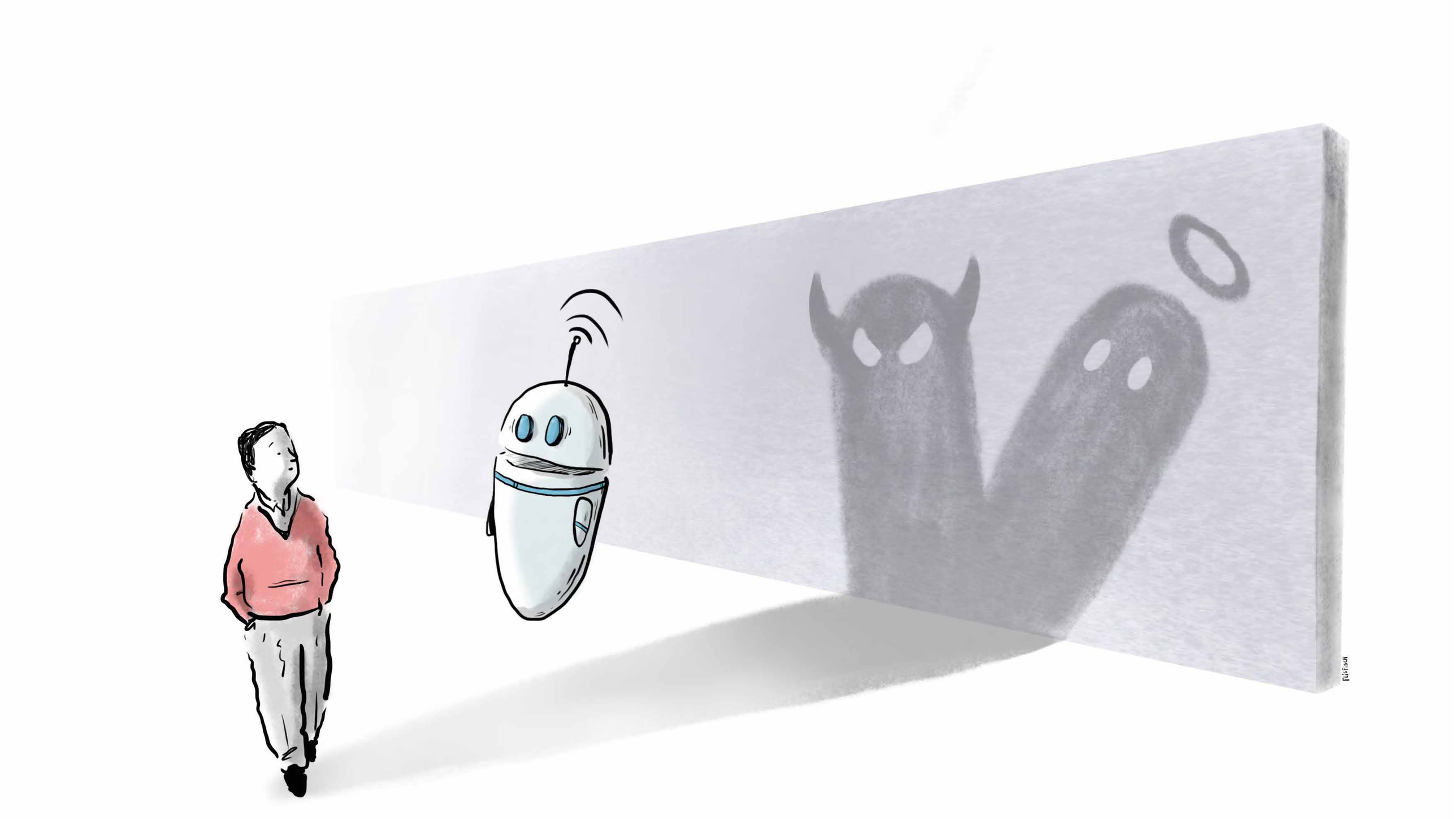
How Humans Judge Machines?

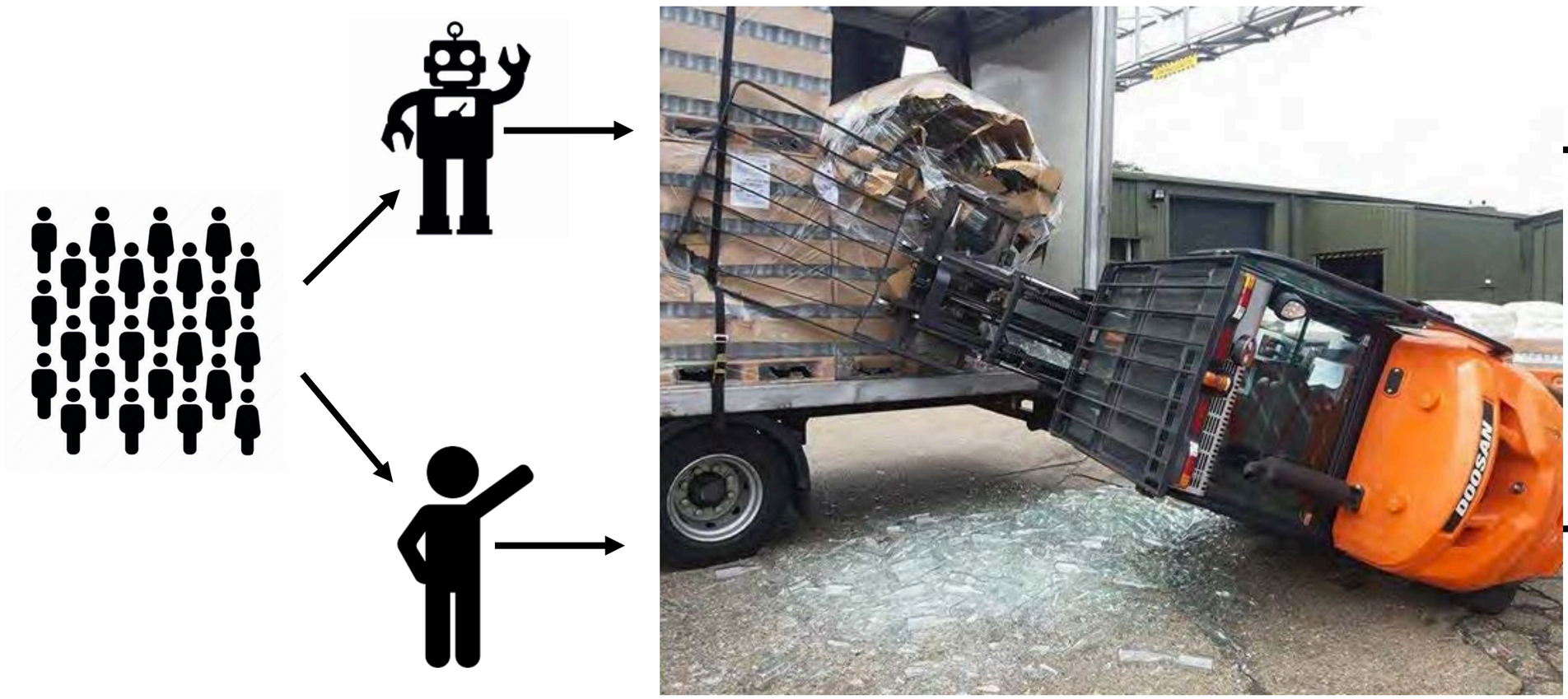
Cesar A. Hidalgo
Augmented Society Chair

ANITI
Université Fédérale
Toulouse Midi-Pyrénées




CÉSAR A. HIDALGO
DIANA ORGHIAN
JORDI ALBO-CANALS
FILIPA DE ALMEIDA
NATALIA MARTIN
HOW HUMANS
JUDGE
MACHINES
HOW HUMANS JUDGE MACHINES
Hidalgo et al.
HOW HUMANS JUDGE MACHINES
Hidalgo et al.

Randomized Controlled Experiments



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?

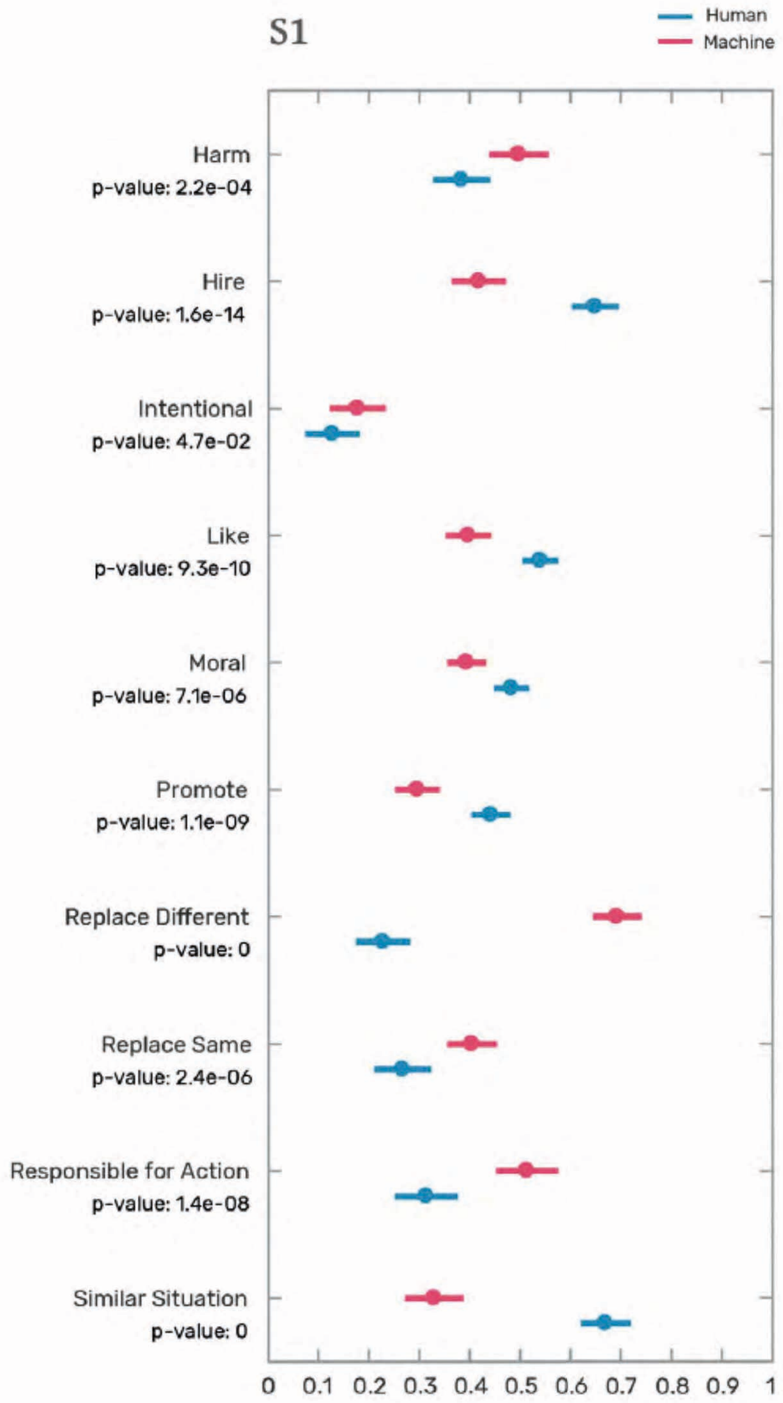
[replace different]

