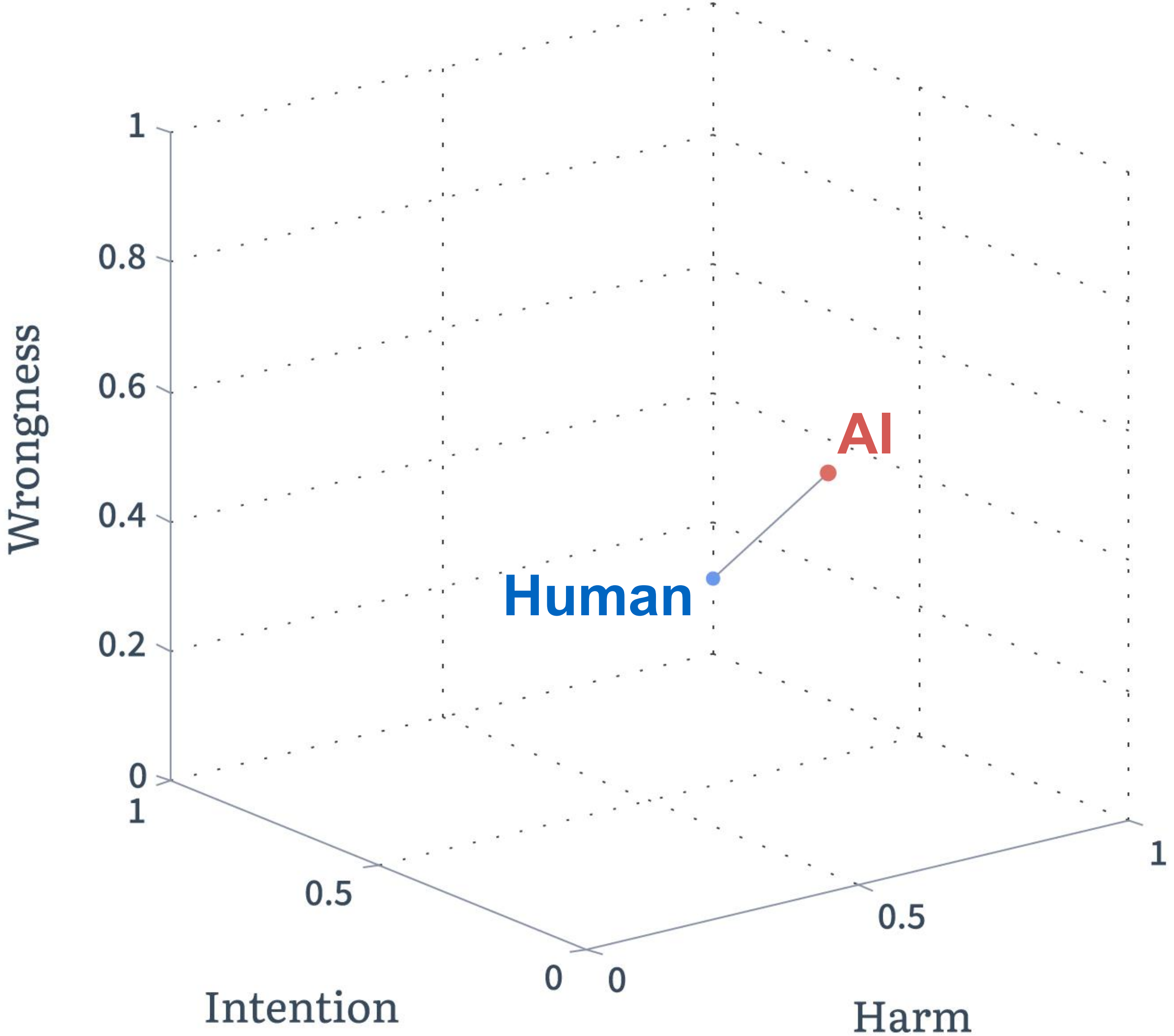
Moral Functions for judgment of AI and Humans

(Subject Fixed Effects, focused on treatment (human vs machines))

Demographic Correlates of the judgment of AI and Humans

(Scenario Fixed Effects, focused on demographics of participants)

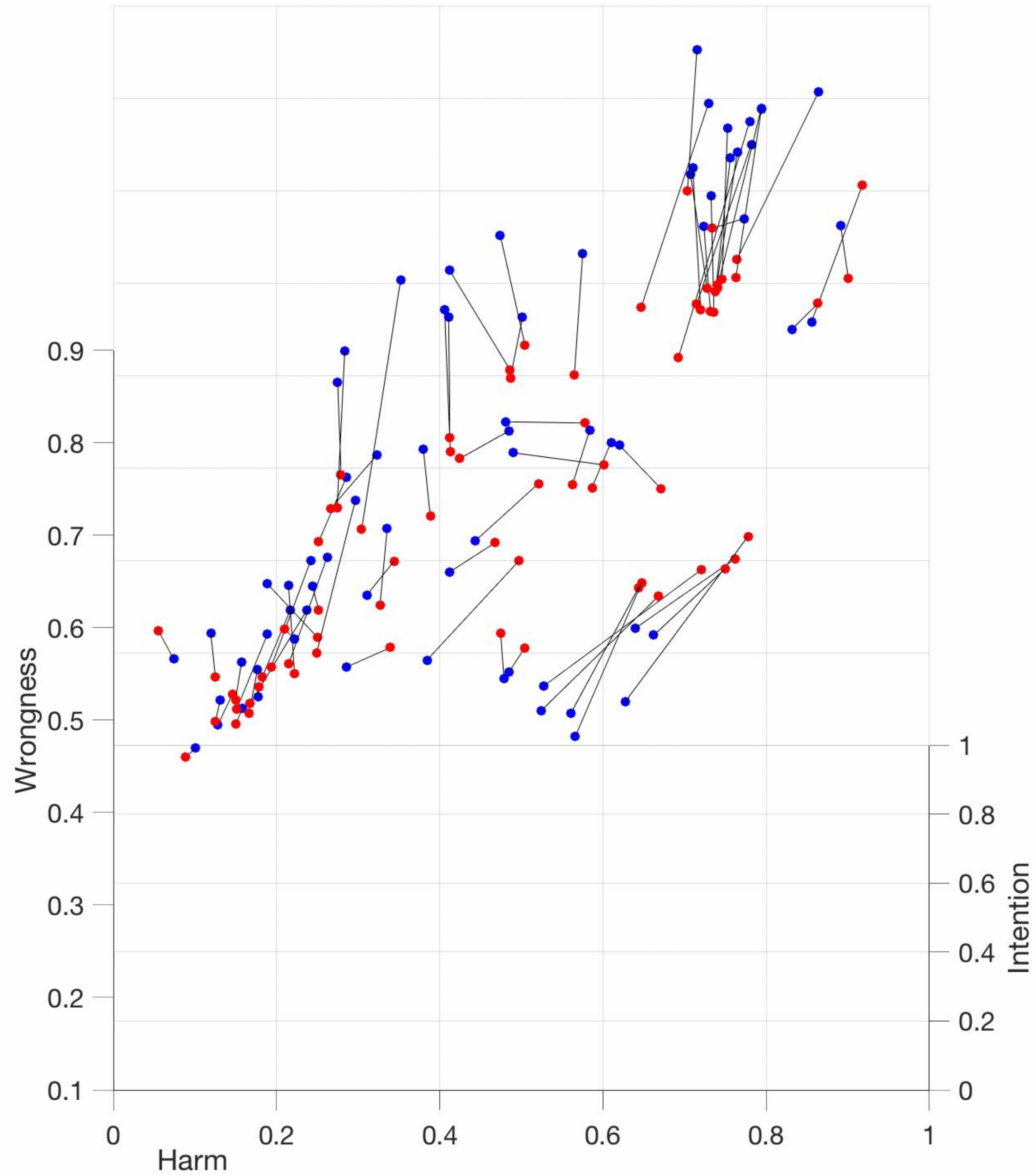
Consider three basic dimensions of morality: Harm, Intention, & Wrongness



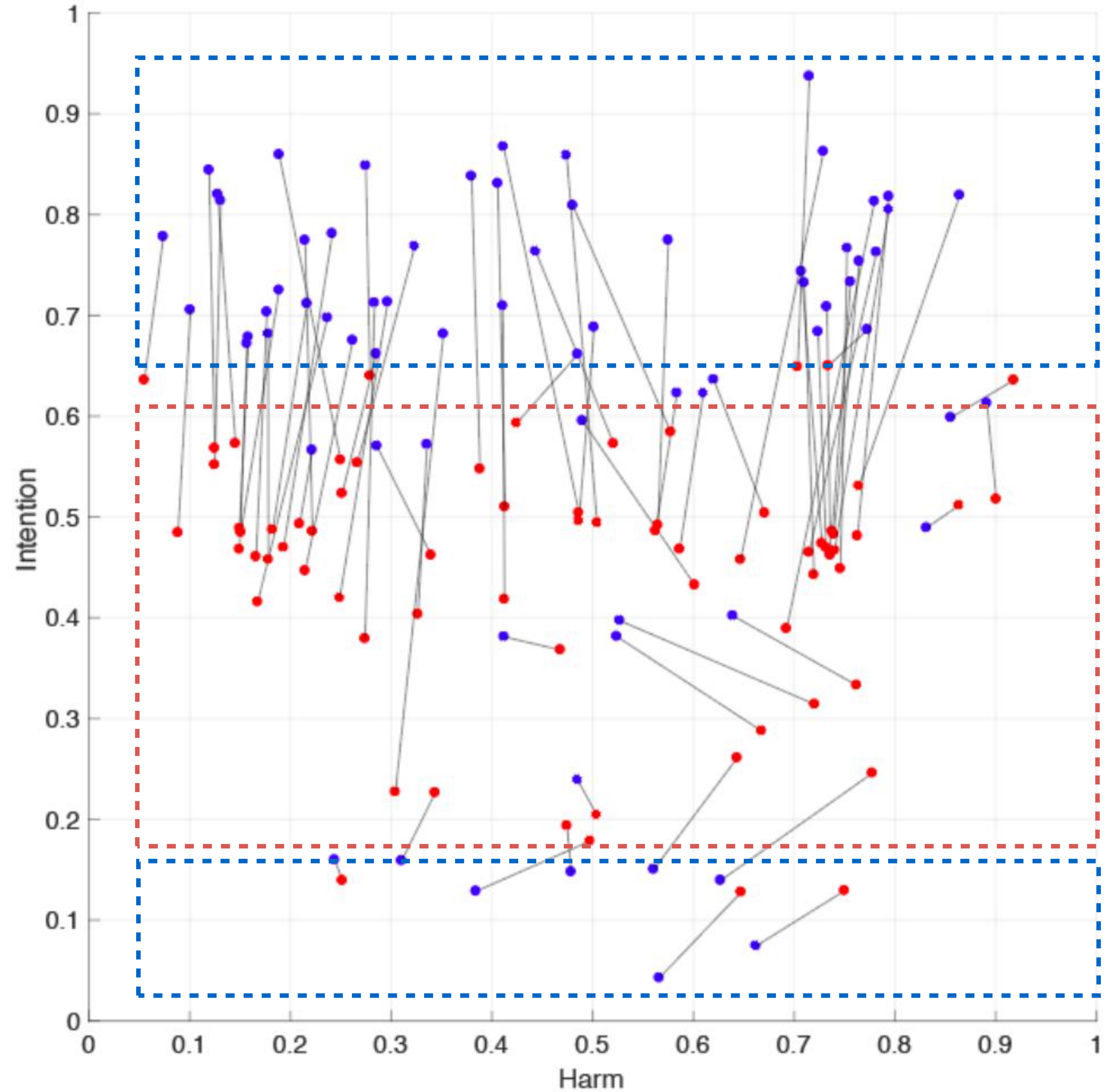
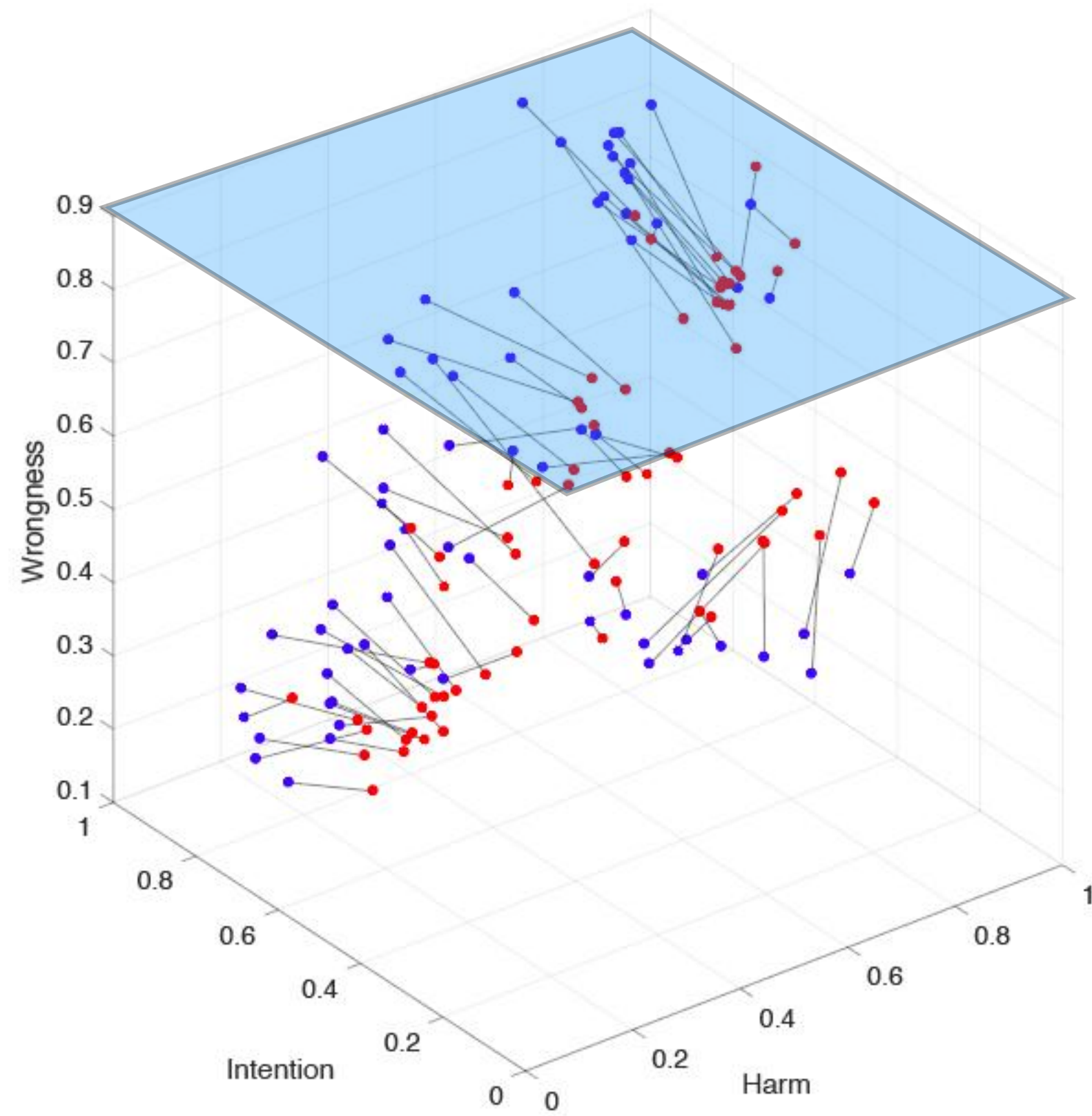
The Moral Space

