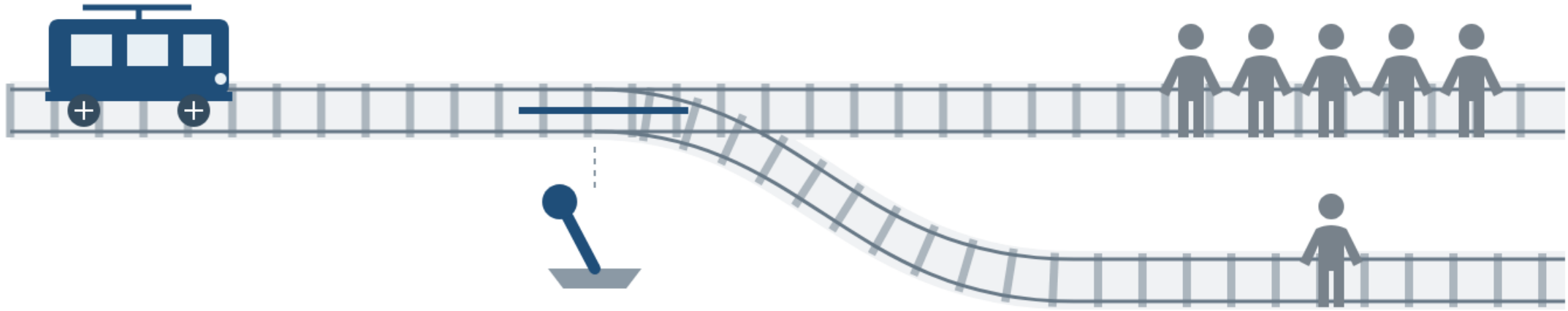


L8

AI Ethics: Values, Fairness, and Accountability



AILT1001 · Artificial Intelligence Literacy I

Three Questions for AI Ethics

Values

What should AI
prioritise?

Three Questions for AI Ethics

Values

What should AI
prioritise?

Fairness

Who benefits?
Who bears the harm?

Three Questions for AI Ethics

Values

What should AI
prioritise?

Fairness

Who benefits?
Who bears the harm?

Accountability

Who must justify
the decision?

Three Questions for AI Ethics

Values

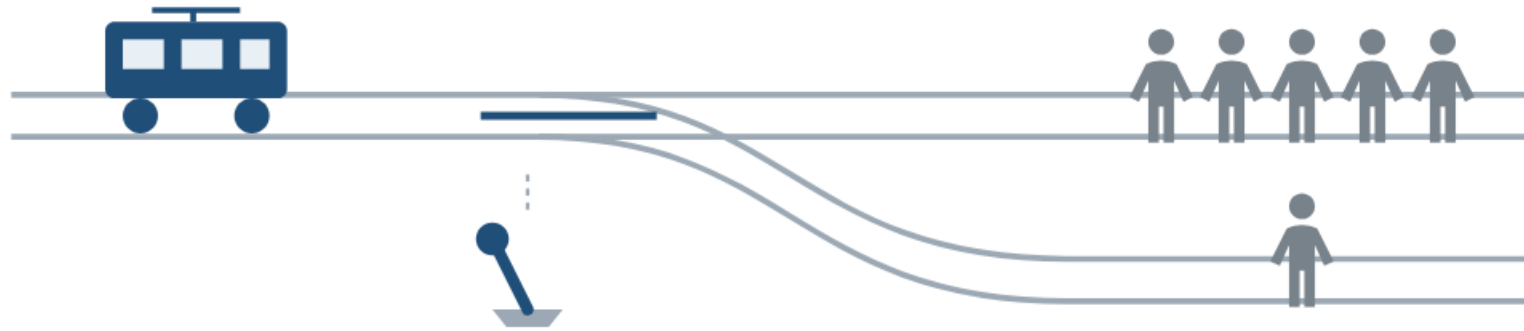
What should AI
prioritise?

Fairness

Who benefits?
Who bears the harm?

Accountability

Who must justify
the decision?