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

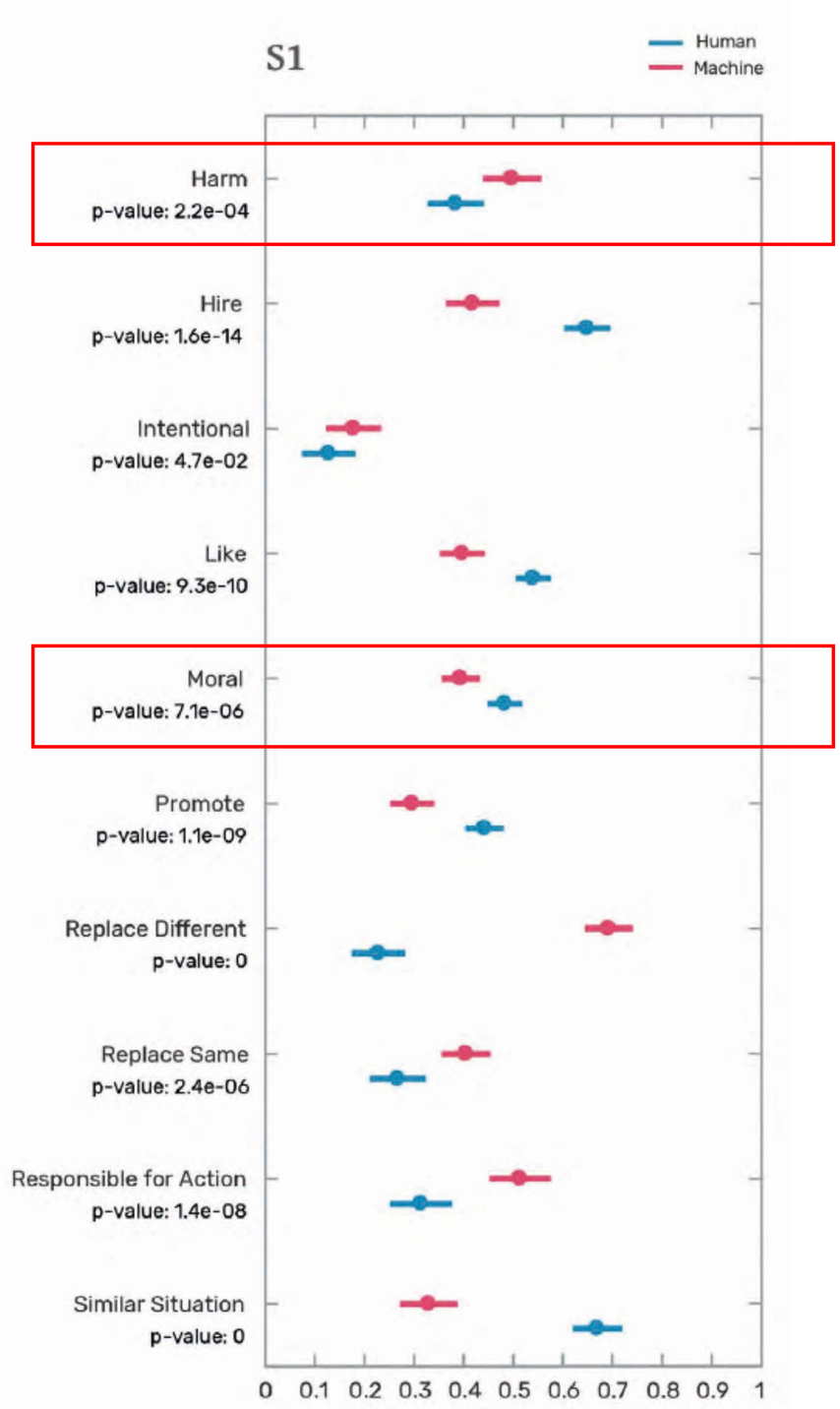
[replace same]

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

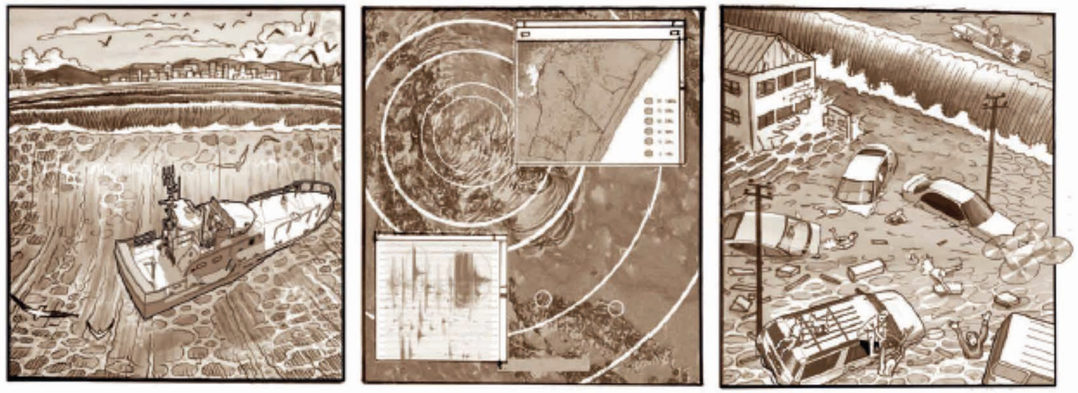
If you were in a **similar situation** as the driver, would you have done the same?

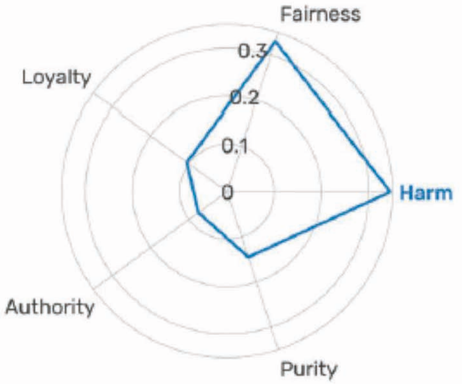

S1
Human
Machine
Harm
p-value: 2.2e-04
Hire
p-value: 1.6e-14
Intentional
p-value: 4.7e-02
Like
p-value: 9.3e-10
Moral
p-value: 7.1e-06
Promote
p-value: 1.1e-09
Replace Different
p-value: 0
Replace Same
p-value: 2.4e-06
Responsible for Action
p-value: 1.4e-08
Similar Situation
p-value: 0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1


JOHN
OOPS...


S1
Human
Machine
Harm
p-value: 2.2e-04
Hire
p-value: 1.6e-14
Intentional
p-value: 4.7e-02
Like
p-value: 9.3e-10
Moral
p-value: 7.1e-06
Promote
p-value: 1.1e-09
Replace Different
p-value: 0
Replace Same
p-value: 2.4e-06
Responsible for Action
p-value: 1.4e-08
Similar Situation
p-value: 0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Consider the following three versions of this moral dilemma:



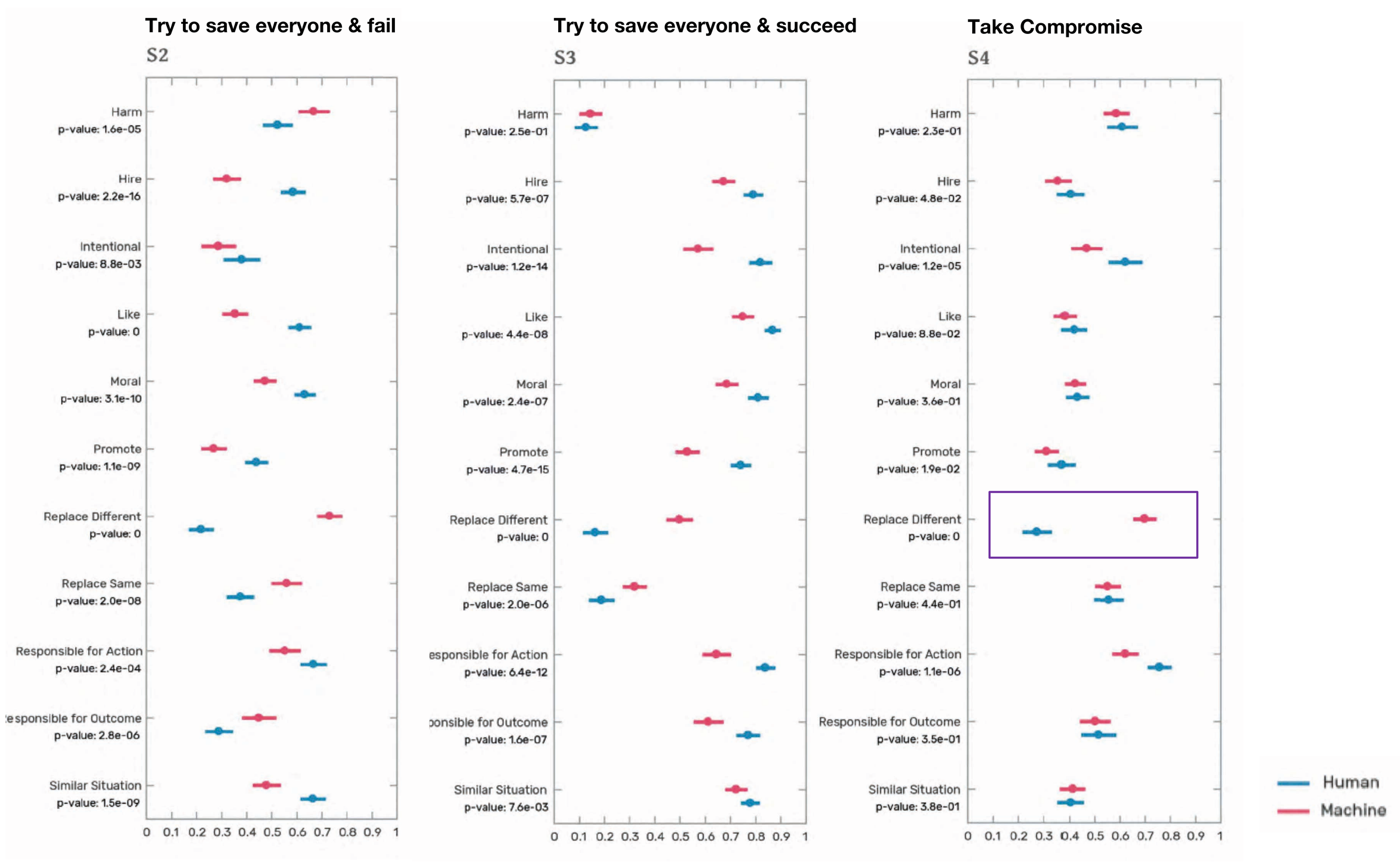


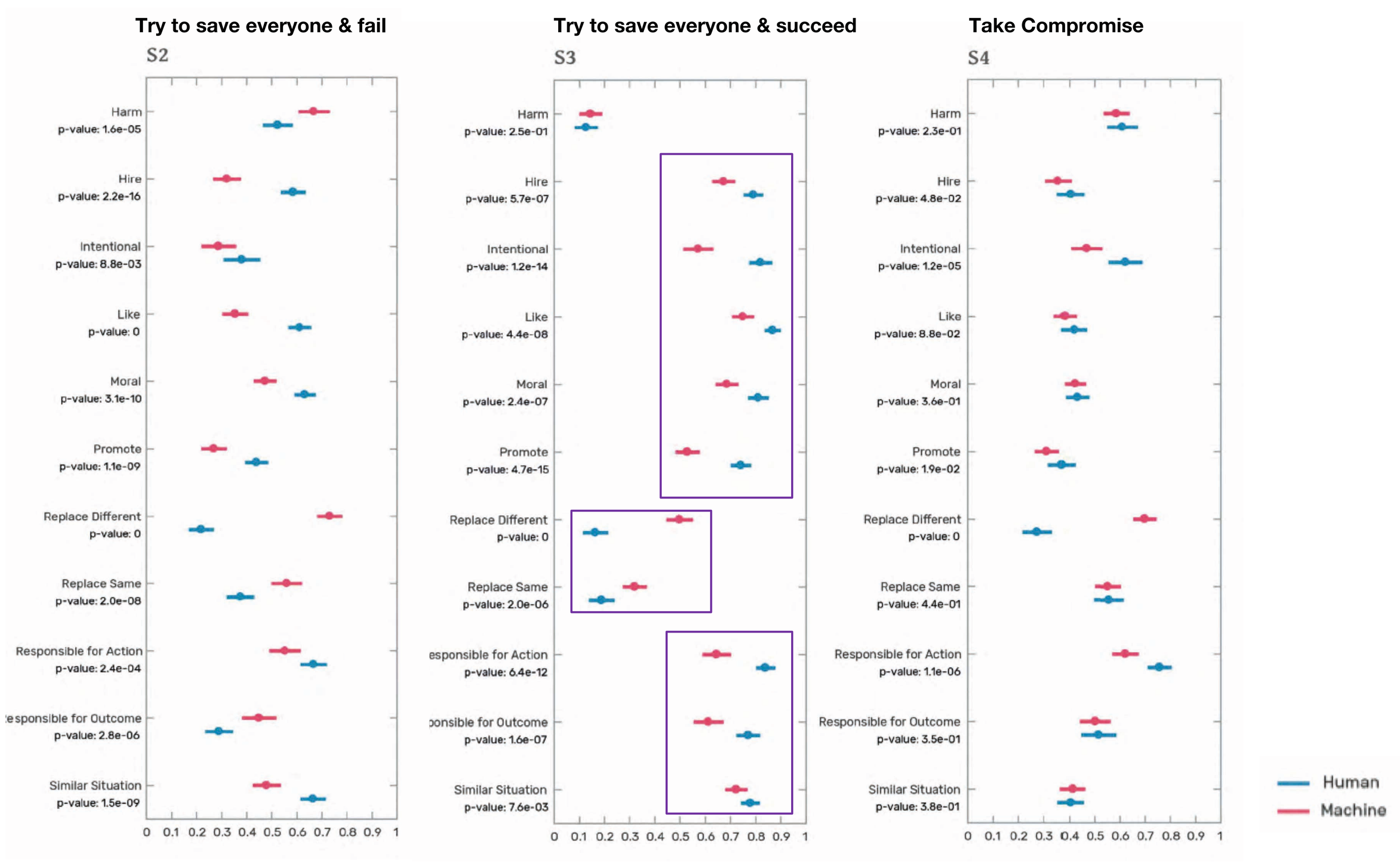
A large tsunami is approaching a coastal town of 10,000 people, with potentially devastating consequences. The [politician/algorithm] responsible for the safety of the town can decide to evacuate everyone, with a 50 percent chance of success, or save 50 percent of the town, with 100 percent success.

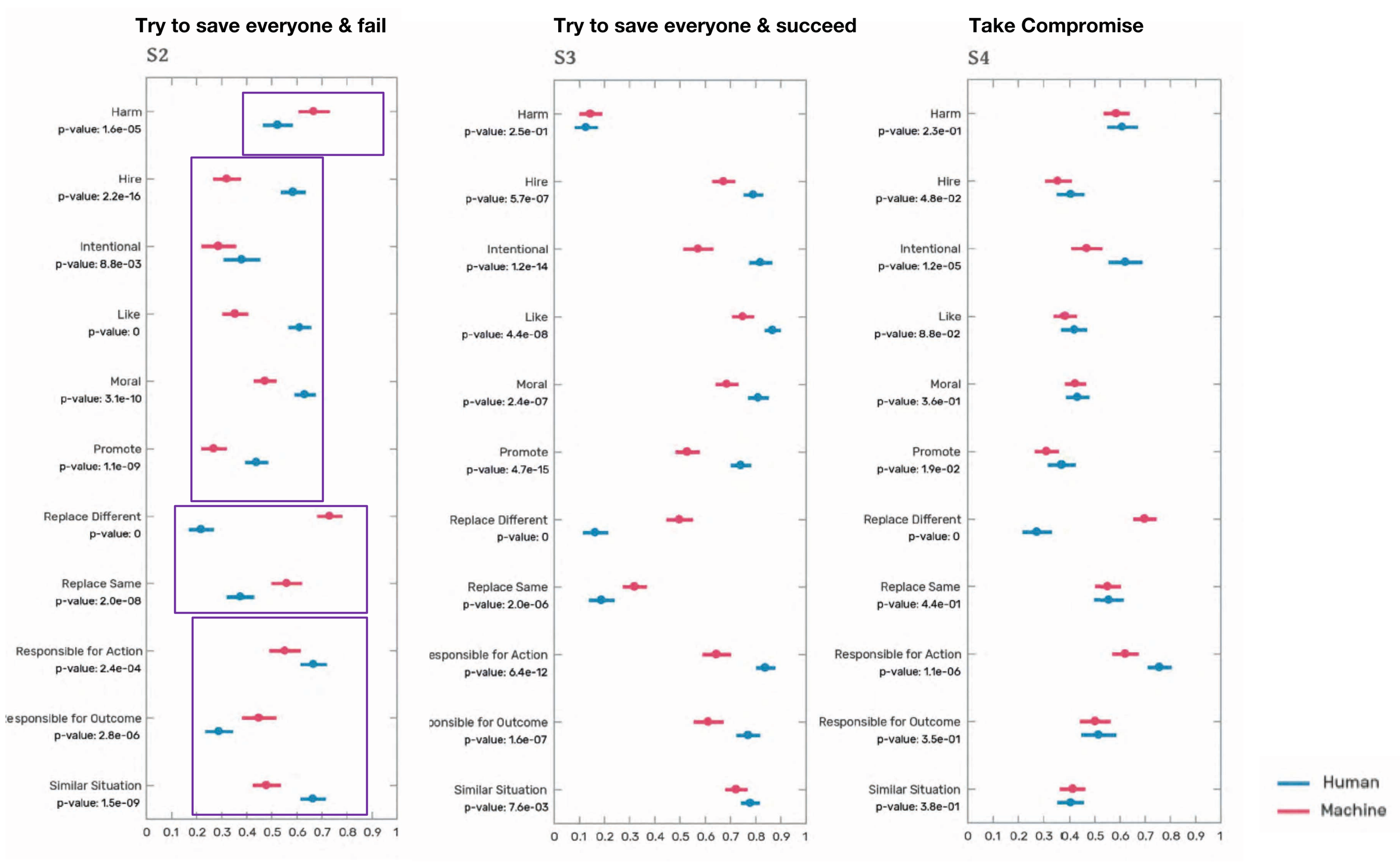
S2 The [politician/algorithm] decides to save everyone, but the rescue effort fails. The town is devastated, and a large number of people die.

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
Harm
p-value: 1.6e-05
Hire
p-value: 2.2e-16
Intentional
p-value: 8.8e-03
Like
p-value: 0
Moral
p-value: 3.1e-10
Promote
p-value: 1.1e-09
Replace Different
p-value: 0
Replace Same
p-value: 2.0e-08
Responsible for Action
p-value: 2.4e-04
esponsible for Outcome
p-value: 2.8e-06
Similar Situation
p-value: 1.5e-09
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Try to save everyone & succeed
S3
Harm
p-value: 2.5e-01
Hire
p-value: 5.7e-07
Intentional
p-value: 1.2e-14
Like
p-value: 4.4e-08
Moral
p-value: 2.4e-07
Promote
p-value: 4.7e-15
Replace Different
p-value: 0
Replace Same
p-value: 2.0e-06
esponsible for Action
p-value: 6.4e-12
onsible for Outcome
p-value: 1.6e-07
Similar Situation
p-value: 7.6e-03
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Take Compromise
S4
Harm
p-value: 2.3e-01
Hire
p-value: 4.8e-02
Intentional
p-value: 1.2e-05
Like
p-value: 8.8e-02
Moral
p-value: 3.6e-01
Promote
p-value: 1.9e-02
Replace Different
p-value: 0
Replace Same
p-value: 4.4e-01
Responsible for Action
p-value: 1.1e-06
Responsible for Outcome
p-value: 3.5e-01
Similar Situation
p-value: 3.8e-01
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Human
Machine

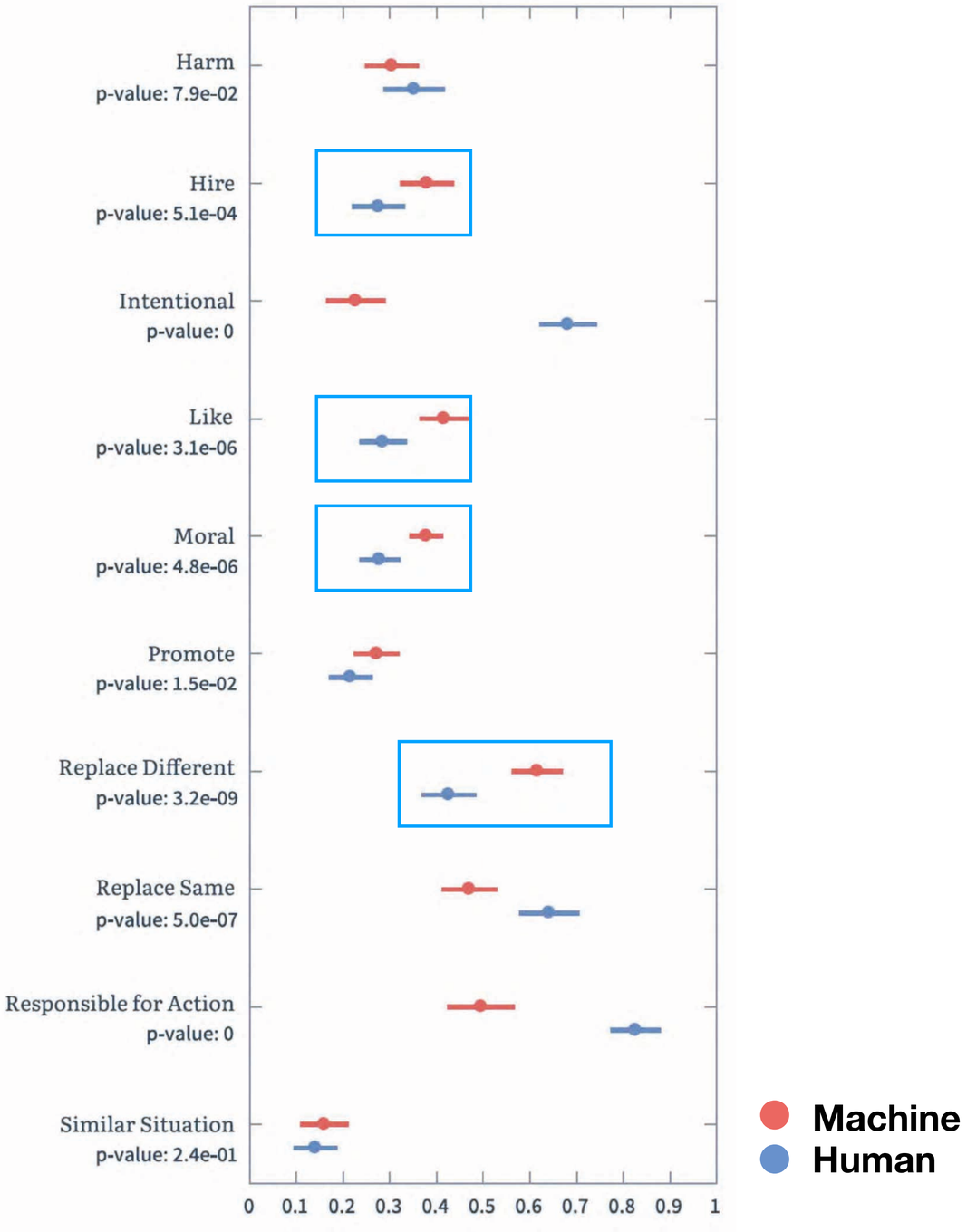

Try to save everyone & fail
S2
Harm
p-value: 1.6e-05
Hire
p-value: 2.2e-16
Intentional
p-value: 8.8e-03
Like
p-value: 0
Moral
p-value: 3.1e-10
Promote
p-value: 1.1e-09
Replace Different
p-value: 0
Replace Same
p-value: 2.0e-08
Responsible for Action
p-value: 2.4e-04
esponsible for Outcome
p-value: 2.8e-06
Similar Situation
p-value: 1.5e-09
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Try to save everyone & succeed
S3
Harm
p-value: 2.5e-01
Hire
p-value: 5.7e-07
Intentional
p-value: 1.2e-14
Like
p-value: 4.4e-08
Moral
p-value: 2.4e-07
Promote
p-value: 4.7e-15
Replace Different
p-value: 0
Replace Same
p-value: 2.0e-06
esponsible for Action
p-value: 6.4e-12
onsible for Outcome
p-value: 1.6e-07
Similar Situation
p-value: 7.6e-03
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Take Compromise
S4
Harm
p-value: 2.3e-01
Hire
p-value: 4.8e-02
Intentional
p-value: 1.2e-05
Like
p-value: 8.8e-02
Moral
p-value: 3.6e-01
Promote
p-value: 1.9e-02
Replace Different
p-value: 0
Replace Same
p-value: 4.4e-01
Responsible for Action
p-value: 1.1e-06
Responsible for Outcome
p-value: 3.5e-01
Similar Situation
p-value: 3.8e-01
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Human
Machine