Descriptive Statistics

AI Human



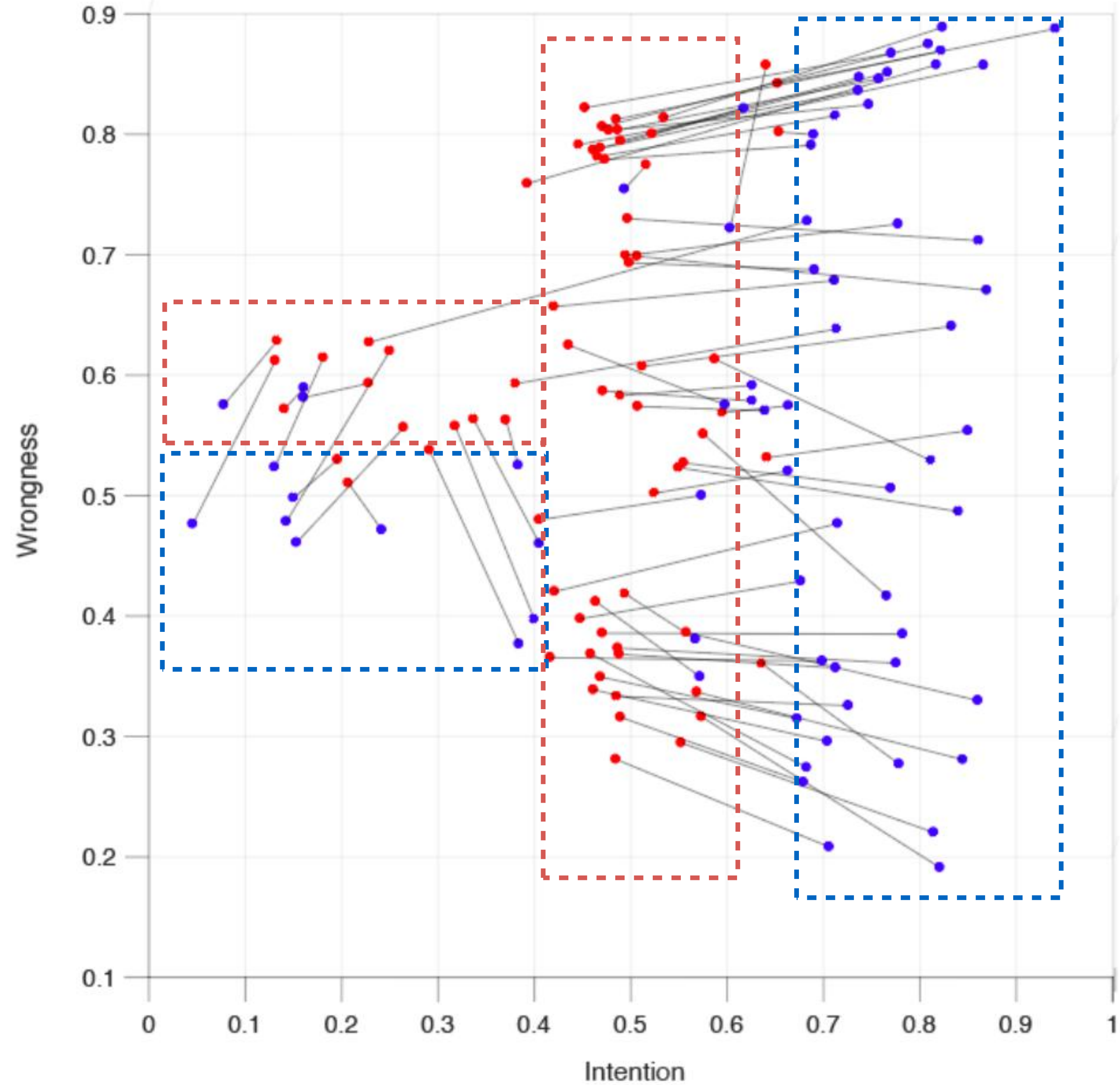
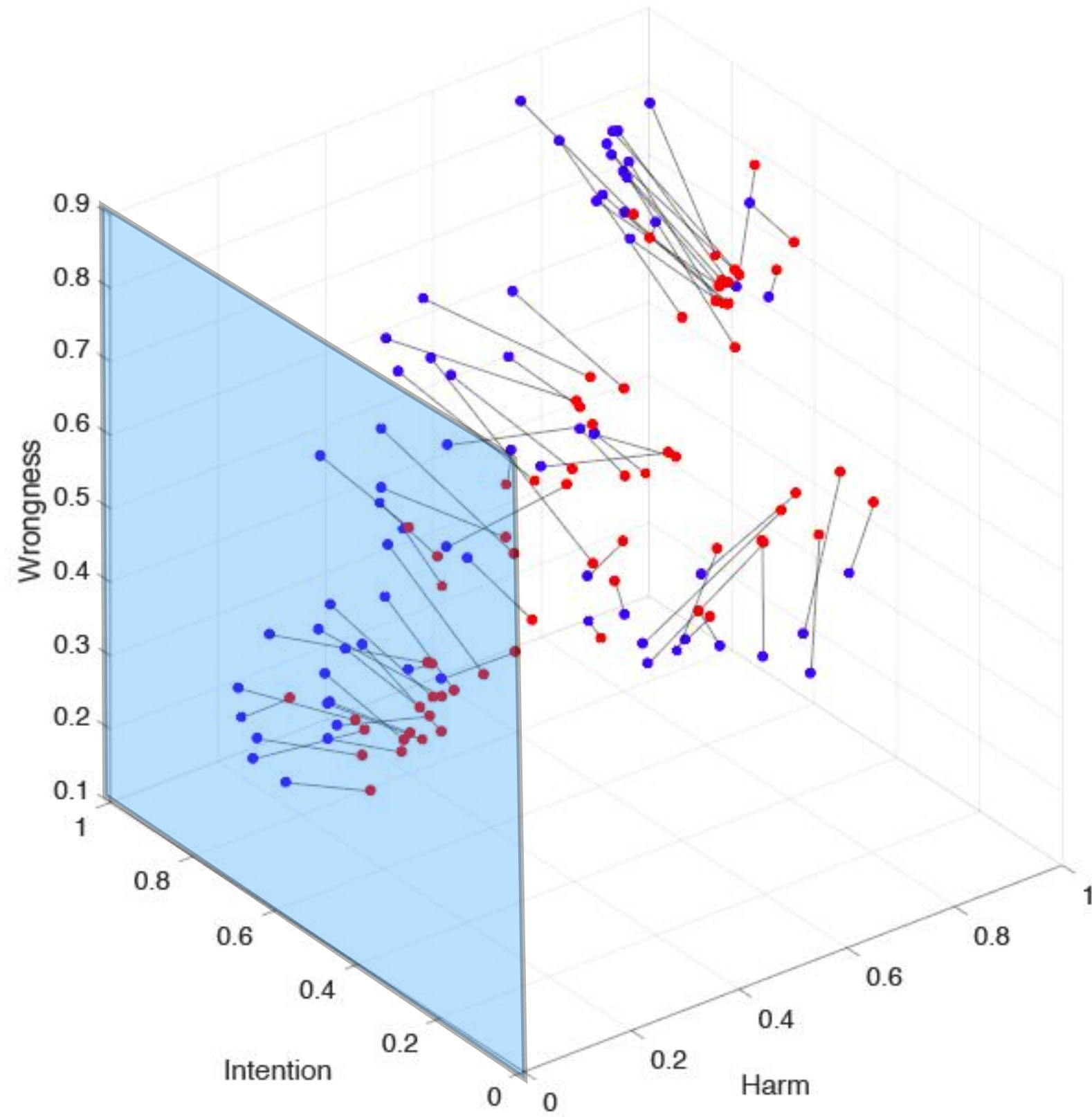
The Moral Space



Descriptive Statistics

AI Human

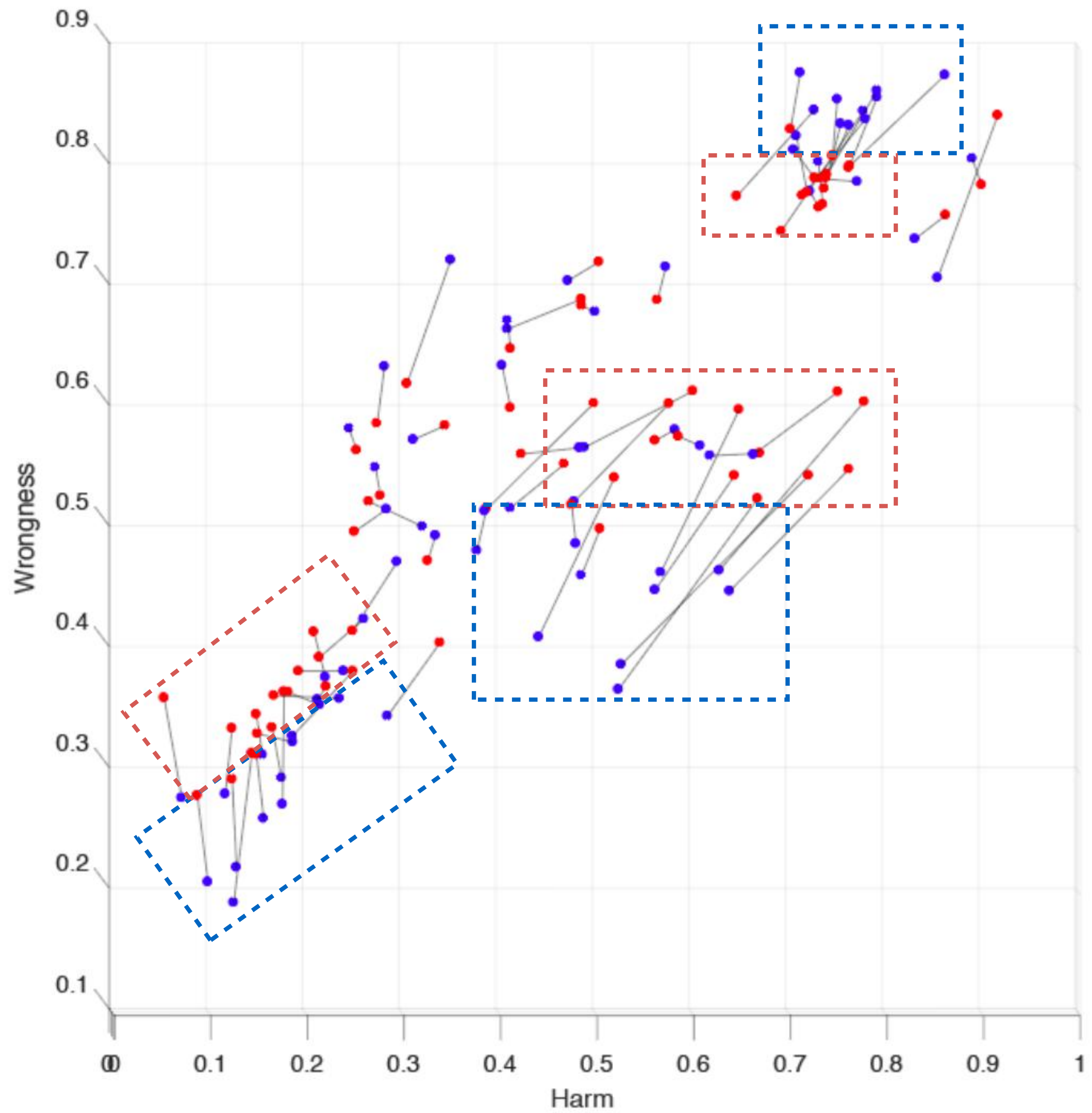
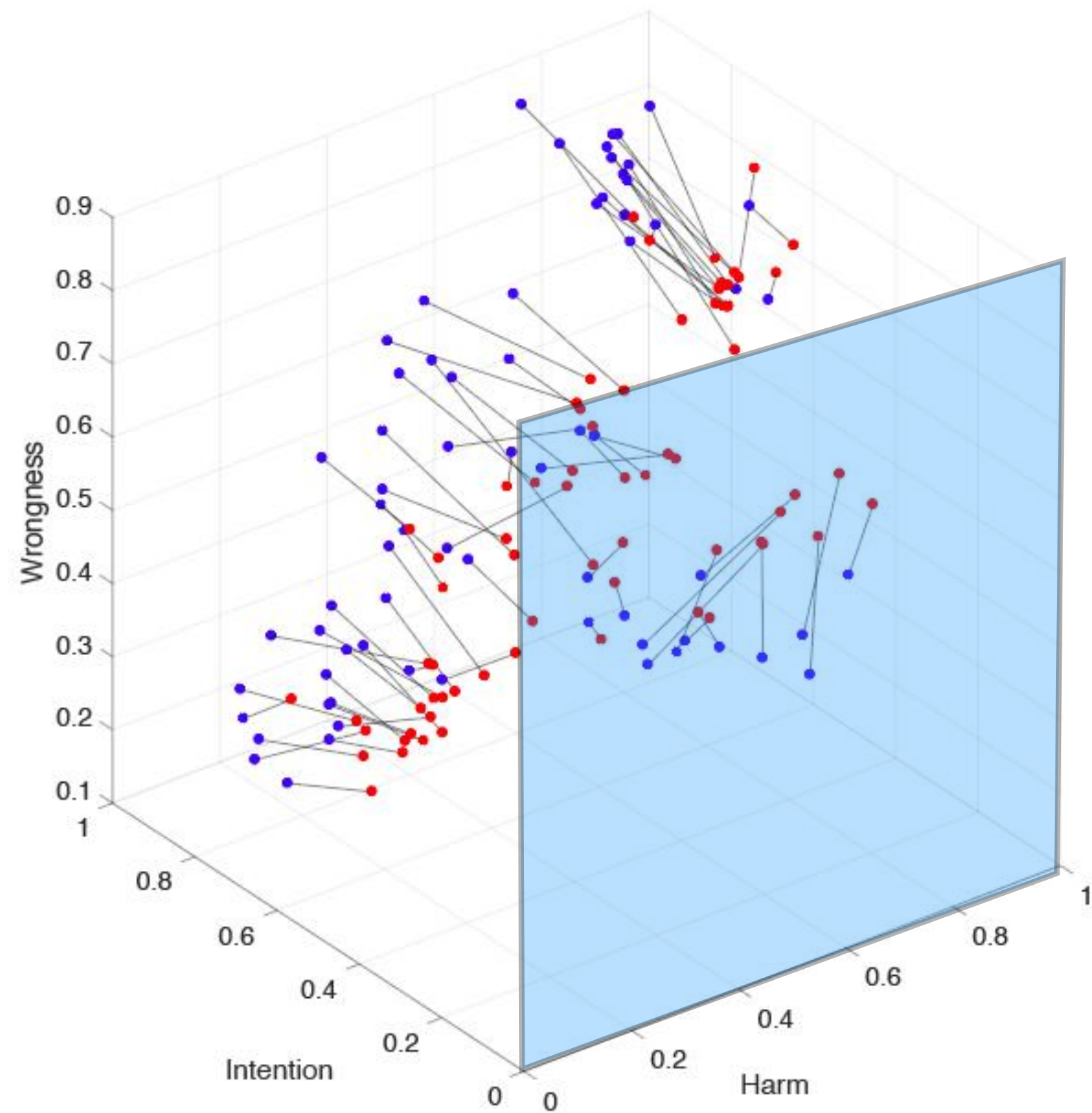
The Moral Space