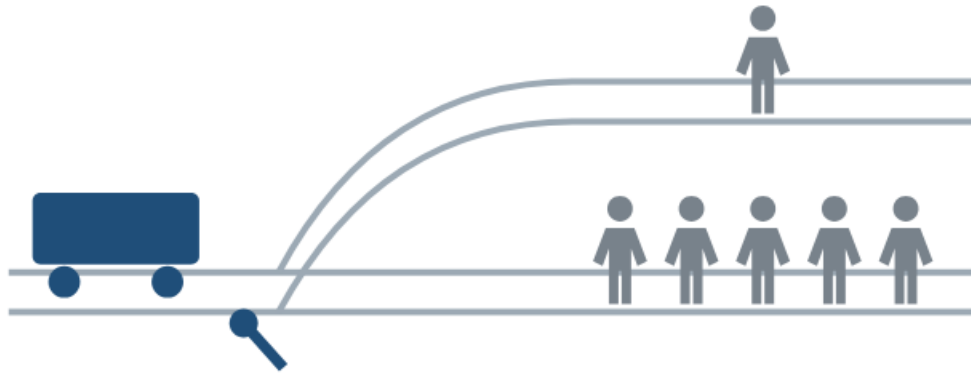
Two Trolley Decisions

Make Two Private Choices

Then tell us what mattered most

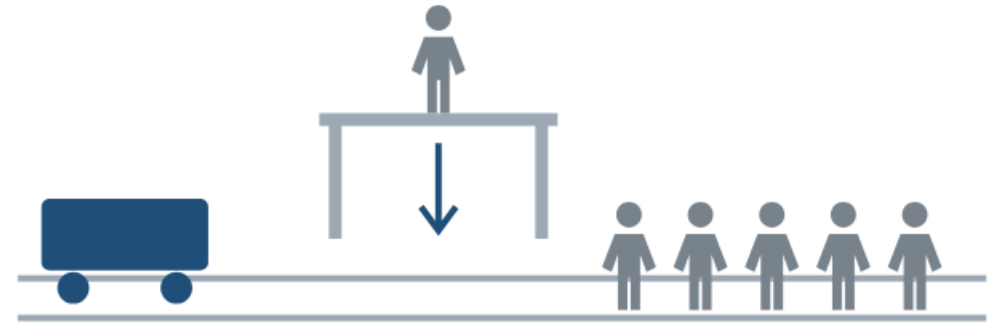
What Changed Between the Scenarios?

Switch



Pull the lever

Footbridge

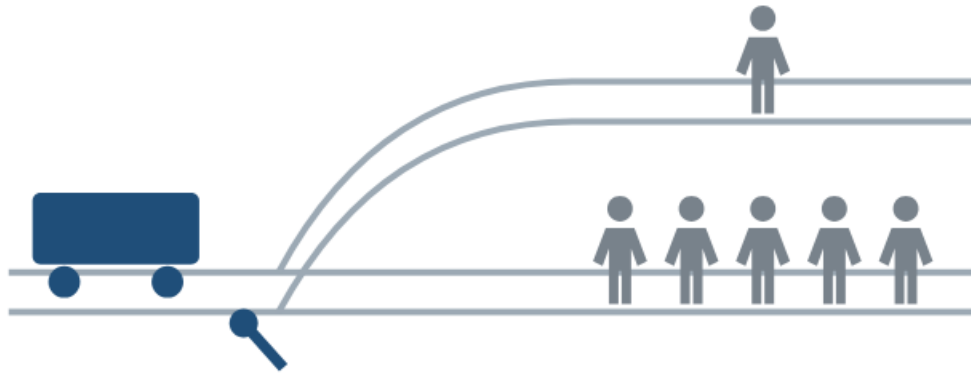


Push the person

If you intervene: five survive, one dies

What Changed Between the Scenarios?