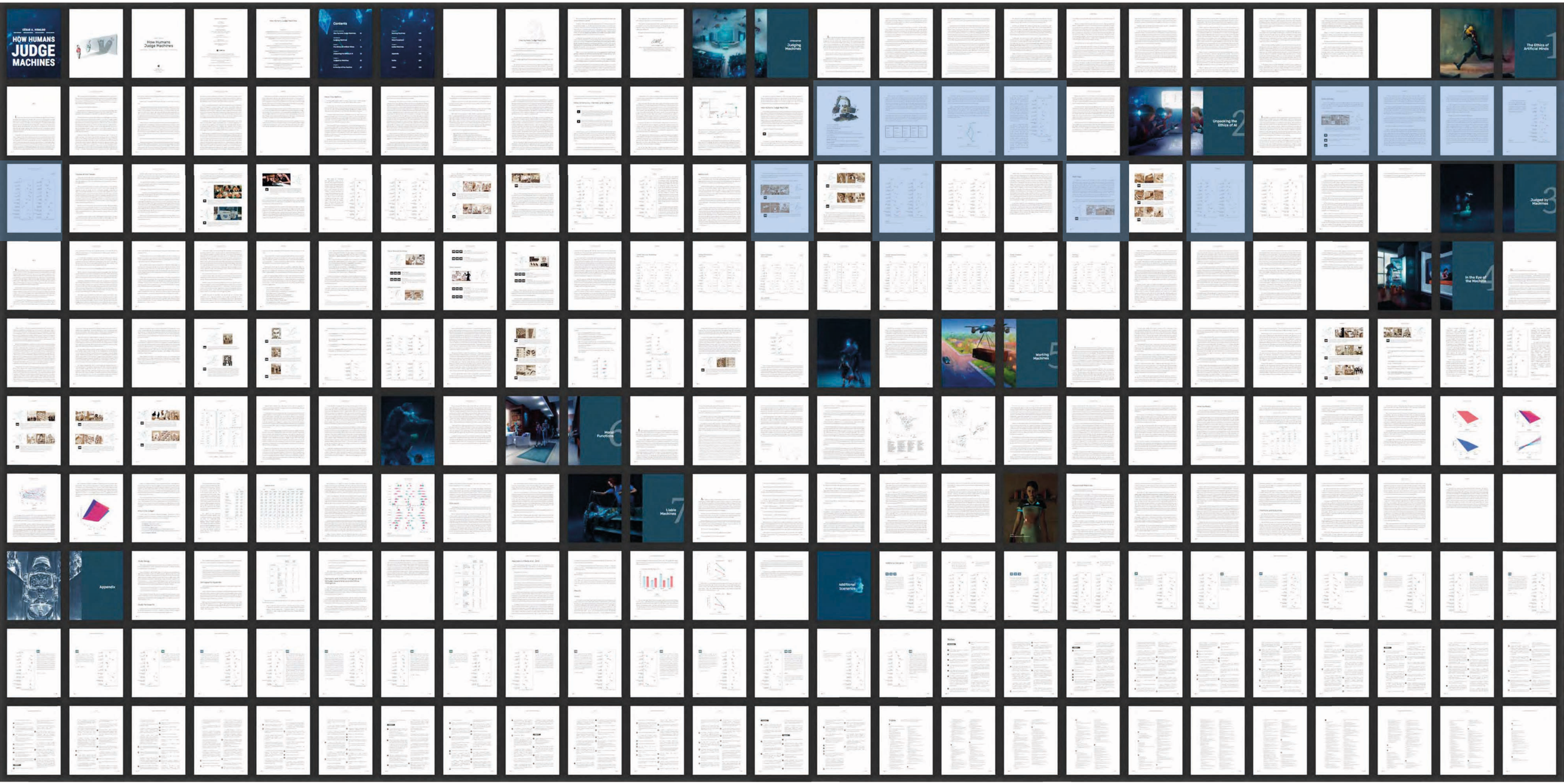

Try to save everyone & fail
S2
Harm
p-value: 1.6e-05
Hire
p-value: 2.2e-16
Intentional
p-value: 8.8e-03
Like
p-value: 0
Moral
p-value: 3.1e-10
Promote
p-value: 1.1e-09
Replace Different
p-value: 0
Replace Same
p-value: 2.0e-08
Responsible for Action
p-value: 2.4e-04
esponsible for Outcome
p-value: 2.8e-06
Similar Situation
p-value: 1.5e-09
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Try to save everyone & succeed
S3
Harm
p-value: 2.5e-01
Hire
p-value: 5.7e-07
Intentional
p-value: 1.2e-14
Like
p-value: 4.4e-08
Moral
p-value: 2.4e-07
Promote
p-value: 4.7e-15
Replace Different
p-value: 0
Replace Same
p-value: 2.0e-06
esponsible for Action
p-value: 6.4e-12
onsible for Outcome
p-value: 1.6e-07
Similar Situation
p-value: 7.6e-03
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Take Compromise
S4
Harm
p-value: 2.3e-01
Hire
p-value: 4.8e-02
Intentional
p-value: 1.2e-05
Like
p-value: 8.8e-02
Moral
p-value: 3.6e-01
Promote
p-value: 1.9e-02
Replace Different
p-value: 0
Replace Same
p-value: 4.4e-01
Responsible for Action
p-value: 1.1e-06
Responsible for Outcome
p-value: 3.5e-01
Similar Situation
p-value: 3.8e-01
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Human
Machine

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


Harm
p-value: 7.9e-02
Hire
p-value: 5.1e-04
Intentional
p-value: 0
Like
p-value: 3.1e-06
Moral
p-value: 4.8e-06
Promote
p-value: 1.5e-02
Replace Different
p-value: 3.2e-09
Replace Same
p-value: 5.0e-07
Responsible for Action
p-value: 0
Similar Situation
p-value: 2.4e-01
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Machine
Human