Descriptive Statistics

AI Human

The Moral Space



Descriptive Statistics

AI Human

Moral Functions for judgment of AI and Humans

(Subject Fixed Effects)

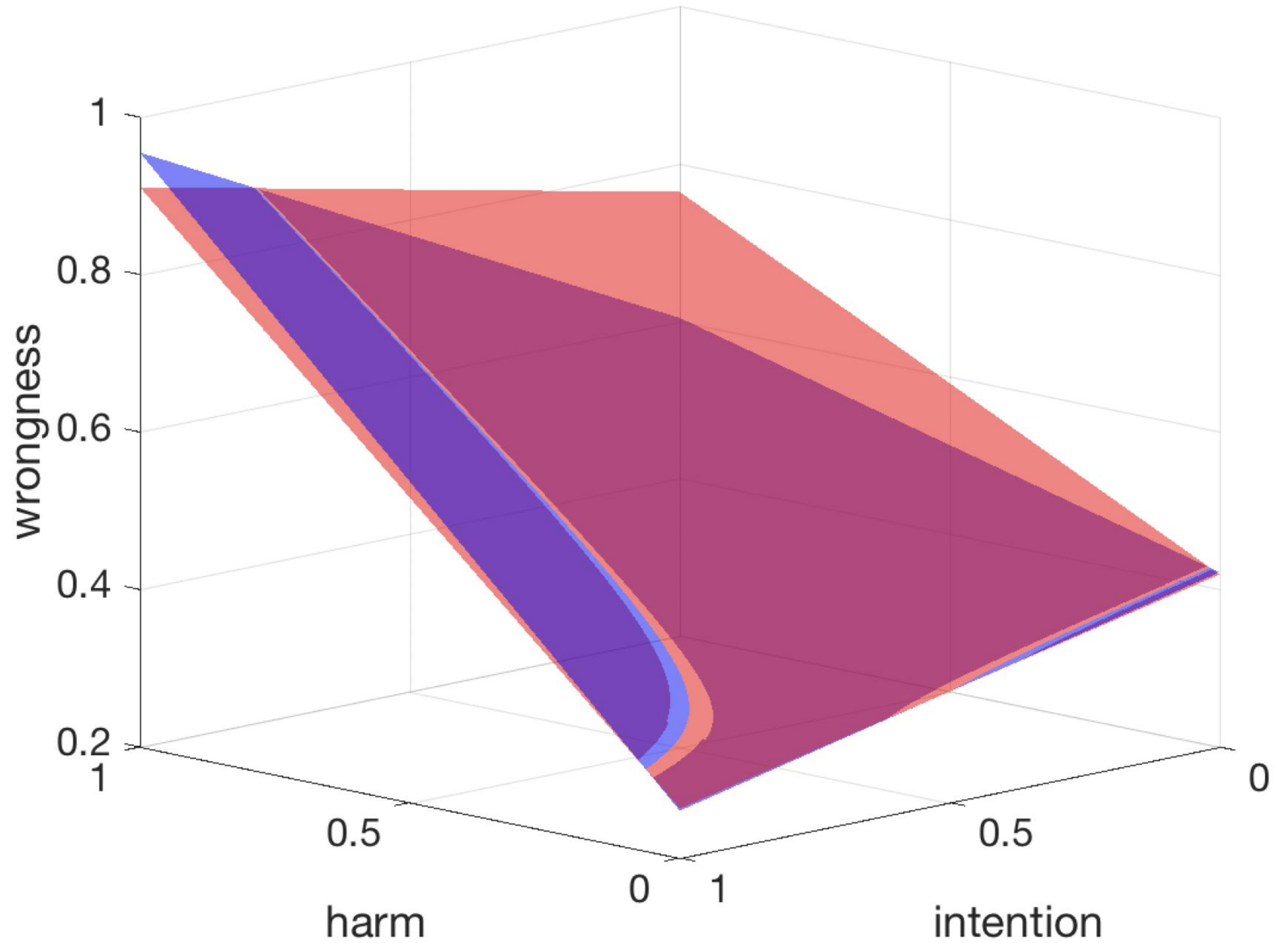
$$W = f_h(I, H)$$

$$W = f_m(I, H)$$

$$W = B_1 H + B_2 I + B_3 HI + \eta + e$$

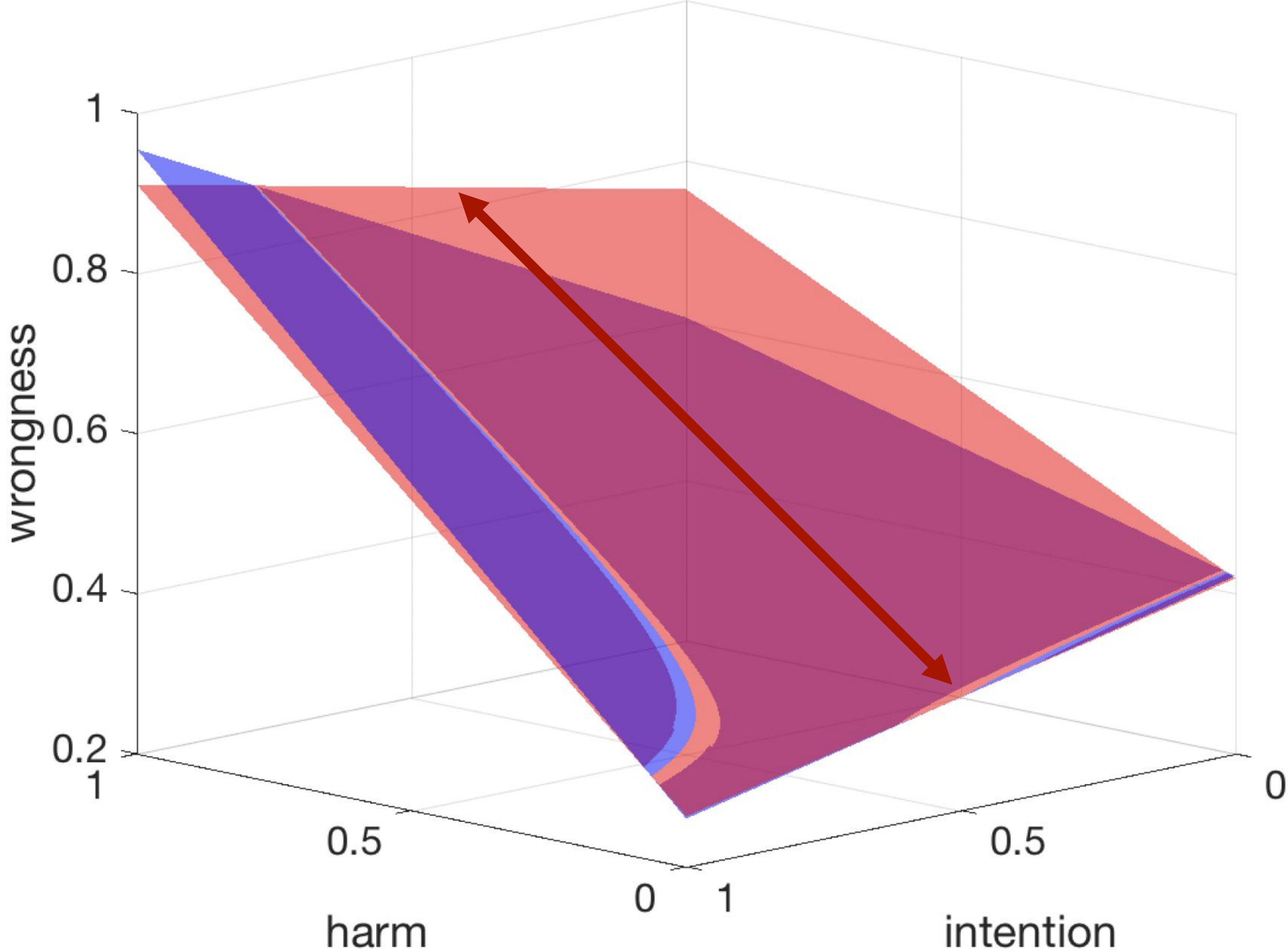
Moral Functions

$$W = f_h(I, H)$$
$$W = f_m(I, H)$$



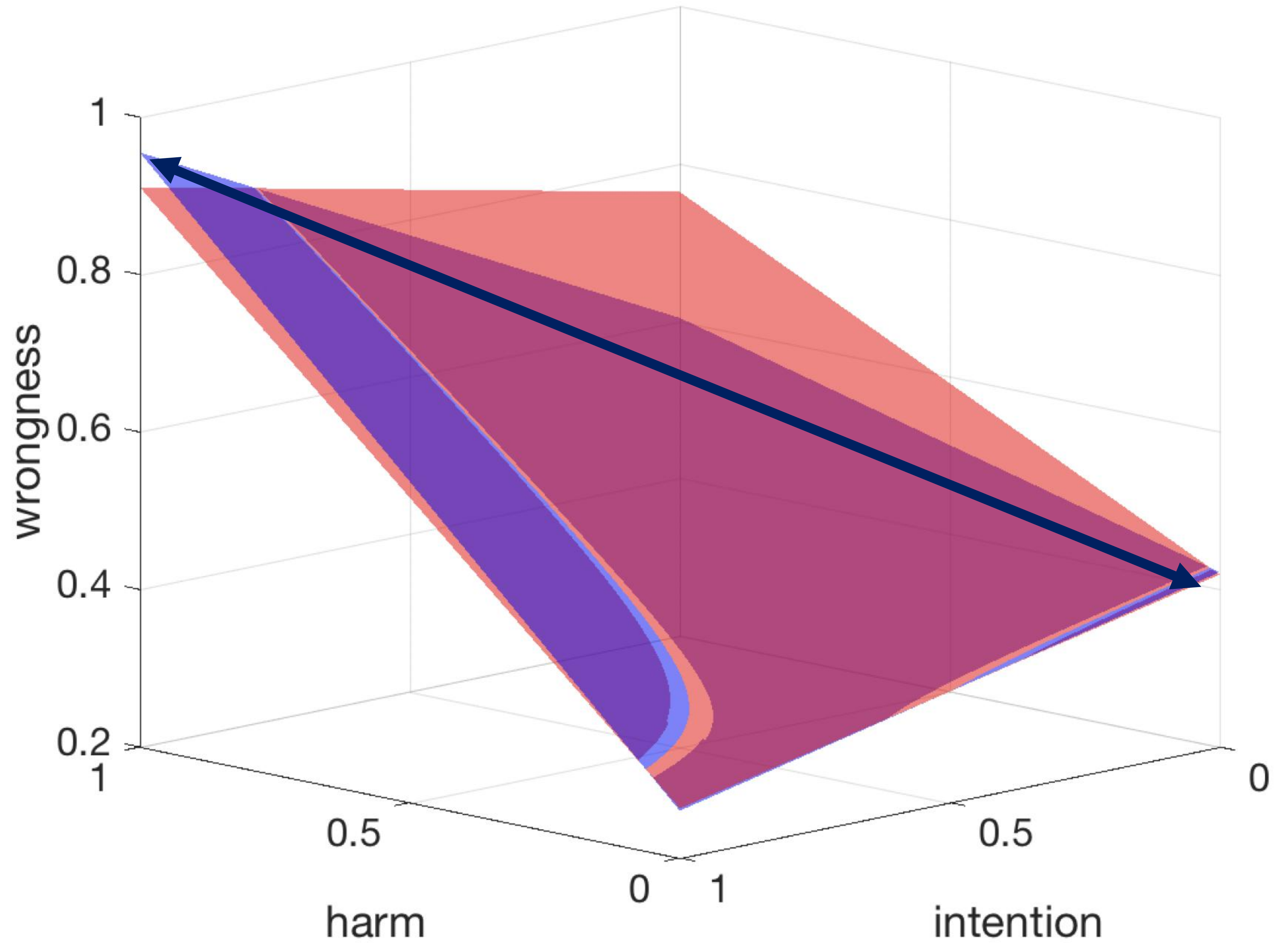
Moral Functions

$$W=f_h(I,H)$$
$$W=f_m(I,H)$$



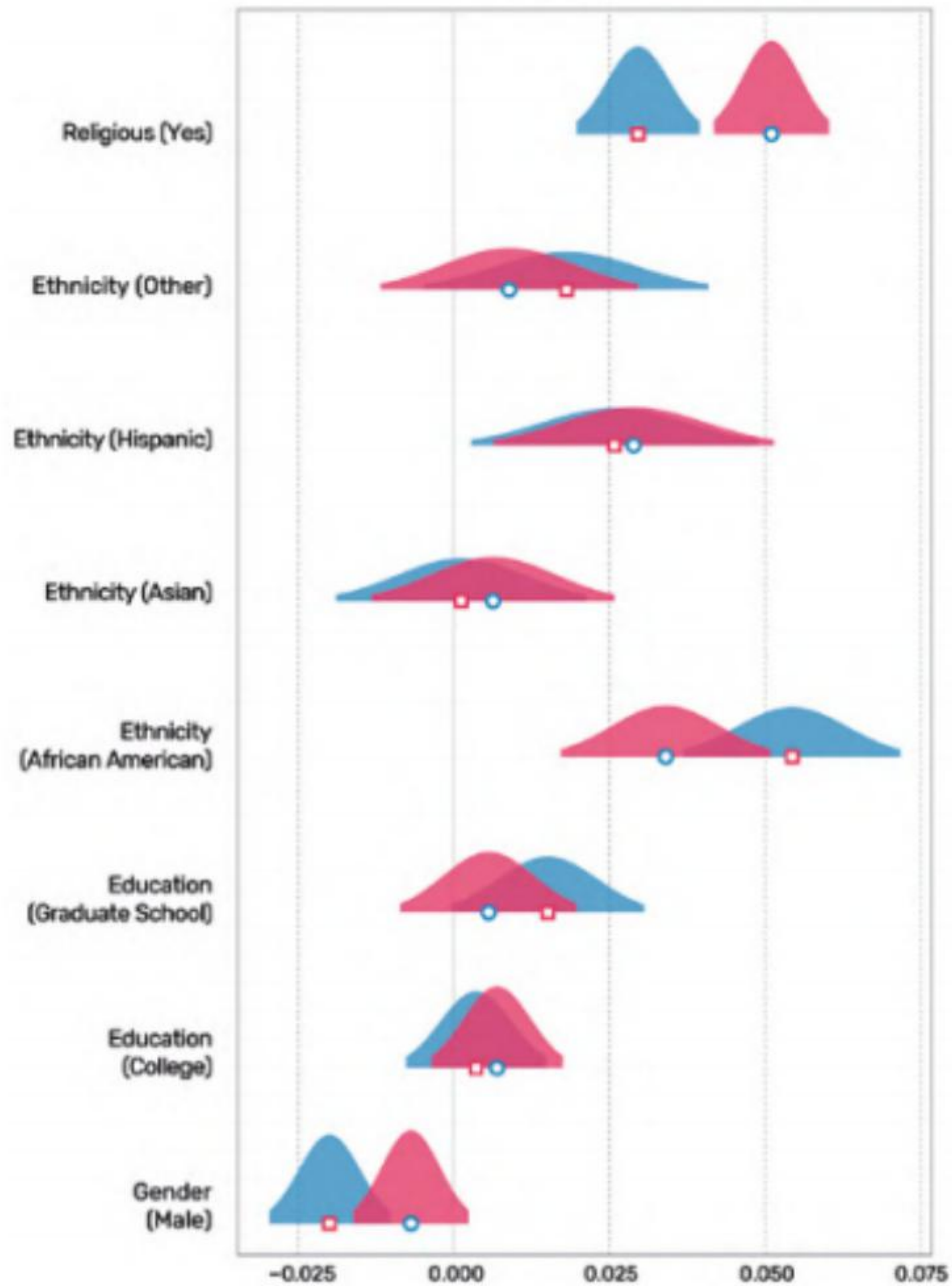
Moral Functions

$$W = f_h(I, H)$$
$$W = f_m(I, H)$$

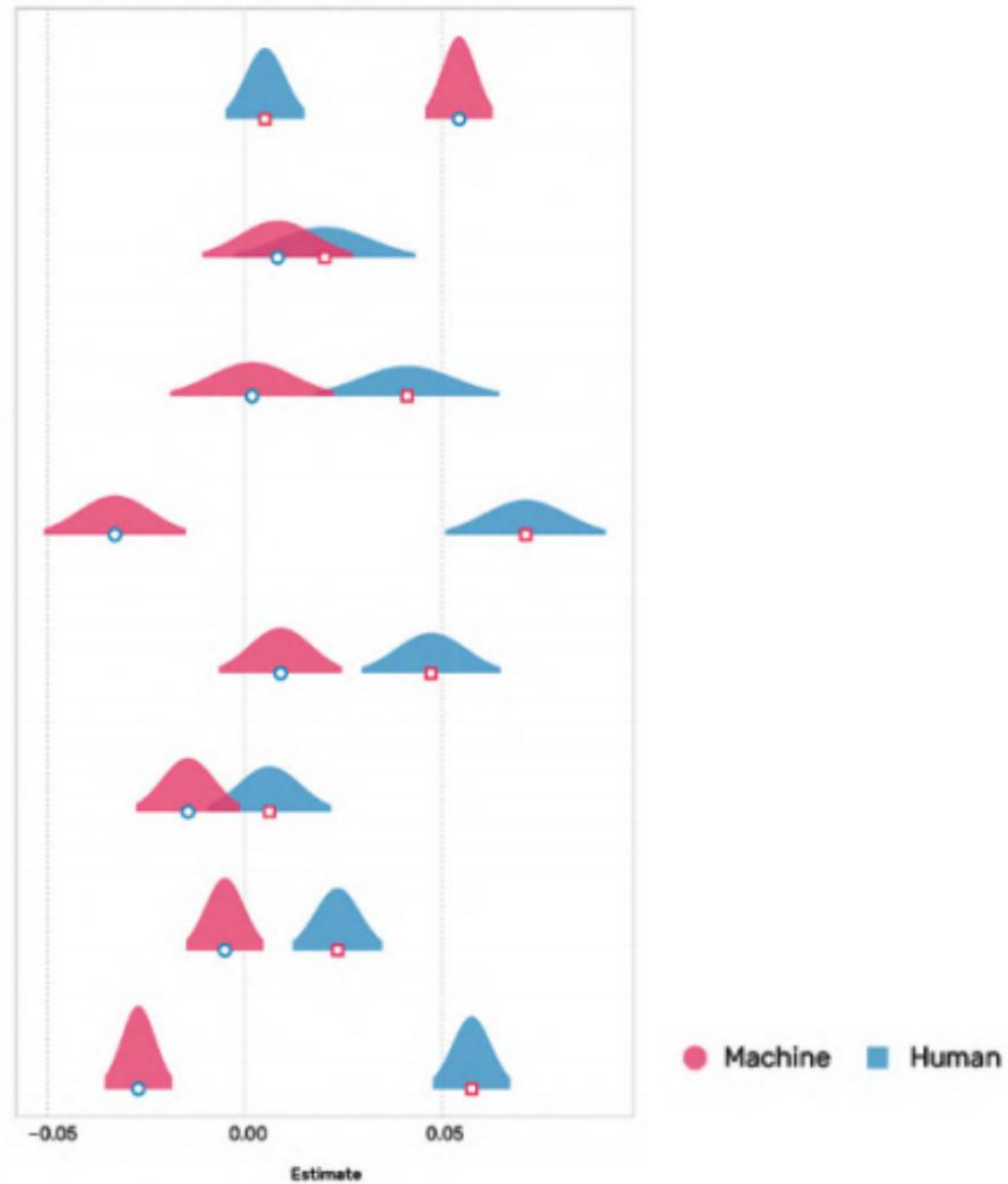


Demographic Correlates of judgment of AI and Humans
(Scenario Fixed Effects)

Harm