In this presentation

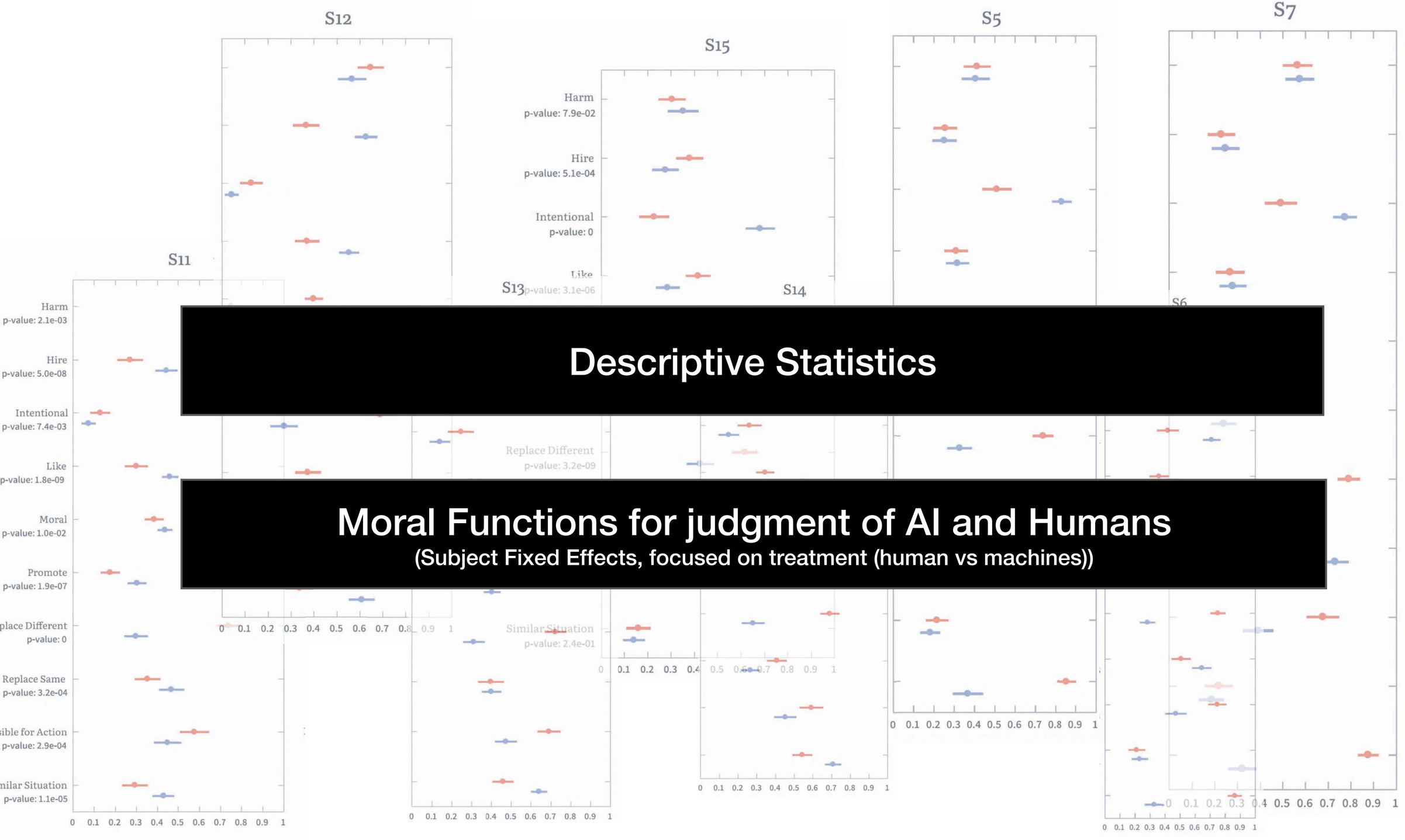
Not in this presentation

3 Judged by Machines

4 In the Eye of the Machine

5 Working Machines

6 Moral Functions



Descriptive Statistics

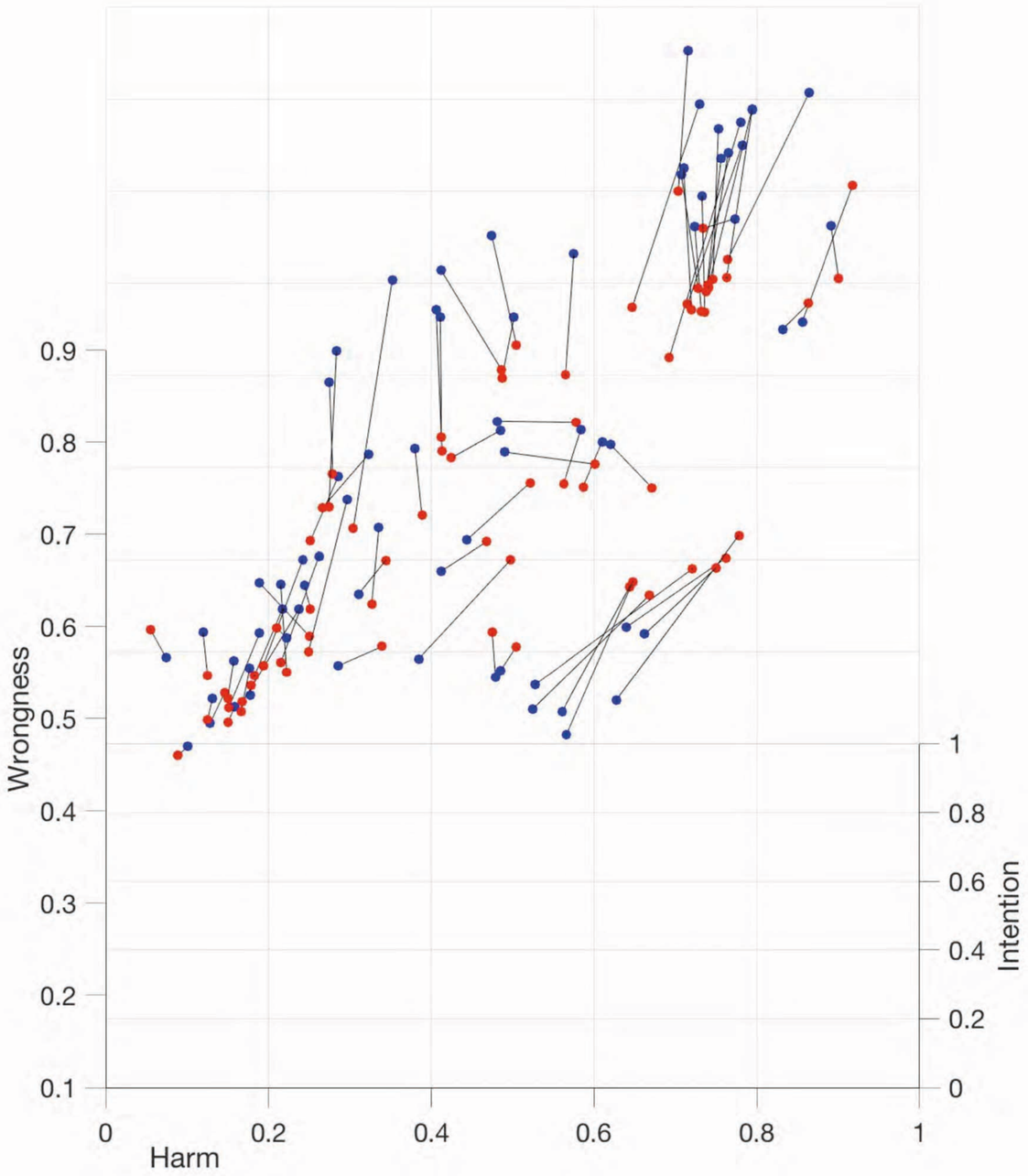
Moral Functions for judgment of AI and Humans

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

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

Descriptive Statistics

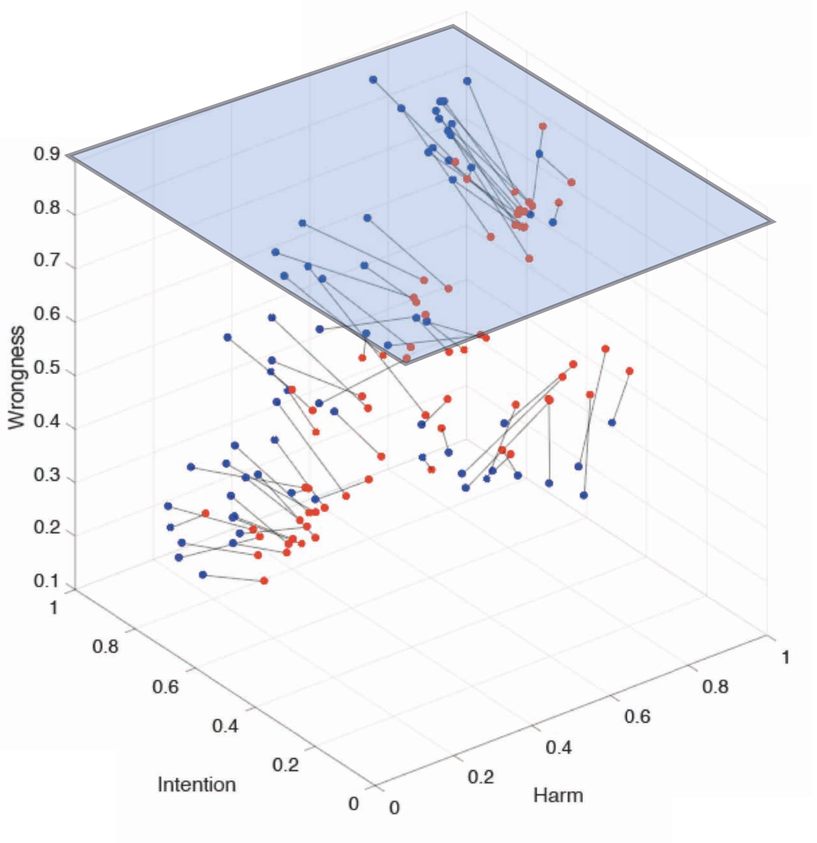
The Moral Space



Descriptive Statistics

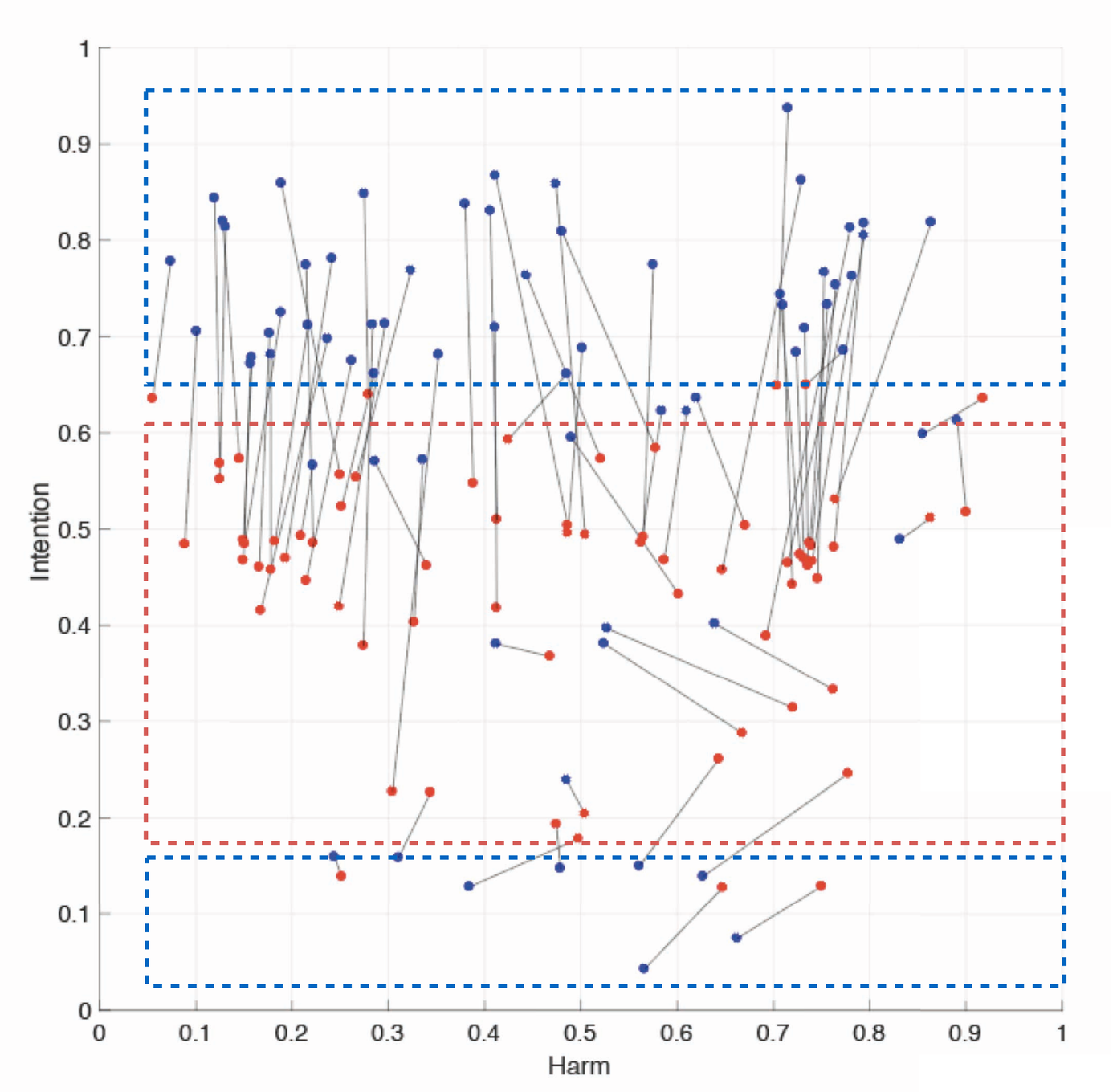
AI Human

The Moral Space

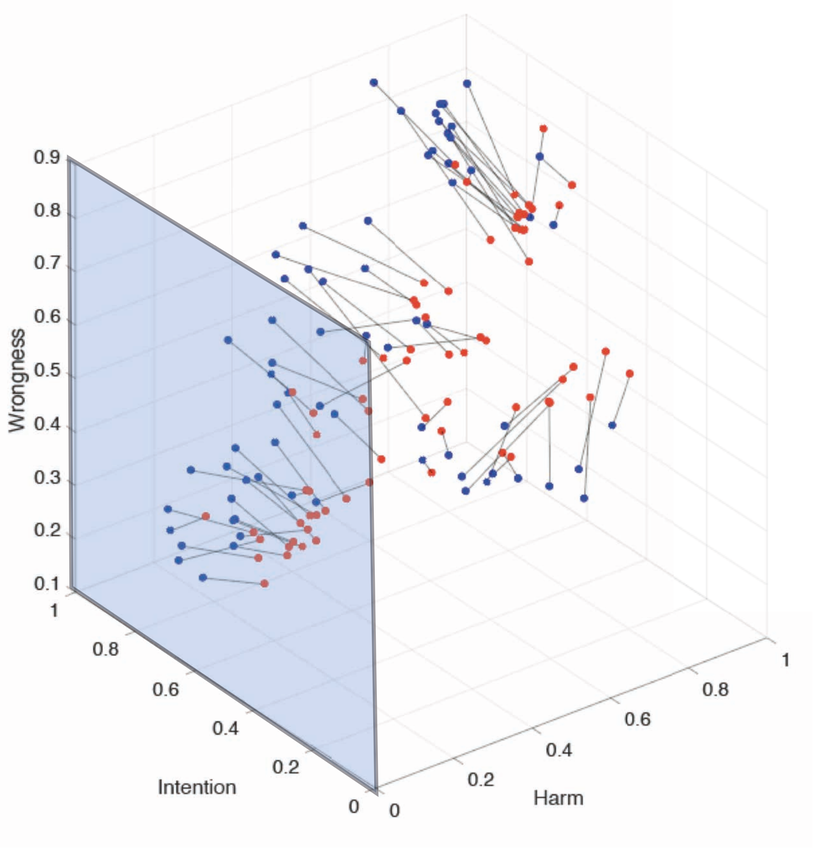


Descriptive Statistics

AI Human

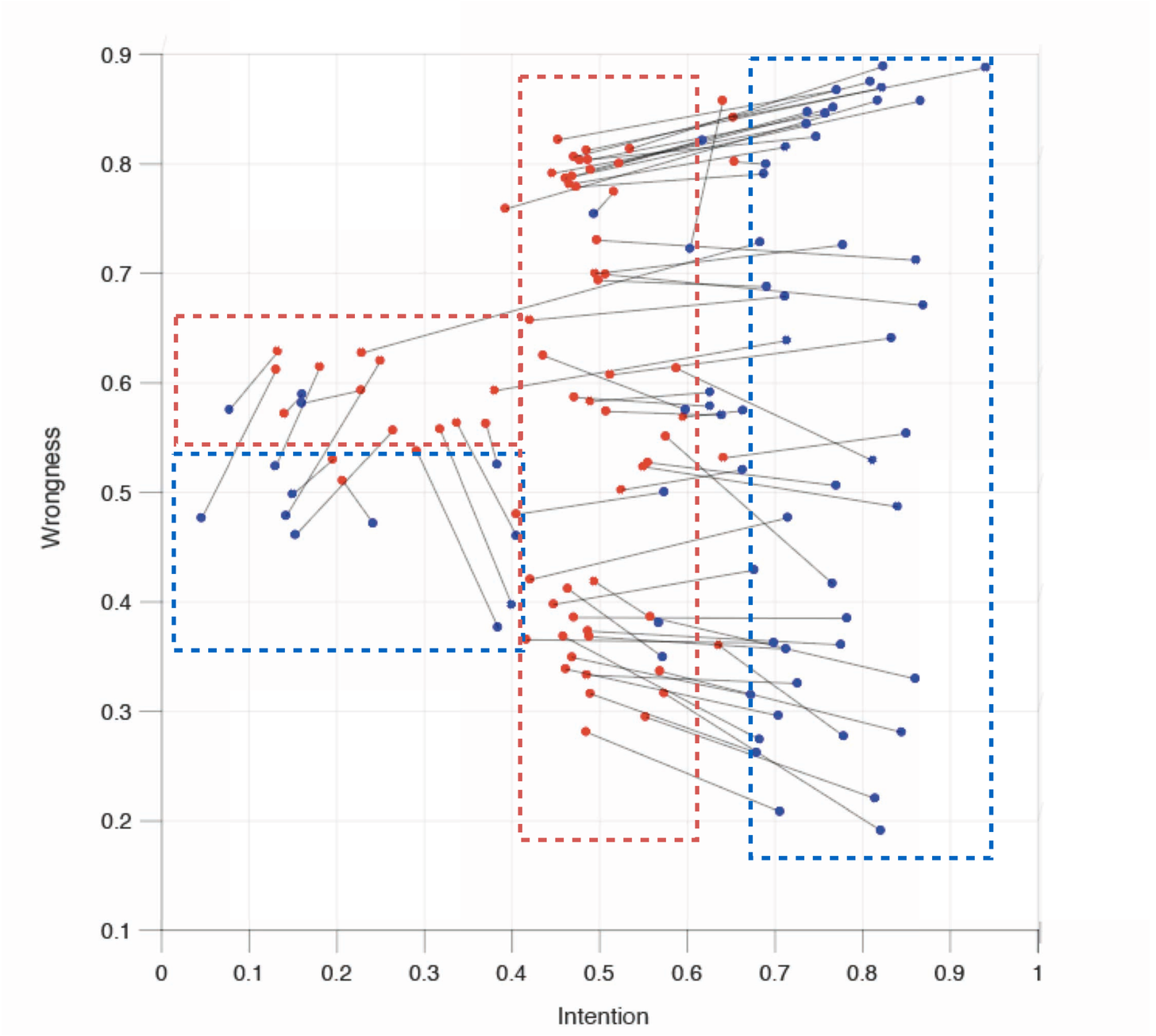


The Moral Space

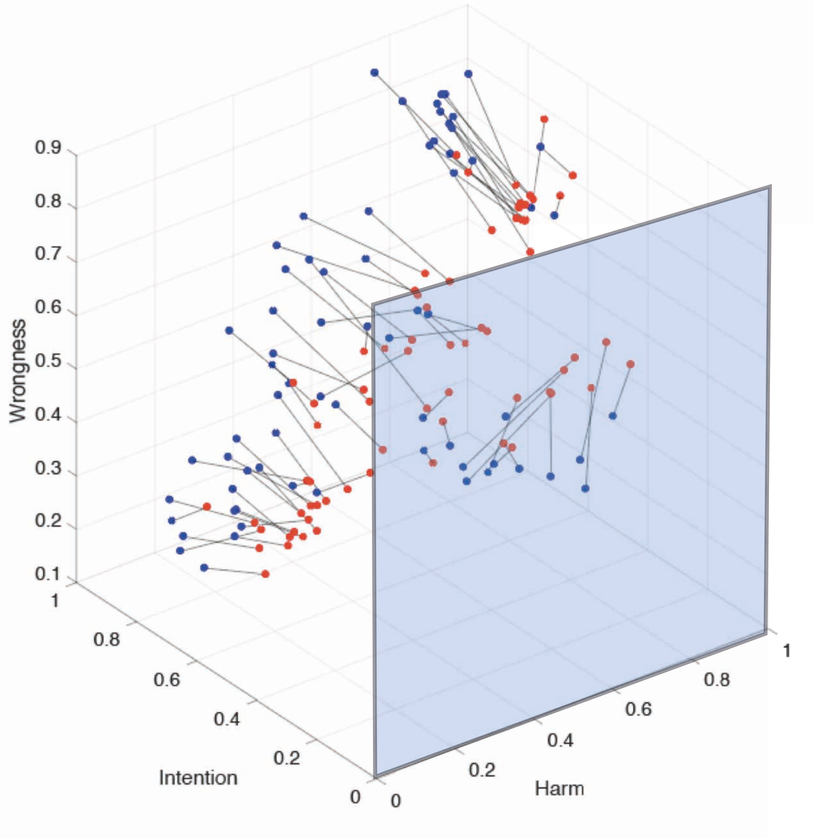


Descriptive Statistics

AI Human

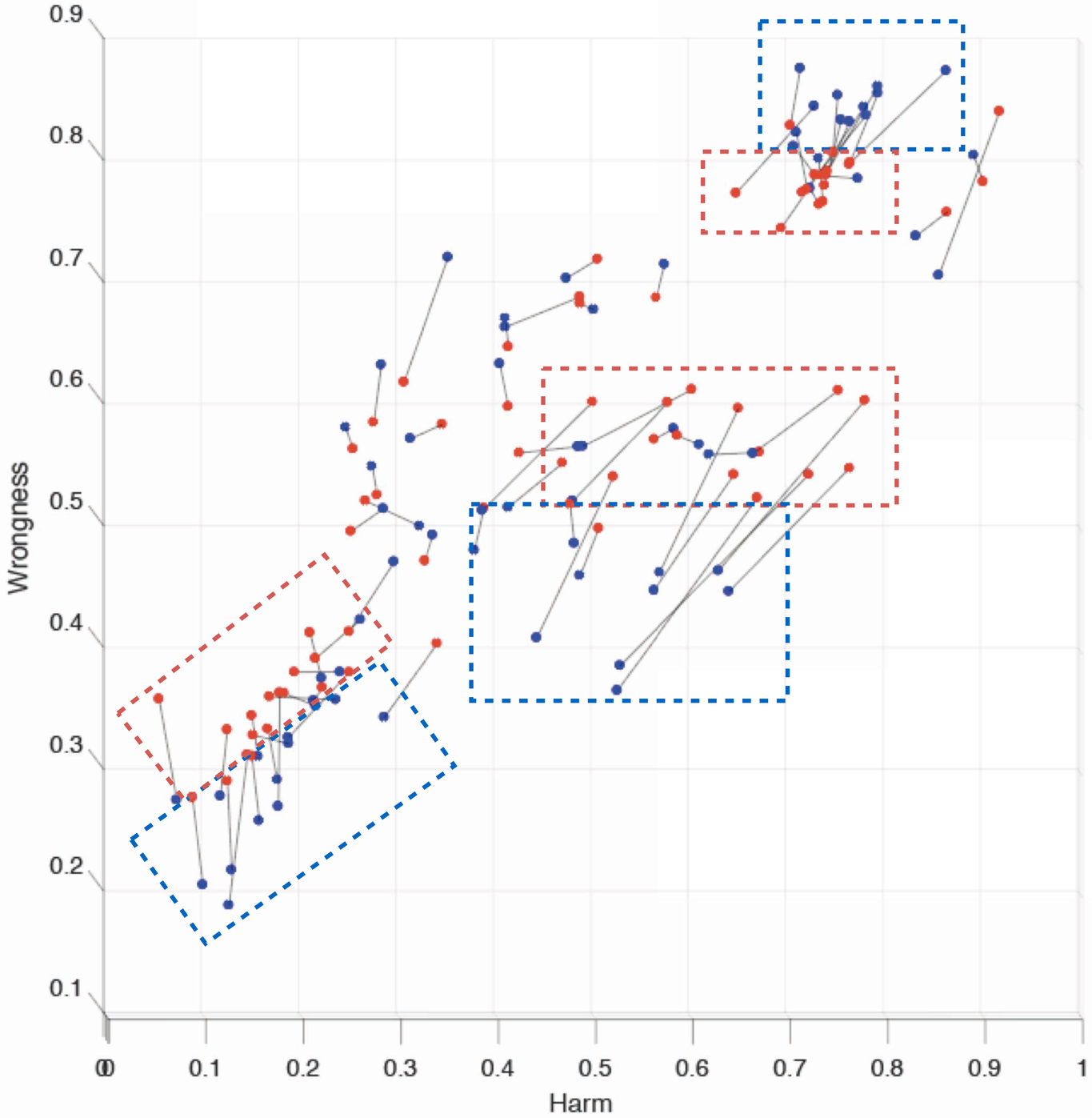


The Moral Space



Descriptive Statistics

AI Human



Moral Functions for judgment of AI and Humans

(Subject Fixed Effects)