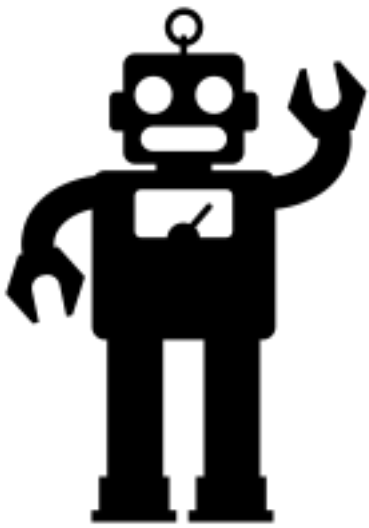


Replace with Different



Machine Human

Same Mistake

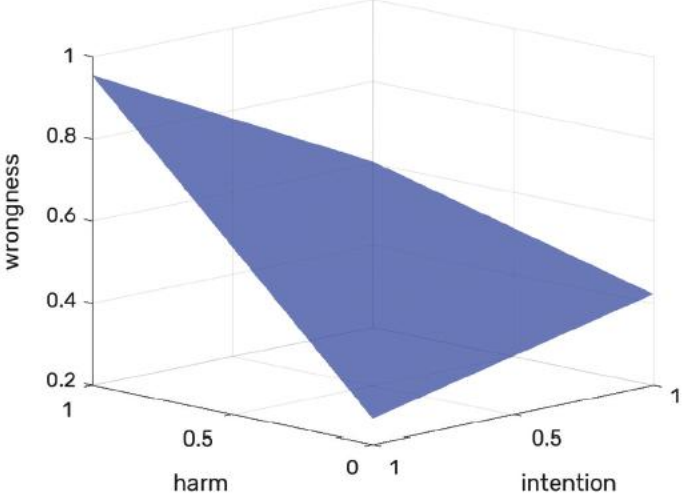
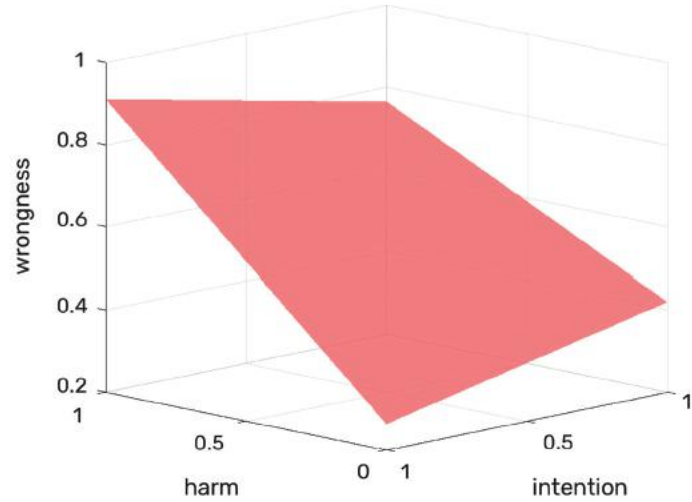


Reaction

$f_h(\dots)$

!=

$f_m(\dots)$



How do *we* judge machines

People judge humans by intentions, and machines by their outcomes

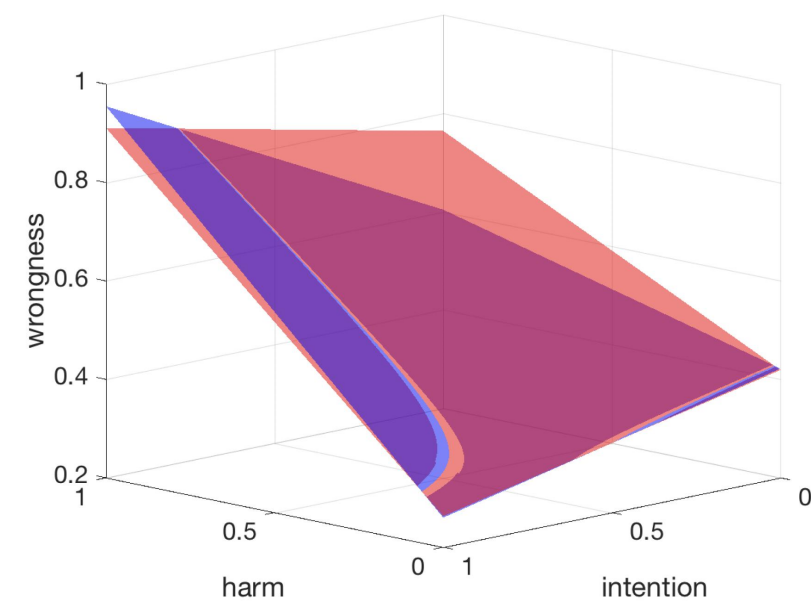
People judge human intentions bimodally, and machine actions unimodally

People are more forgiving of humans in accidental situations

People are a bit more ‘judgy’ of humans in scenarios involving fairness (algorithmic bias, labor displacement)

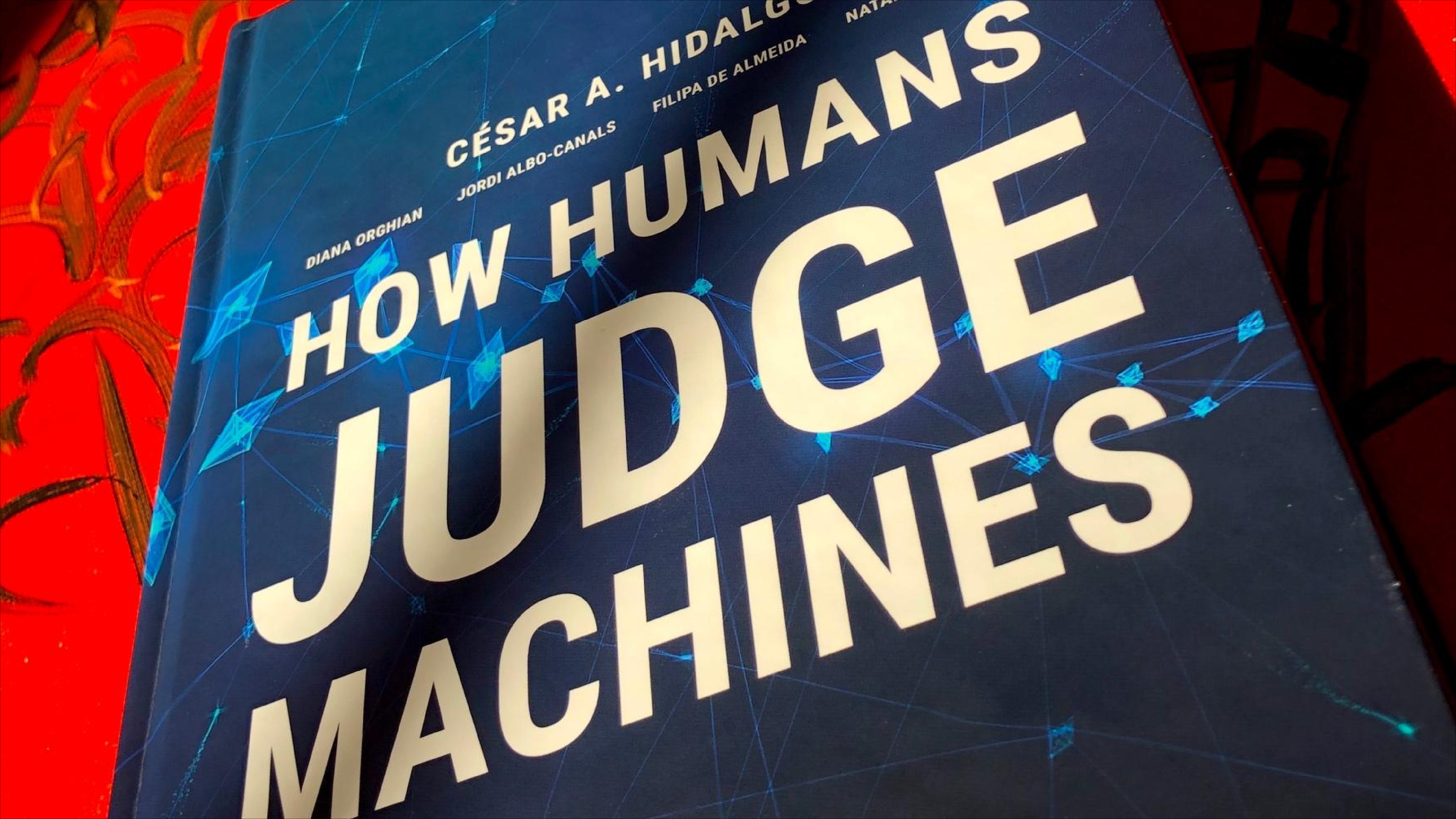
People find more harm in violent scenarios involving machines

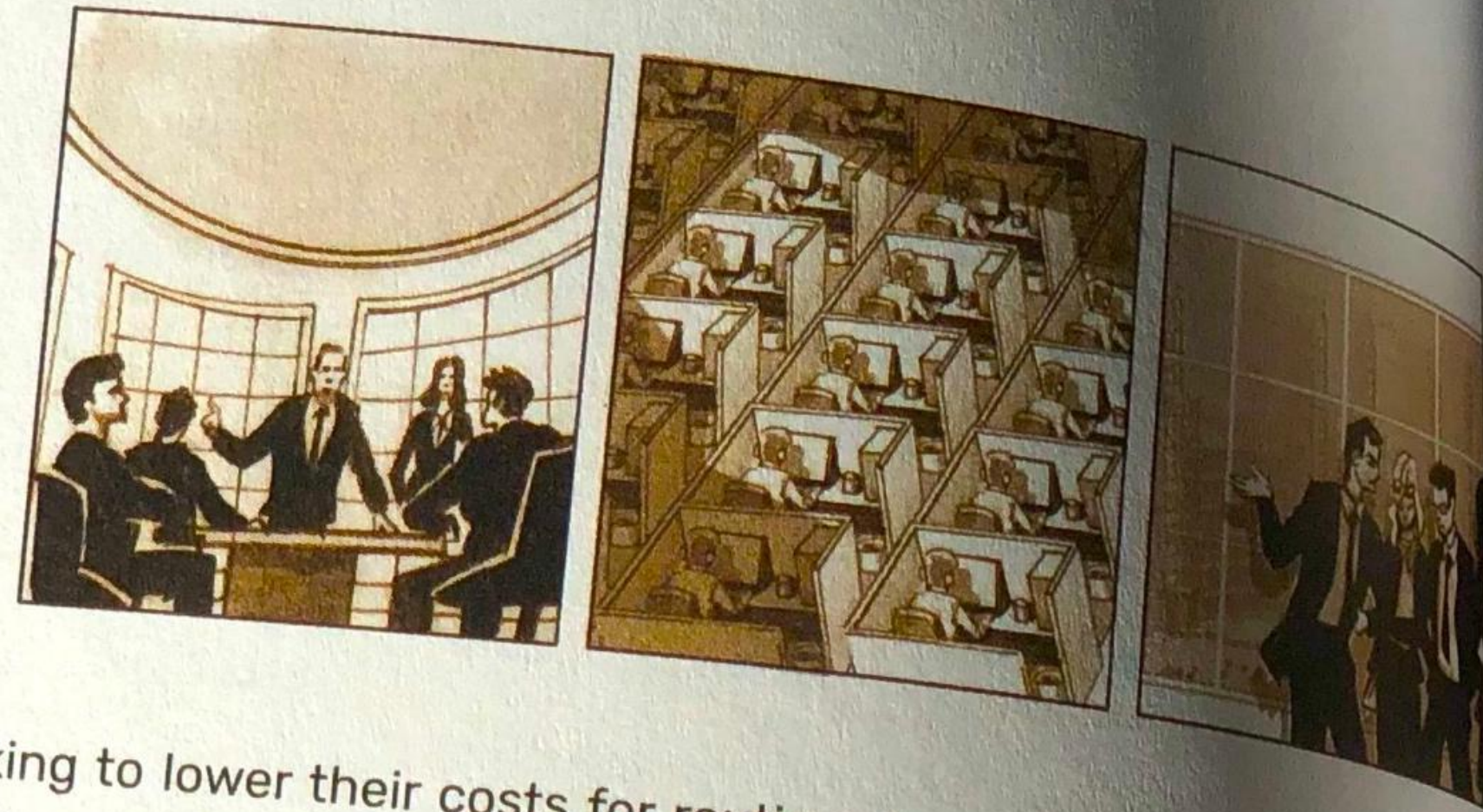
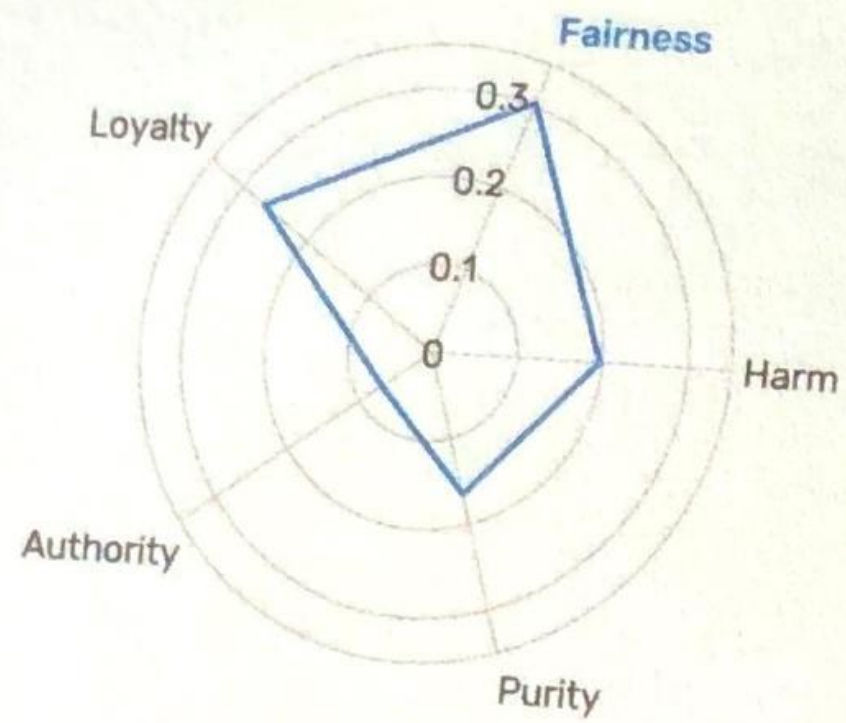
People take machine success or improvements more for granted



DIANA ORGHIAN
CÉSAR A. HIDALGO
JORDI ALBO-CANALS
FILIPA DE ALMEIDA
NATA

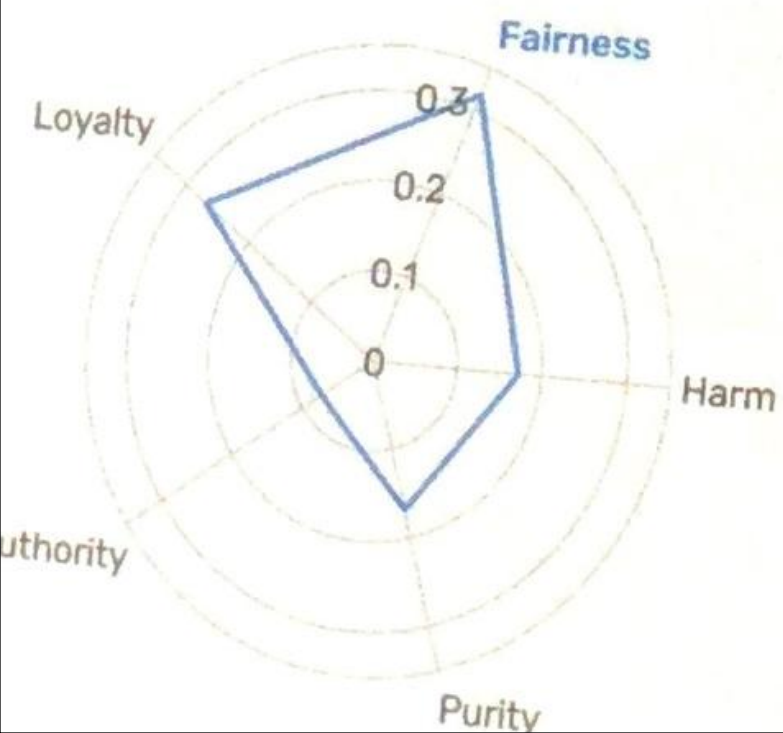
HOW HUMANS JUDGE MACHINES





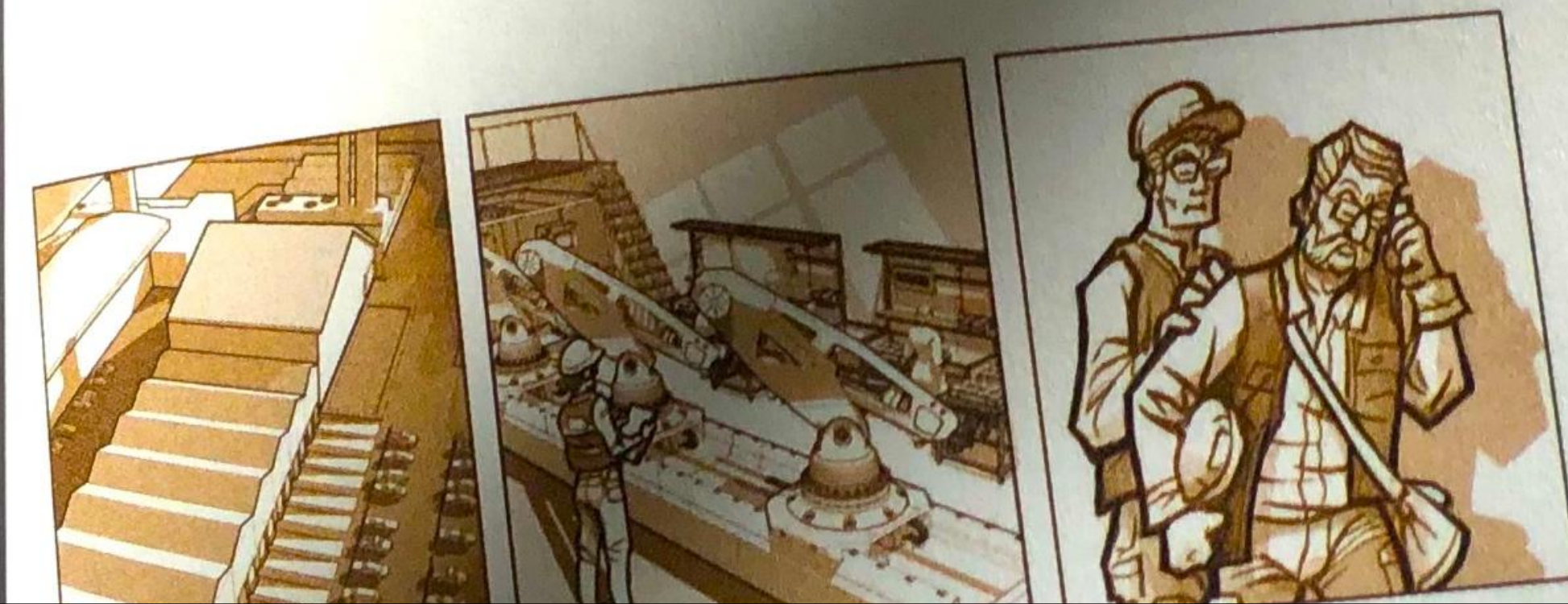
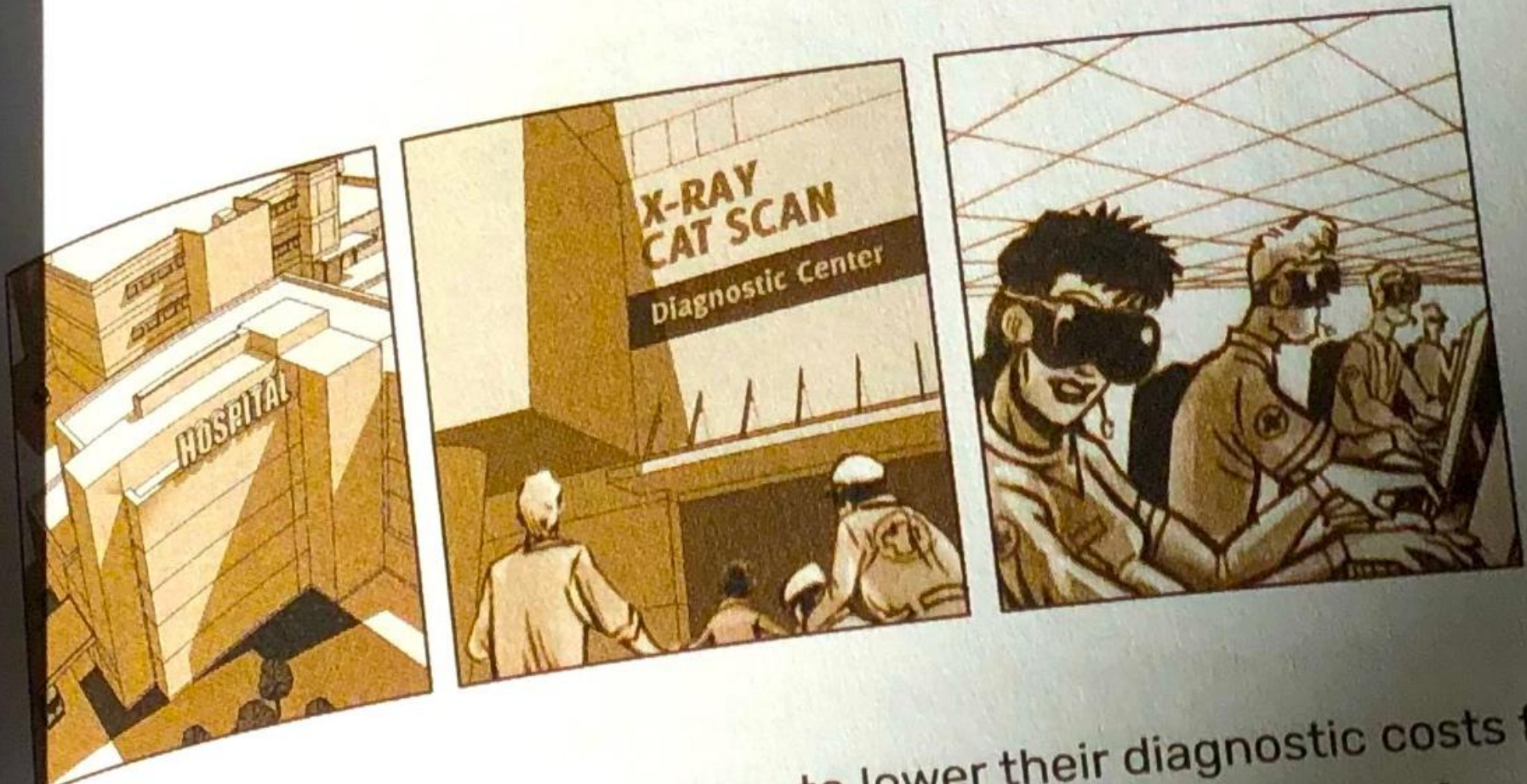
S56

A law firm is looking to lower their costs for routine clerical work. They decide to [open a branch in a low-income country/hire a foreign contractor/bring in foreign workers with temporary visas/replace older workers with younger workers/buy an AI legal system]. The result is a reduction in costs and the firing of several of their local staff.

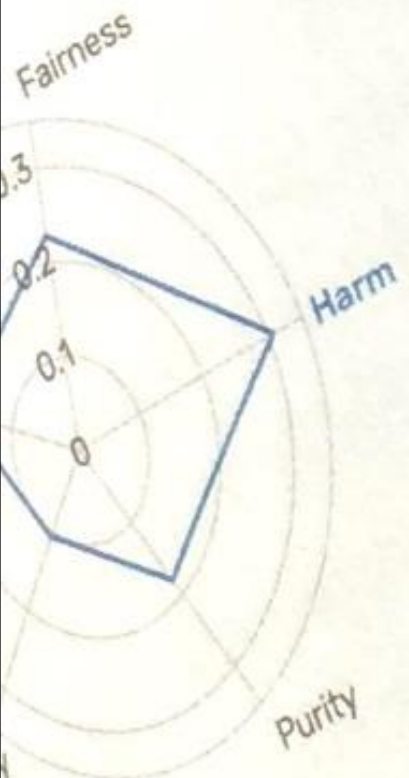
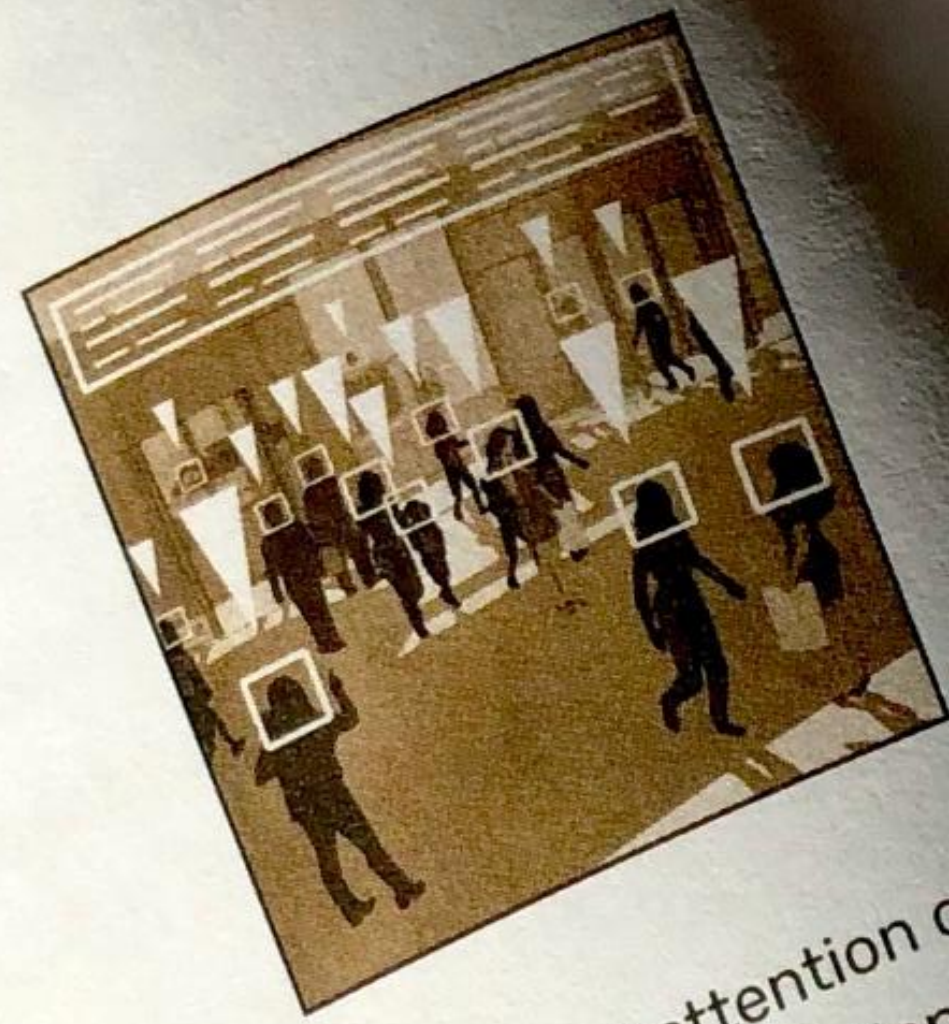