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

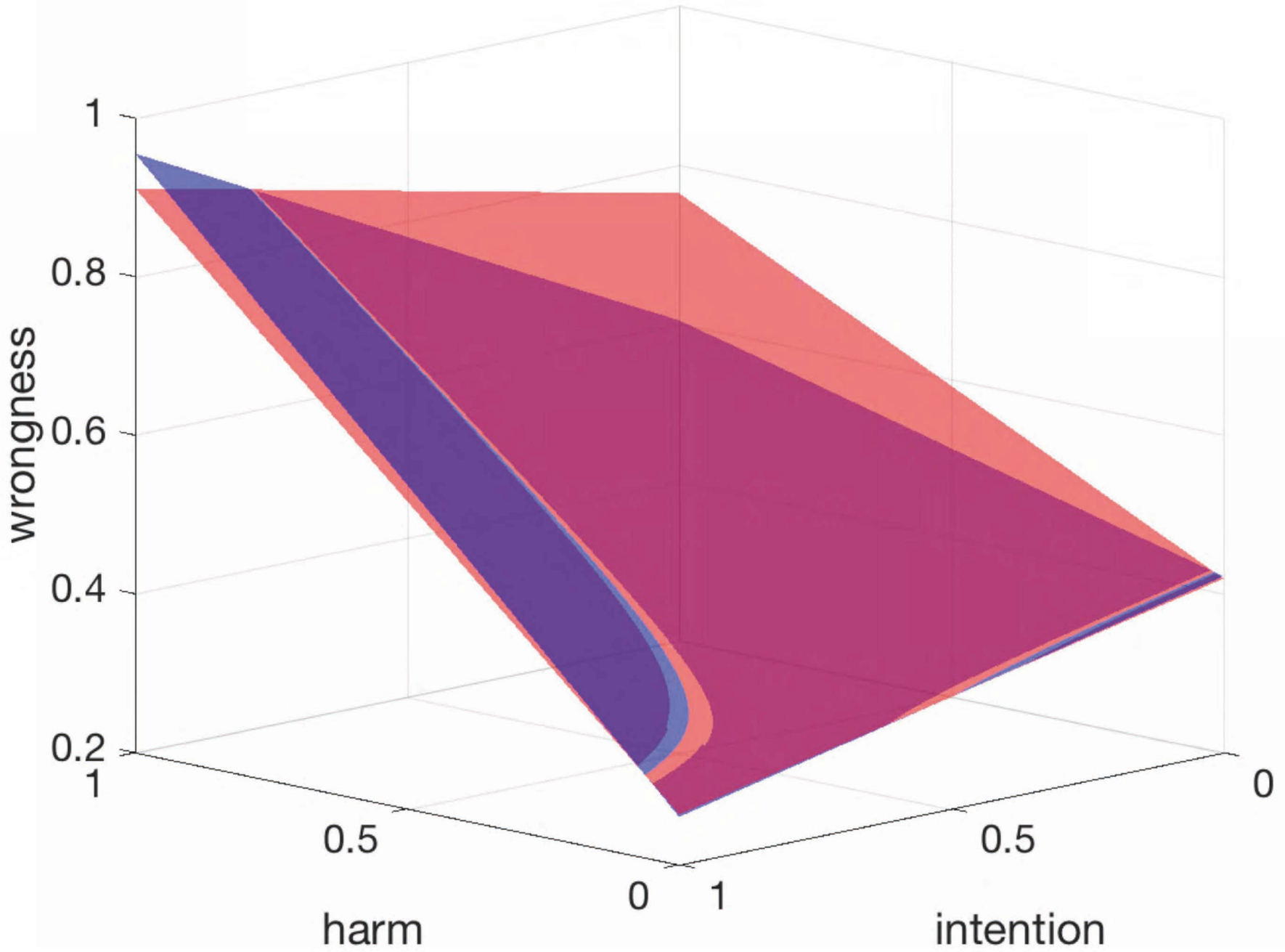
$$W = f_m(I, H)$$

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

Moral Functions

$W=f_h(I,H)$

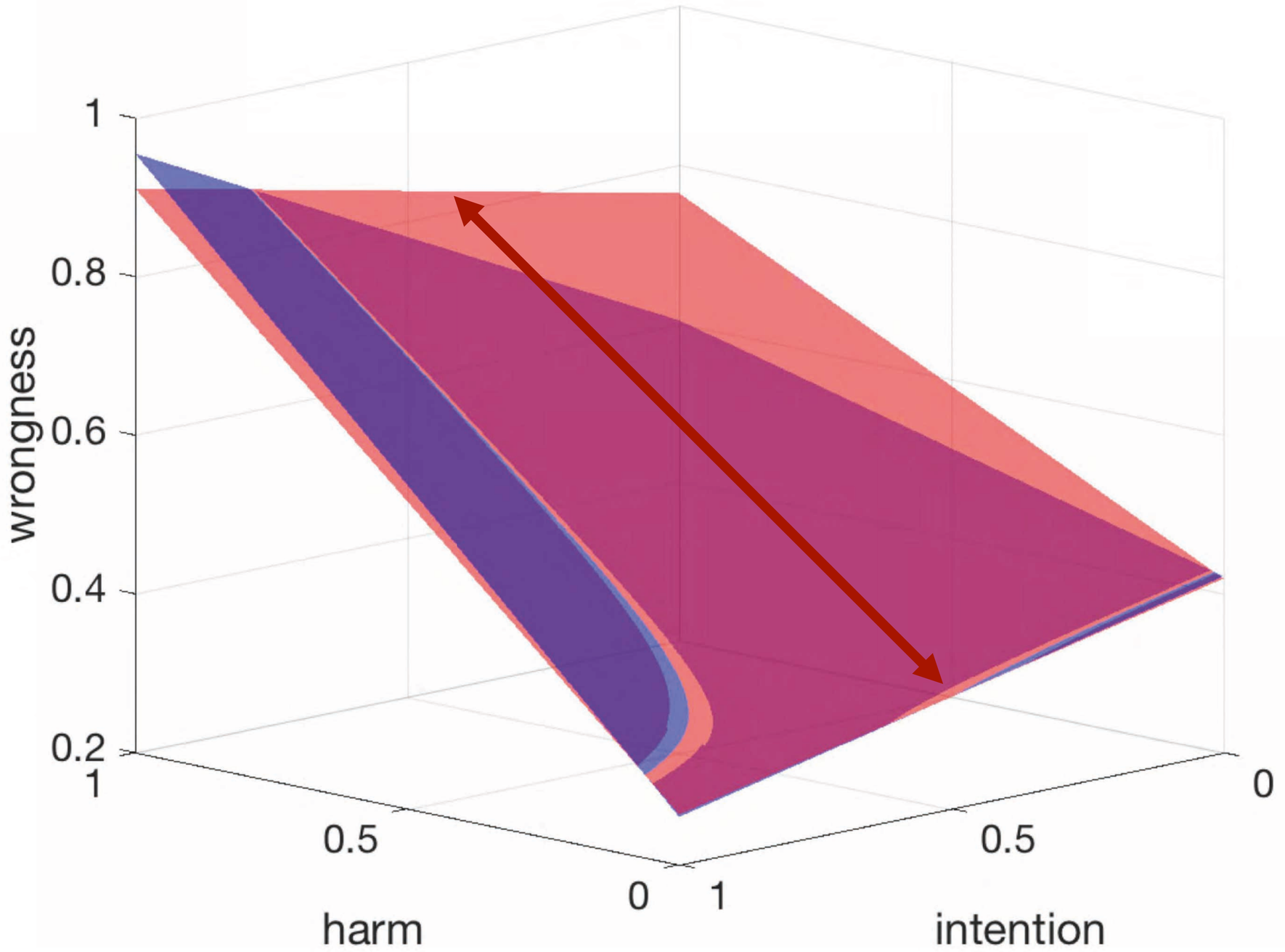
$W=f_m(I,H)$



Moral Functions

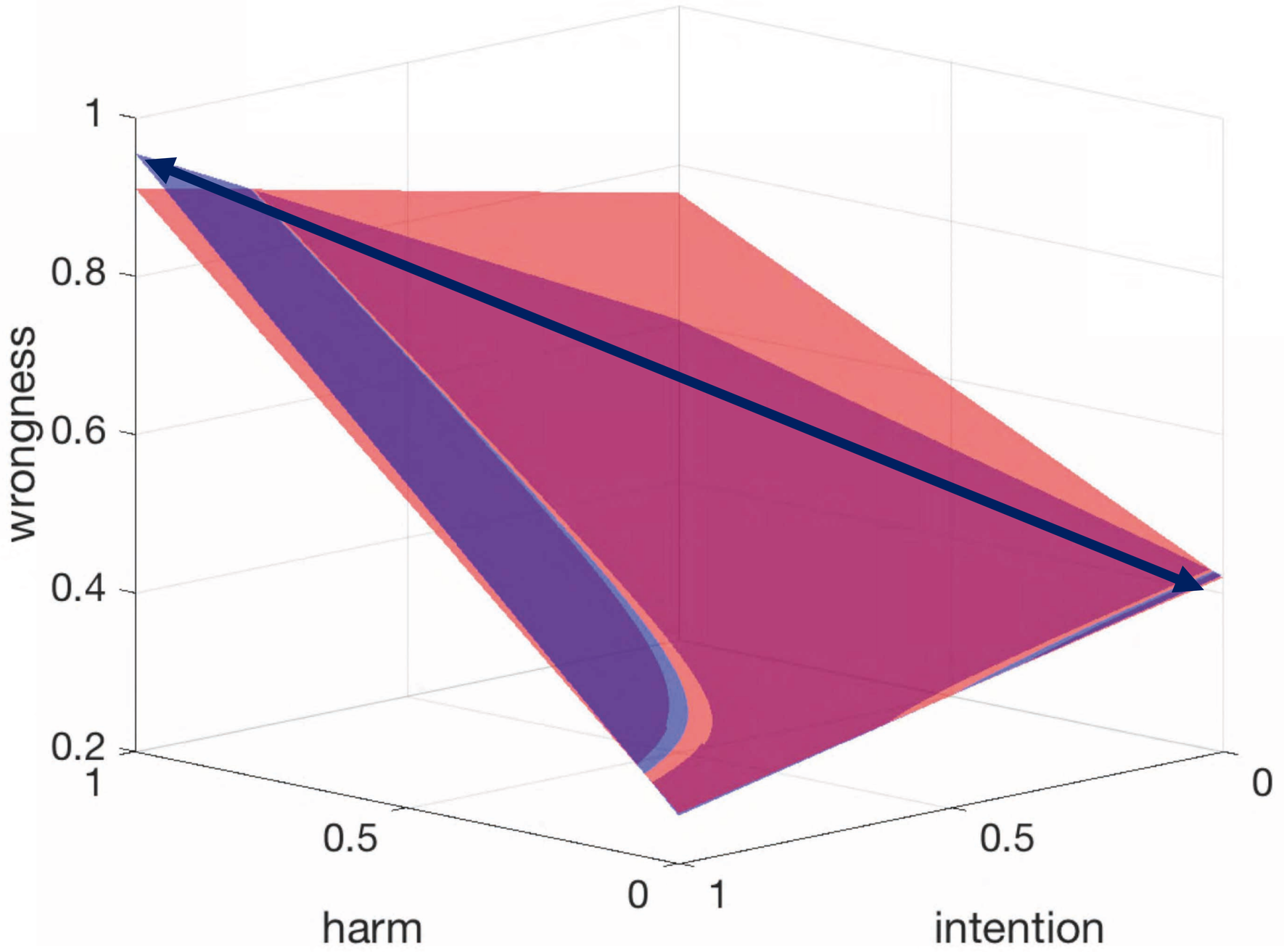
$W=f_h(I,H)$

$W=f_m(I,H)$



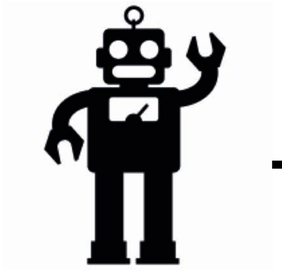
Moral Functions

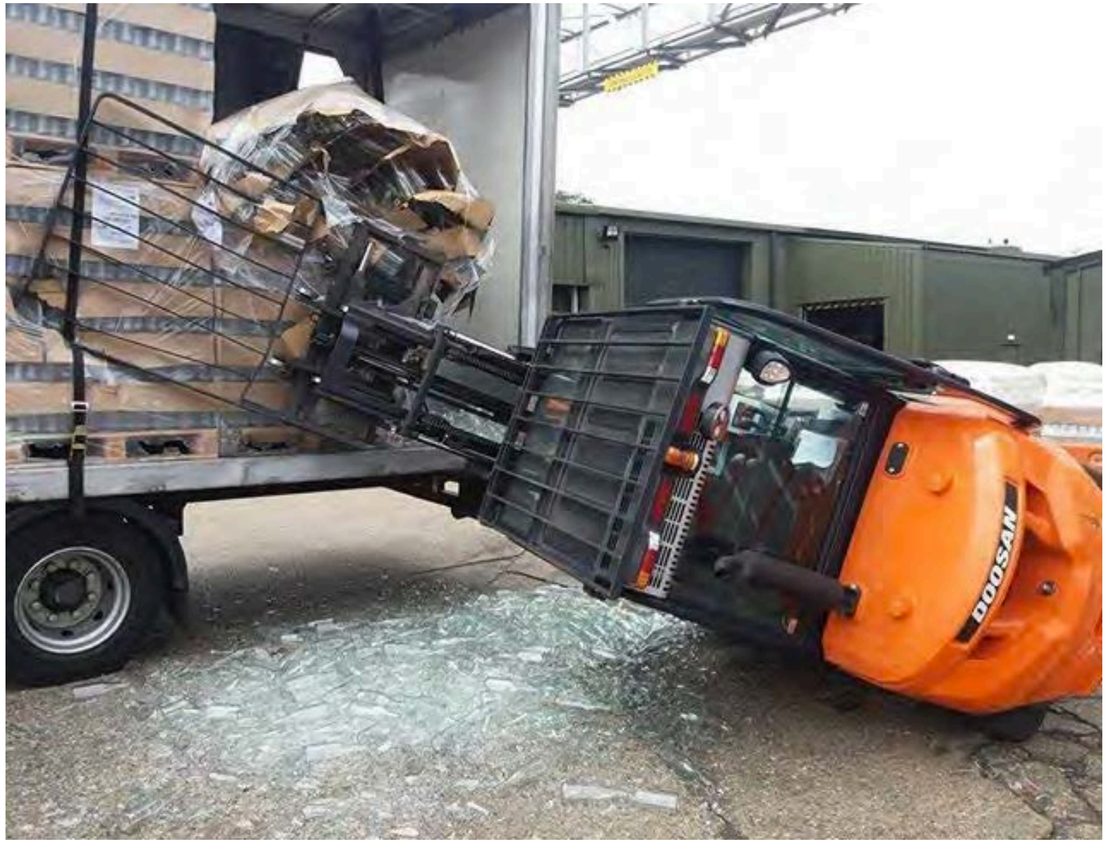
$$W=f_h(I,H)$$
$$W=f_m(I,H)$$



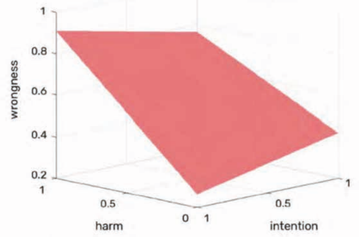
Same Mistake

Reaction



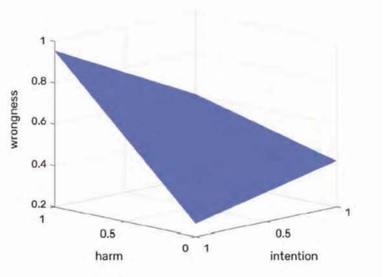


$f_h(\ldots)$



!=

$f_m(\ldots)$



How do *we* judge machines

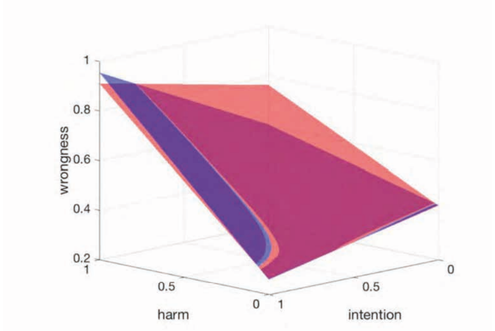
People judge humans by intentions, and machines by their outcomes

People judge human intentions bimodally, and machine actions unimodally

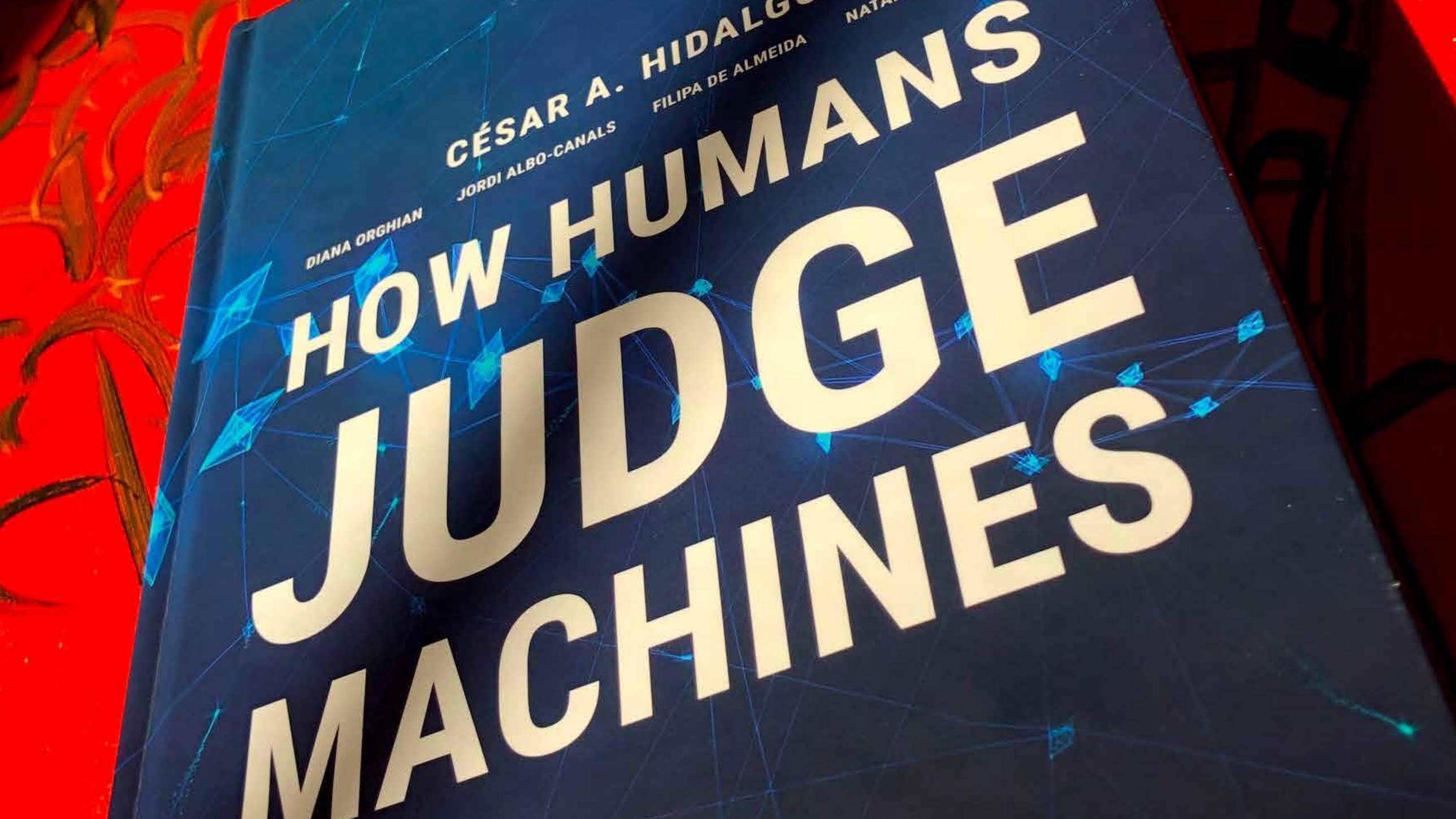
People are more forgiving of humans in accidental situations

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


CÉSAR A. HIDALGO
DIANA ORGHIAN
JORDI ALBO-CANALS
FILIPA DE ALMEIDA
HOW HUMANS
JUDGE
MACHINES

30 Episode Free Online Video Course

EPISODE 1: HOW HUMANS JUDGE MACHINES: INTRODUCTION

EPISODE 2: HOW HUMANS JUDGE MACHINES: POSITIVE AND NORMATIVE PHILOSOPHY

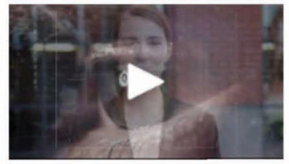
EPISODE 3: HOW HUMANS JUDGE MACHINES: MORAL STATUS AND MORAL AGENTS

EPISODE 4: HOW HUMANS JUDGE MACHINES: STRONG AND WEAK AI

EPISODE 17: HOW HUMANS JUDGE MACHINES: DISCUSSING BIAS

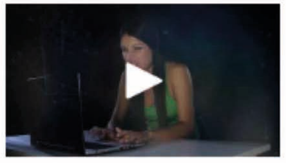
EPISODE 18: HOW HUMANS JUDGE MACHINES: PRIVACY (INTRODUCTION)

EPISODE 19: HOW HUMANS JUDGE MACHINES: DIFFERENTIAL PRIVACY



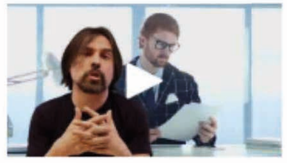