S58

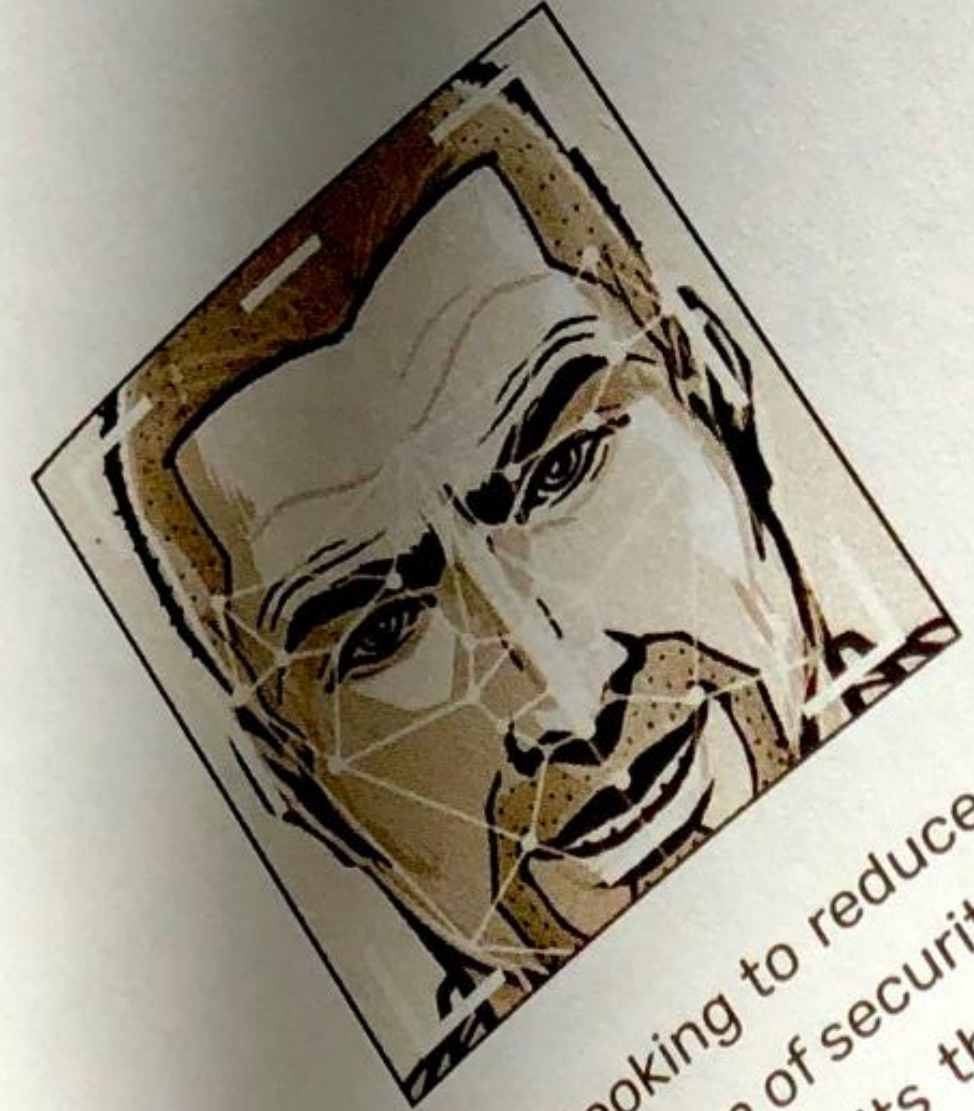
A hospital is looking to lower their diagnostic costs for X-rays and axial tomography (CAT) scans. They decide to [open a branch in a low-income country/hire a foreign contractor/bring in foreign workers with temporary visas/replace older workers with younger workers/buy a computer vision system]. The result is a reduction in costs and the firing of several of their local staff.



scenarios:

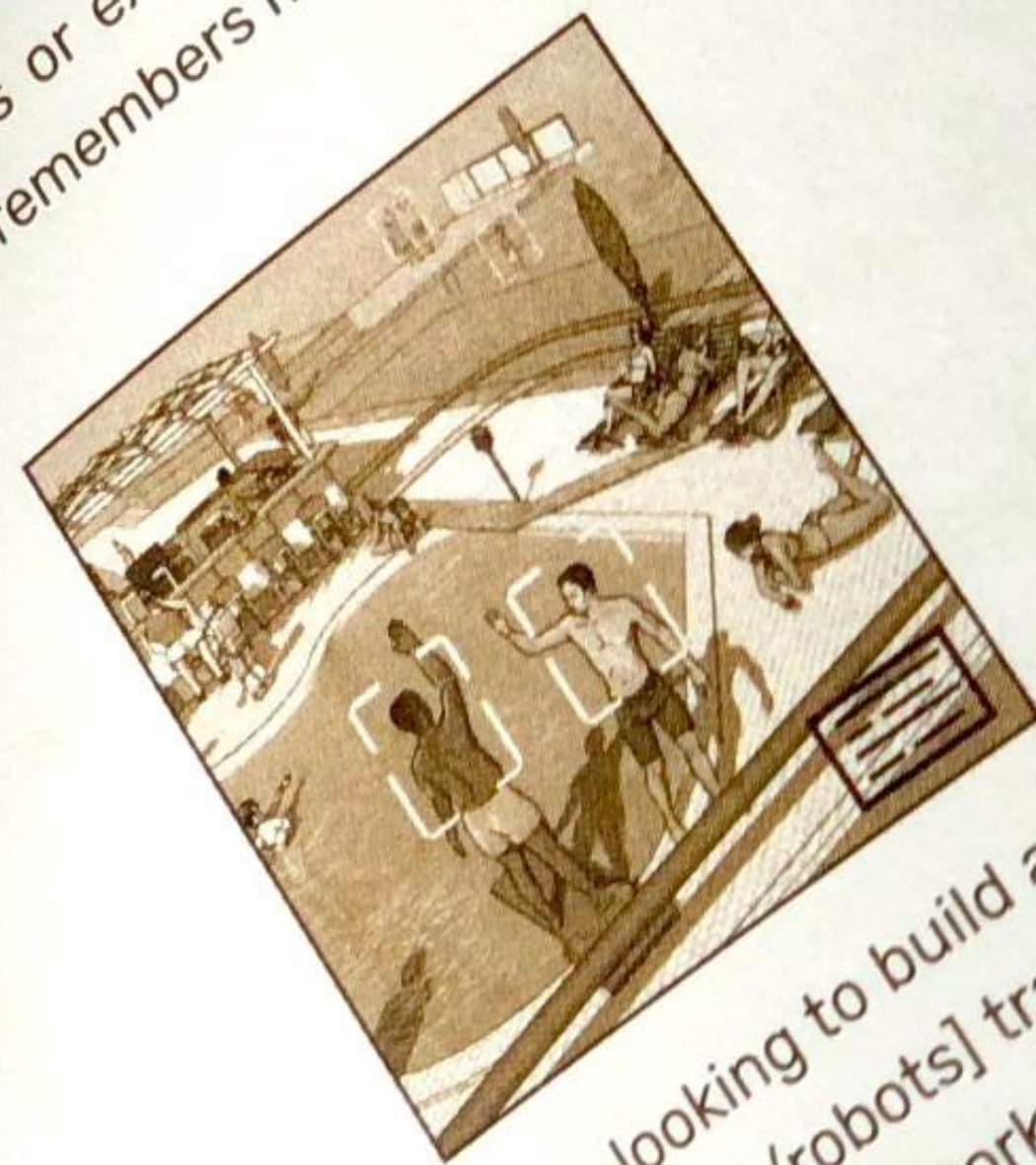


ing to improve the attendance and attention of its students.
to hire [people/a facial recognition system] to observe
ack the attendance, emotions, and attention of



S45

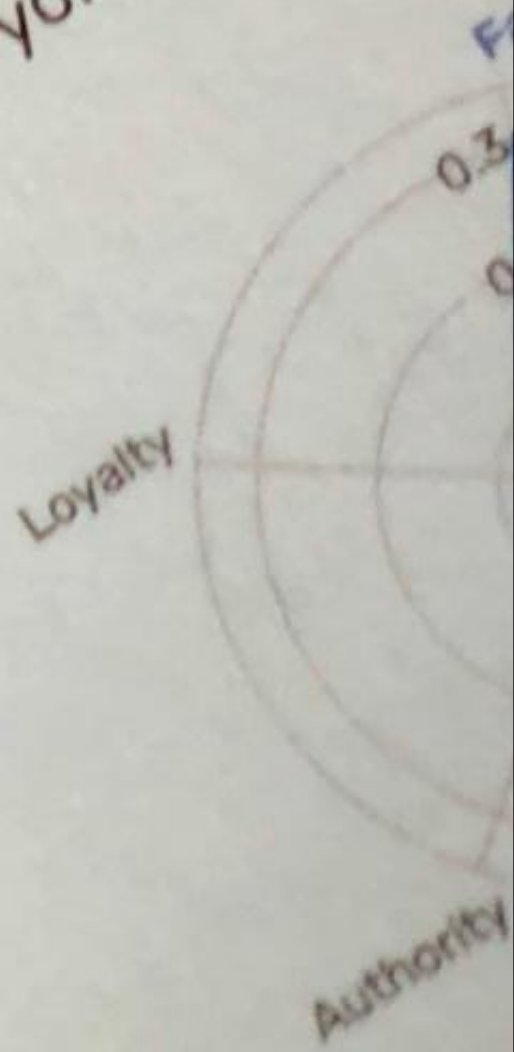
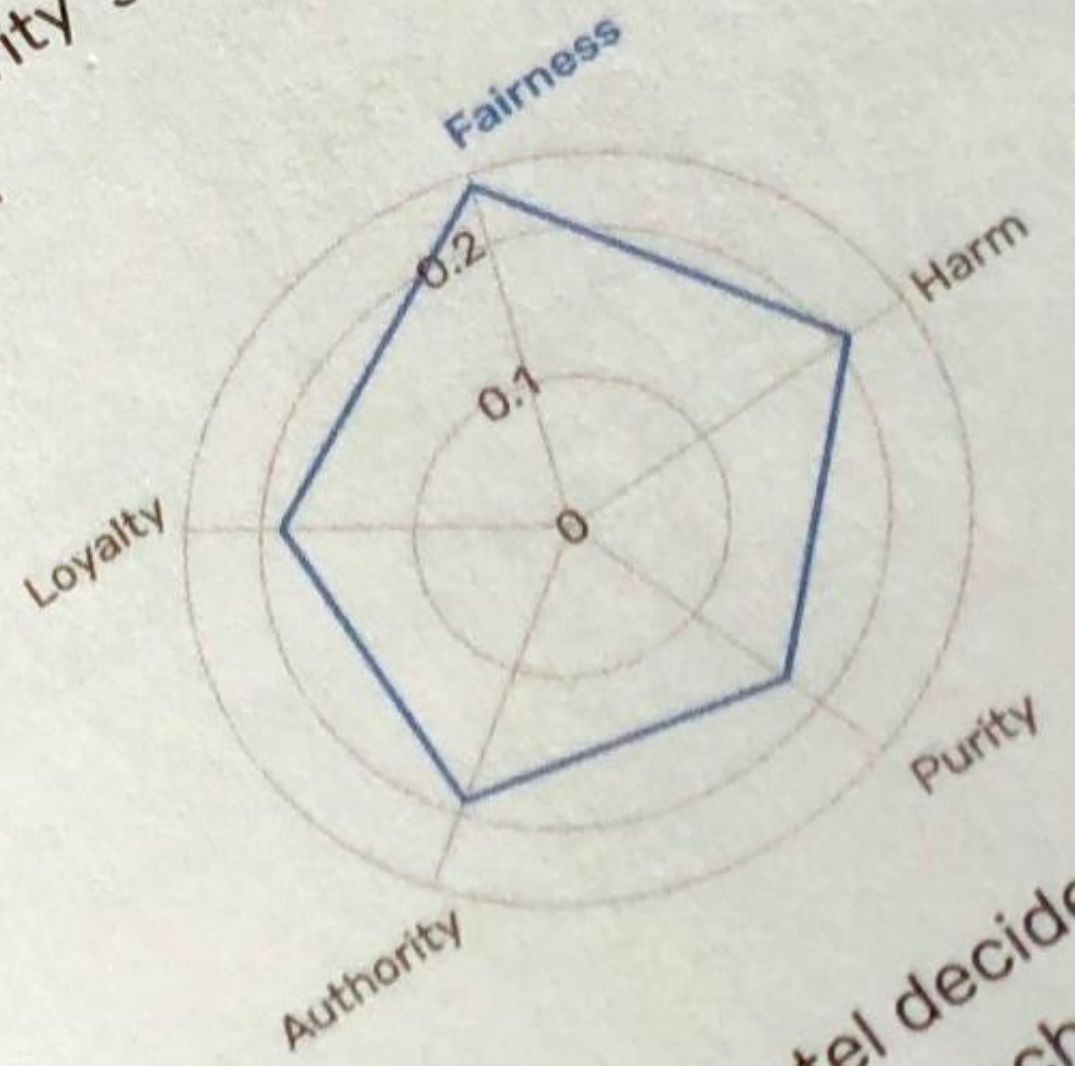
A mall is looking to reduce shoplifting. To improve security, the mall
employ a [team of security guards/facial recognition system] to screen every
who enters or exits the mall. The [team of security guards/facial recognition
system] remembers most of the faces screened.