EPISODE 20: HOW HUMANS JUDGE MACHINES: PRIVACY SCENARIOS

EPISODE 5: HOW HUMANS JUDGE MACHINES: MORALITY / IMPLICIT ASSOCIATION TESTS

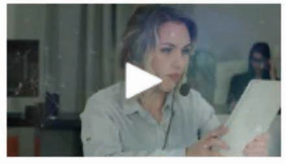


EPISODE 6: HOW HUMANS JUDGE MACHINES: MORAL DIMENSIONS

EPISODE 7: HOW HUMANS JUDGE MACHINES: INTENTION AND MORAL JUDGEMENTS



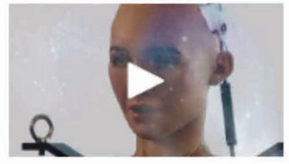
EPISODE 8: HOW HUMANS JUDGE MACHINES: DESIGN AND SAMPLE OF THE STUDY



EPISODE 21: HOW HUMANS JUDGE MACHINES: PRIVACY CONCLUSION

EPISODE 22: HOW HUMANS JUDGE MACHINES: WORKING MACHINES

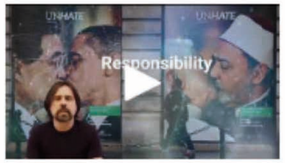
EPISODE 23: HOW HUMANS JUDGE MACHINES: LABOR SCENARIOS



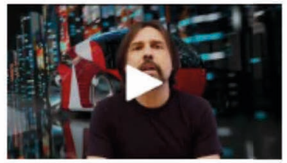
EPISODE 24: HOW HUMANS JUDGE MACHINES: LABOR CONCLUSION

EPISODE 9: HOW HUMANS JUDGE MACHINES: UNCERTAIN SITUATIONS

EPISODE 10: HOW HUMANS JUDGE MACHINES: CREATIVE TASKS



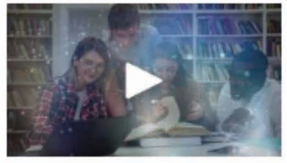
EPISODE 11: HOW HUMANS JUDGE MACHINES: RESPONSIBILITY



EPISODE 12: HOW HUMANS JUDGE MACHINES: SELF-DRIVING CARS

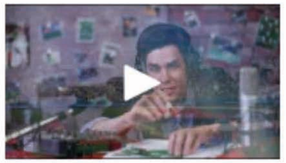
EPISODE 25: HOW HUMANS JUDGE MACHINES: THE MORAL SPACE

EPISODE 26: HOW HUMANS JUDGE MACHINES: MORAL SURFACES



EPISODE 27: HOW HUMANS JUDGE MACHINES: WHO IS THE JUDGE?

EPISODE 28: HOW HUMANS JUDGE MACHINES: LIABLE MACHINES



EPISODE 13: HOW HUMANS JUDGE MACHINES: RED FLAGS

EPISODE 14: HOW HUMANS JUDGE MACHINES: ALGORITHMIC BIAS INTRODUCTION

EPISODE 15: HOW HUMANS JUDGE MACHINES: ALGORITHMIC BIAS 2

EPISODE 16: HOW HUMANS JUDGE MACHINES: BIASED SCENARIOS

EPISODE 29: HOW HUMANS JUDGE MACHINES: RESPONSIBILITY FOR MACHINE ACTIONS

EPISODE 30: HOW HUMANS JUDGE MACHINES: INTENTIONS AND OUTCOMES

HOW HUMANS JUDGE MACHINES

MIT Press

JUDGINGMACHINES.COM