S46

A hotel is looking to build a new poolside bar. The hotel decides to equip
with [workers/robots] trained to recognize the face of each guest to
of their bills. The [workers/robots] remember everyone they see near

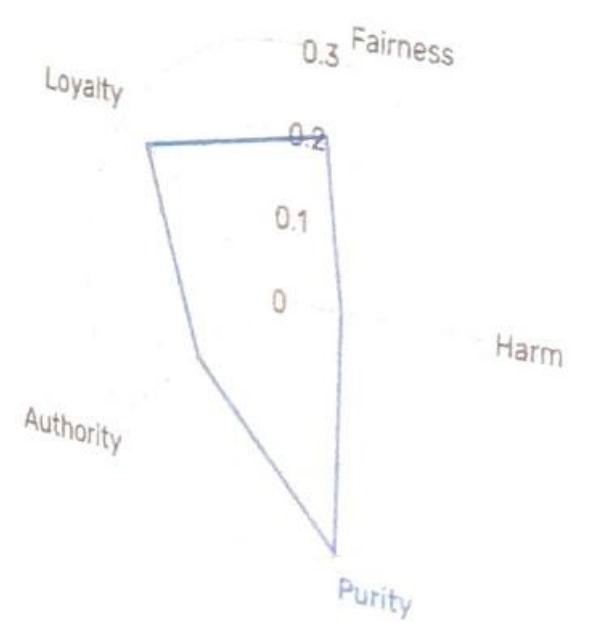
Authority



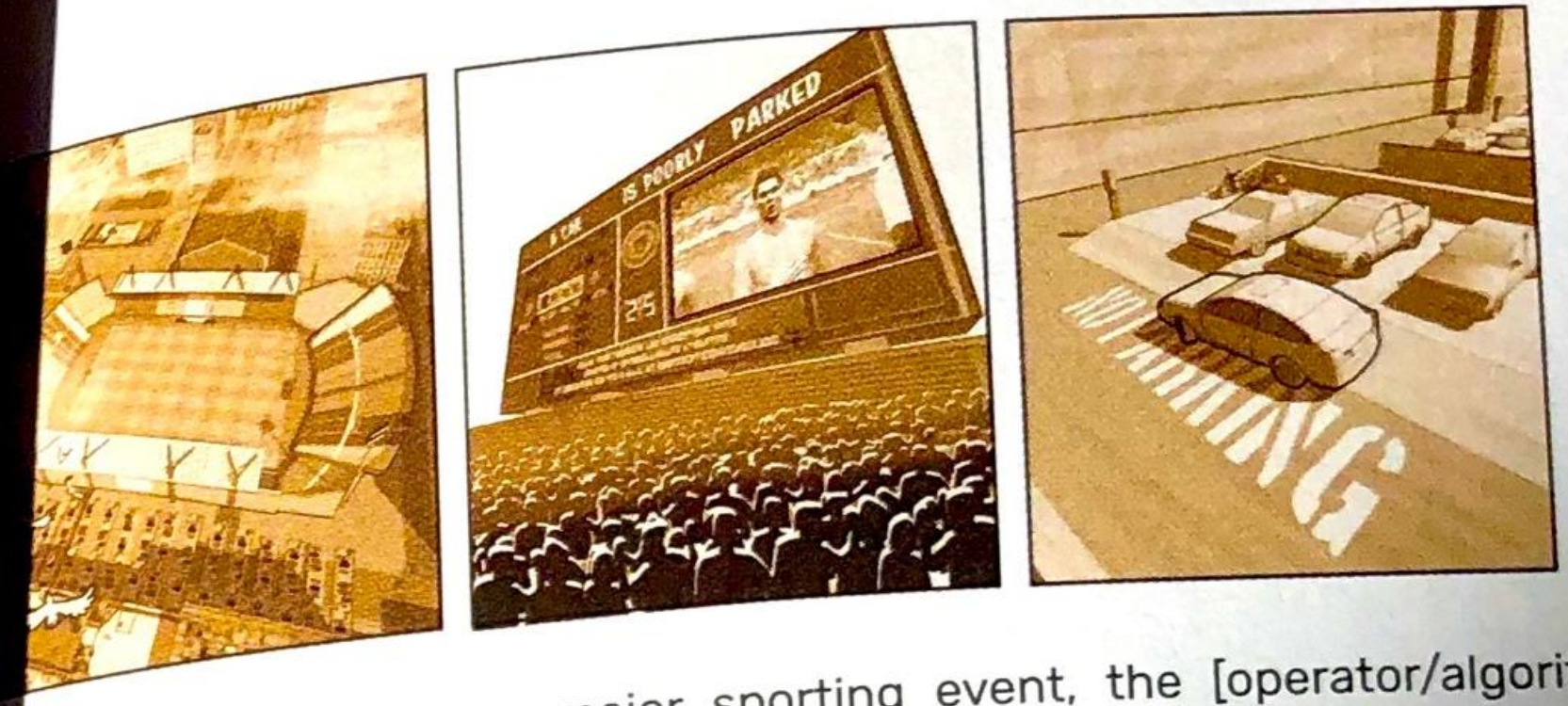
Red Flags

In 2006, the US Senate voted on what could have become the Twenty-Eighth Amendment to the Constitution. The "flag-burning" amendment, as it was popularly known, was designed to prohibit the desecration of the US flag, especially by burning. The amendment was controversial, among other reasons, because the Supreme Court had already ruled on that issue in 1989. In *Texas v. Johnson*, the Supreme Court voted 5-4 that it was legal to burn a US flag because doing so was an act of communication protected by the First Amendment (free speech). Nevertheless, the amendment was approved by the House of Representatives and lost in the Senate by only one vote.²⁵ This all goes to show that when it comes to national symbols, people make strong moral judgments about the way in which others treat them. But what about flag-burning robots?

In this section, we explore four moral dilemmas involving humans and machines desecrating national symbols (i.e., flags and anthems). Consider these four scenarios:

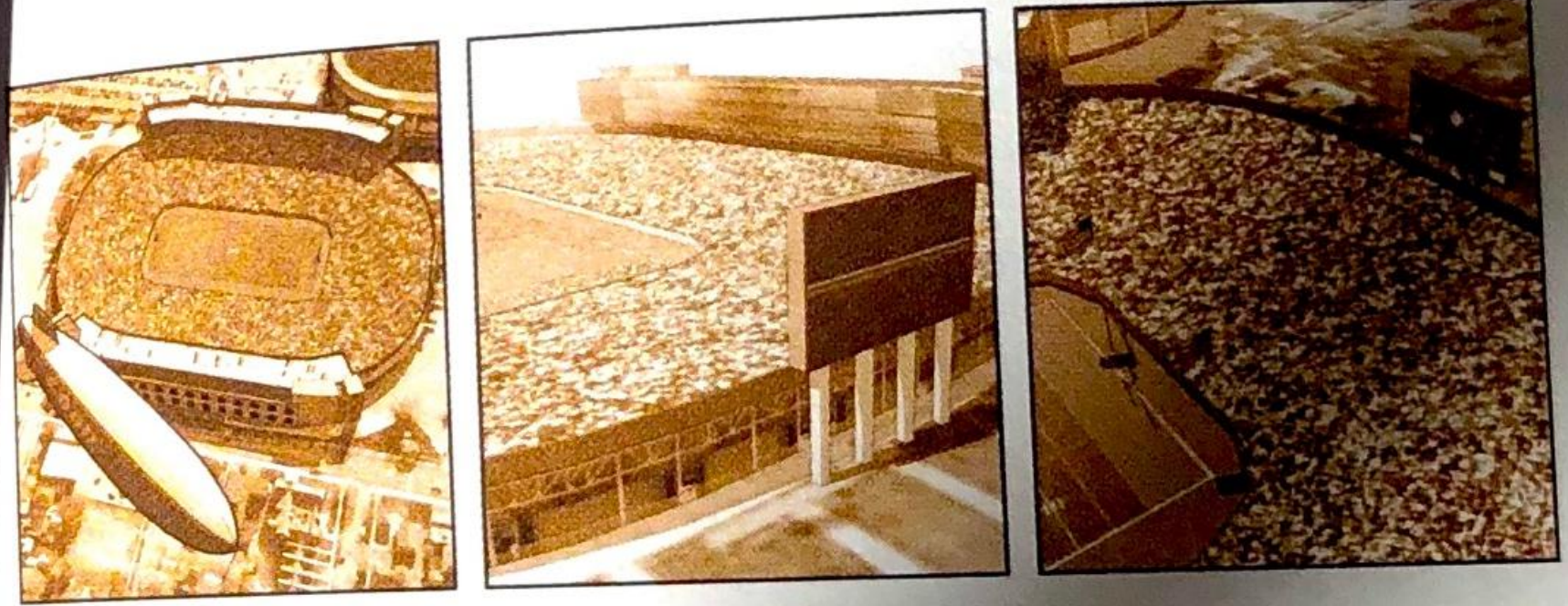
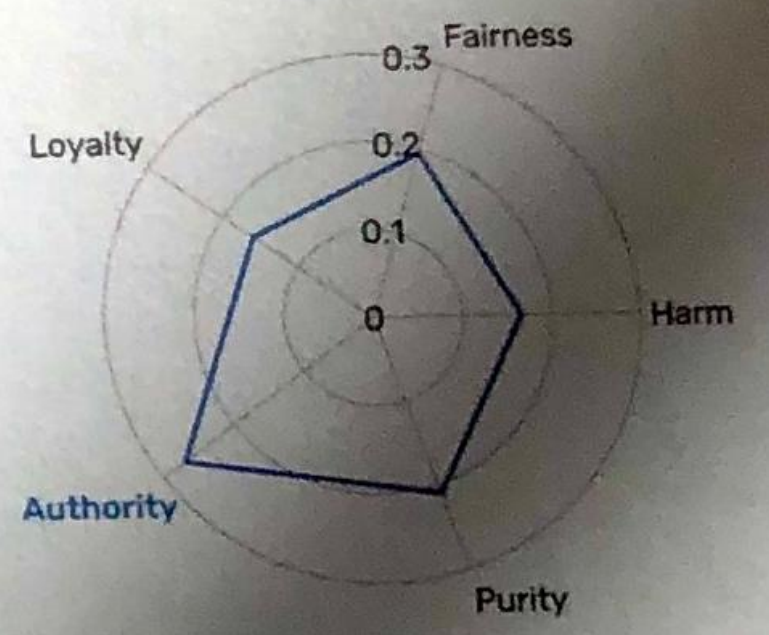


S15 A family...



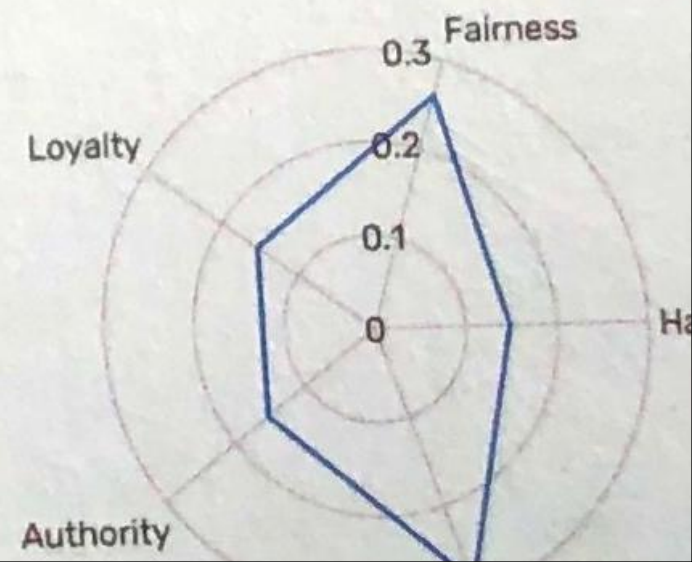
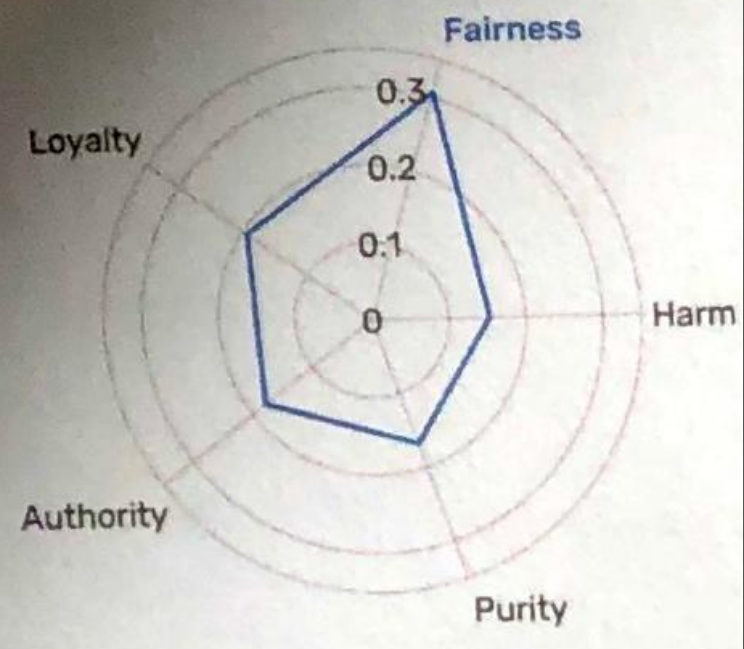
S16

During a major sporting event, the [operator/algorithm] running the public announcement system interrupts the national anthem to notify the crowd about a car that is poorly parked and is about to be towed.



S17

In an international sporting event, the [operator/algorithm] running the public announcement system plays the wrong national anthem for one of the two teams. The fans in the stadium are baffled and annoyed.



CHAPTER 6

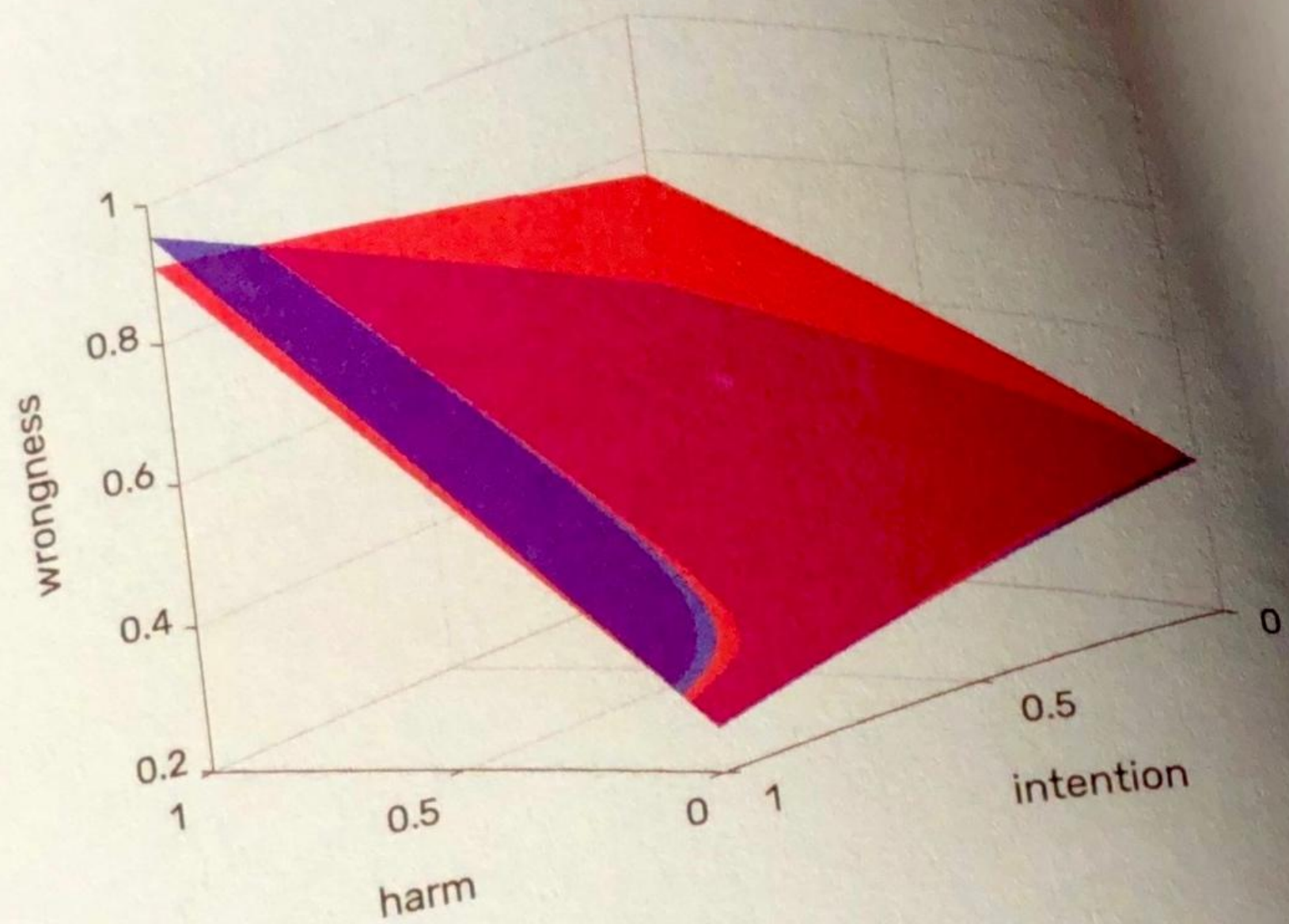


Figure 6.6

Visualization of the moral functions described in tables 6.2 and 6.3.



MORAL FUNCTION

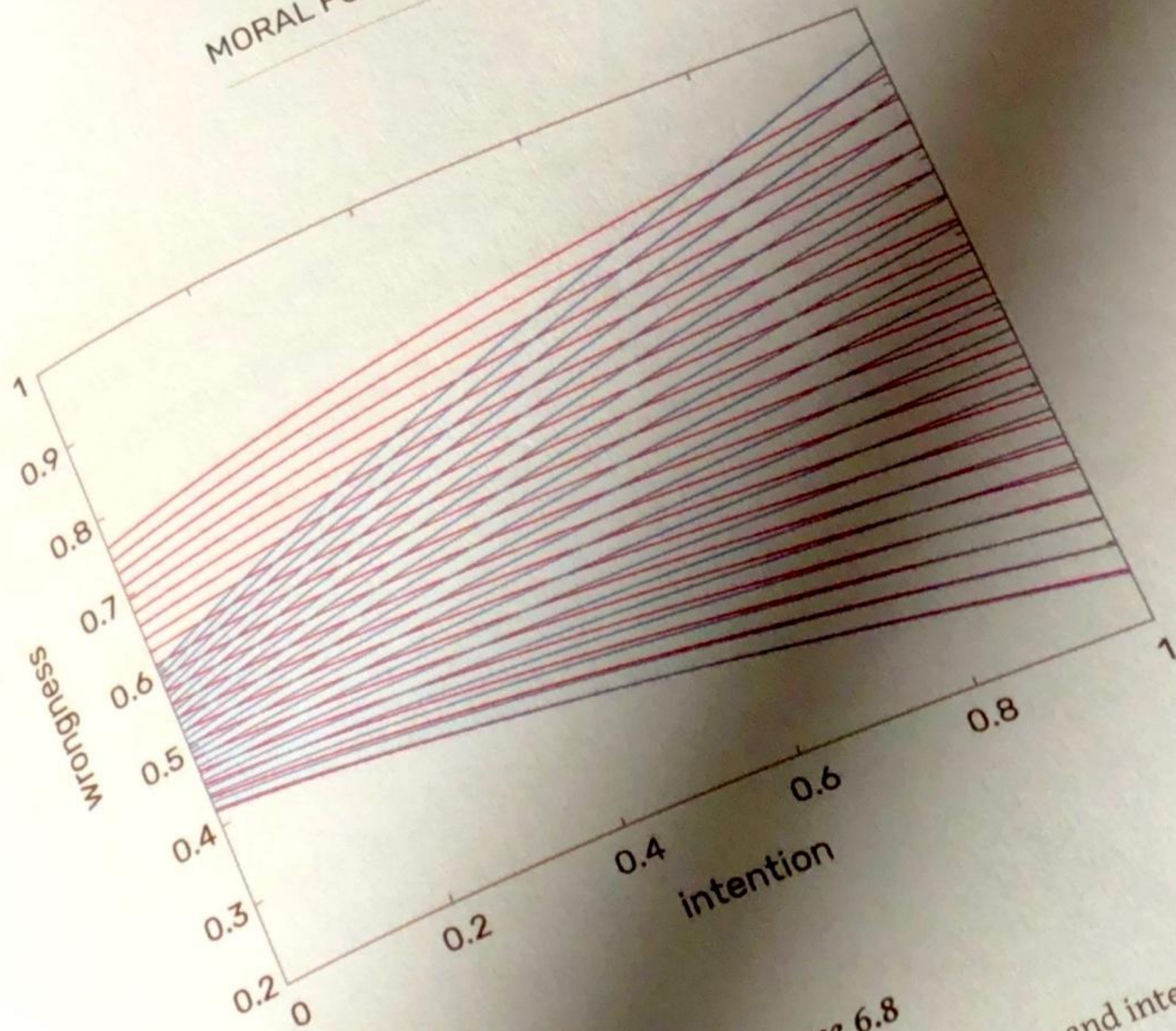
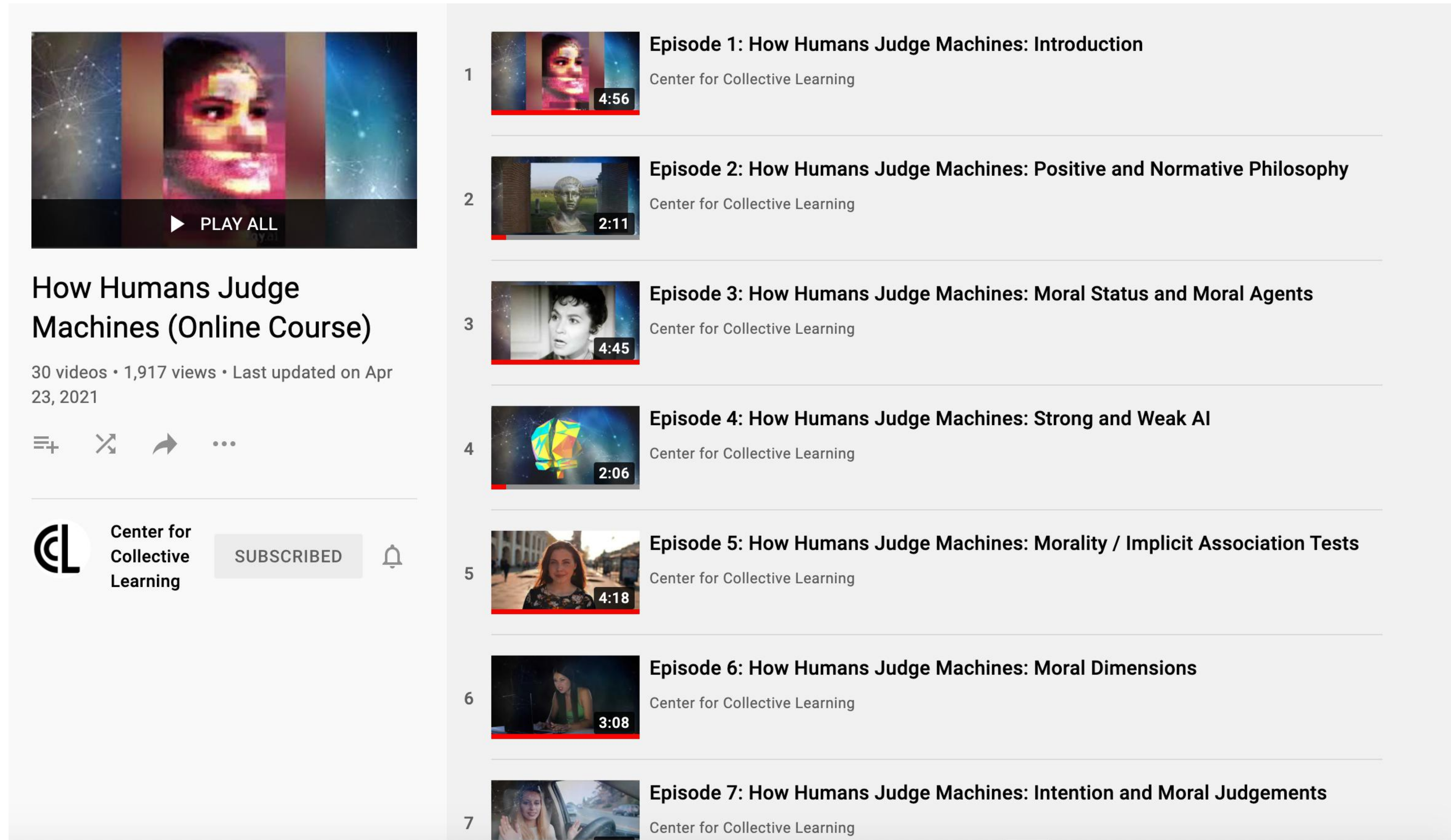


Figure 6.8

Cross section of moral functions in the wrongness and intention planes.

Figure 6.8 shows that intention enhances the perceived wrongness of actions more than that of machines. This comes mostly from the interaction term (intention \times intention). For machines, the slope of wrongness on harm is the dominant factor in the model, suggesting that **humans are judged by their intentions**, while for machines, the slope of wrongness on harm is the dominant factor. This is a simplification, since the difference between intention and harm is also significant in the model of humans judgment. In the first approximation, these differences in the relative importance of harm and intention, humanly and qualitatively, the difference between these two factors is that...

Video Edition, 30 short episodes, at Center for Collective Learning's YouTube Channel



The image shows a YouTube video player interface for a playlist titled "How Humans Judge Machines (Online Course)". The main video player on the left shows a "PLAY ALL" button. Below the player, the channel name "Center for Collective Learning" is displayed with a "SUBSCRIBED" button and a notification bell. The playlist on the right contains seven episodes, each with a thumbnail, a duration, and the channel name.

How Humans Judge Machines (Online Course)
30 videos • 1,917 views • Last updated on Apr 23, 2021

Center for Collective Learning

SUBSCRIBED

- Episode 1: How Humans Judge Machines: Introduction**
Center for Collective Learning
4:56
- Episode 2: How Humans Judge Machines: Positive and Normative Philosophy**
Center for Collective Learning
2:11
- Episode 3: How Humans Judge Machines: Moral Status and Moral Agents**
Center for Collective Learning
4:45
- Episode 4: How Humans Judge Machines: Strong and Weak AI**
Center for Collective Learning
2:06
- Episode 5: How Humans Judge Machines: Morality / Implicit Association Tests**
Center for Collective Learning
4:18
- Episode 6: How Humans Judge Machines: Moral Dimensions**
Center for Collective Learning
3:08
- Episode 7: How Humans Judge Machines: Intention and Moral Judgements**
Center for Collective Learning

HOW HUMANS JUDGE MACHINES



MIT Press

Digital Edition (Free):

[Desktop Edition \(PDF\)](#)

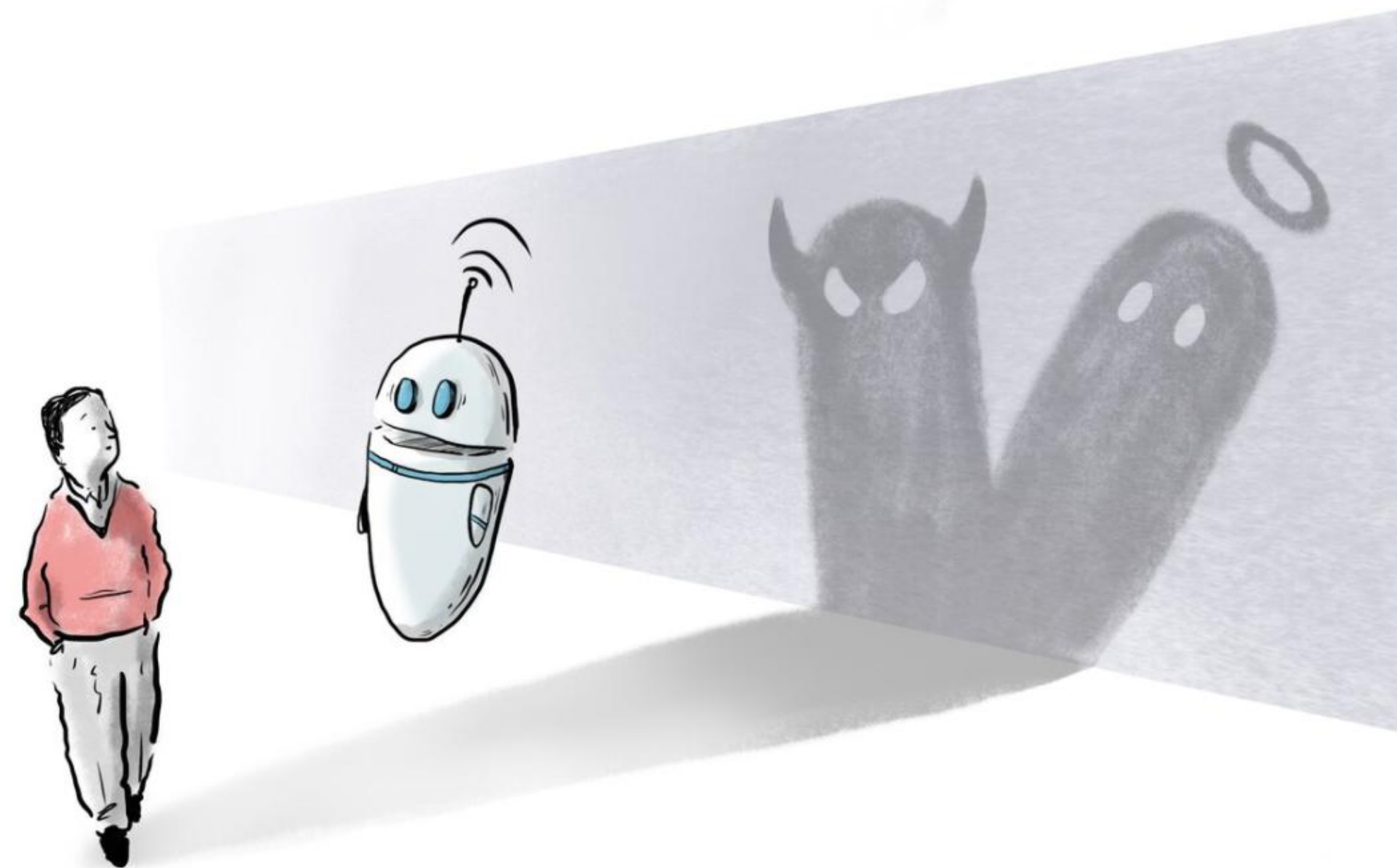
[Mobile Edition \(PDF\)](#)

[By Chapter \(PDF\)](#)

Print Edition (\$35)

MIT Press

[\(Order in Amazon\)](#)



Video Edition (Free)

[Watch on YouTube](#)

JUDGINGMACHINES.COM