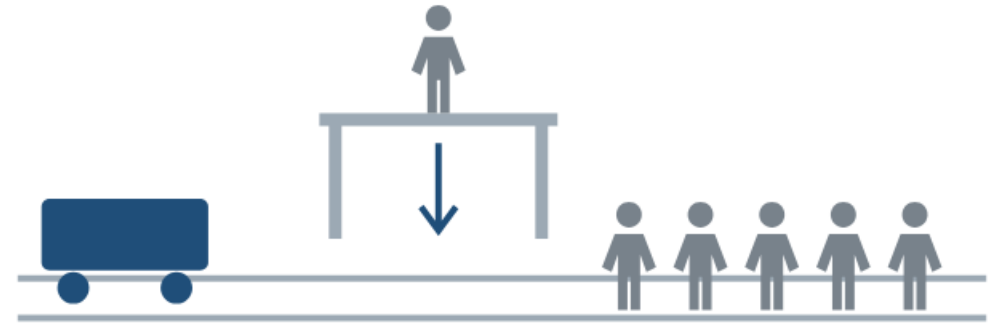
Switch



Pull the lever

Redirect a threat

Footbridge



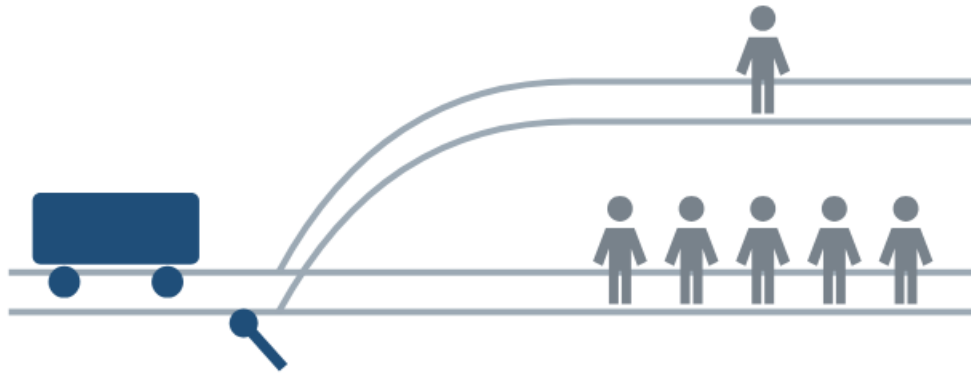
Push the person

Use a person as the means

If you intervene: five survive, one dies

What Changed Between the Scenarios?

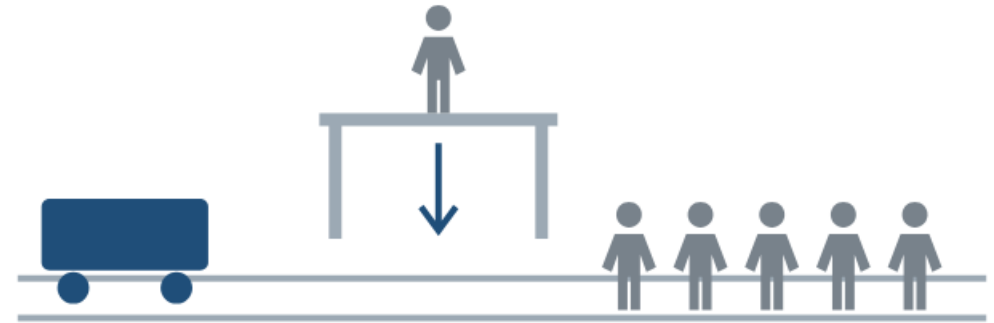
Switch



Pull the lever

Redirect a threat

Footbridge



Push the person

Use a person as the means

If you intervene: five survive, one dies

Saving more lives matters. How we save them matters too.

Reasons Behind the Choice

“We should save the most lives”

Consequences

Reasons Behind the Choice

“We should save the most lives”

Consequences

“We have duties to help and not to harm”

Duties

Reasons Behind the Choice

“We should save the most lives”

Consequences

“We have duties to help and not to harm”

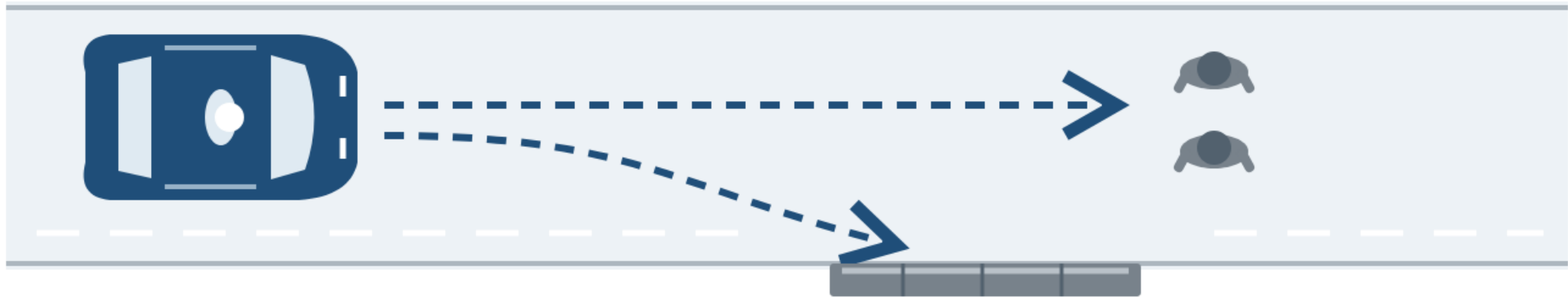
Duties

“The person being sacrificed has rights”

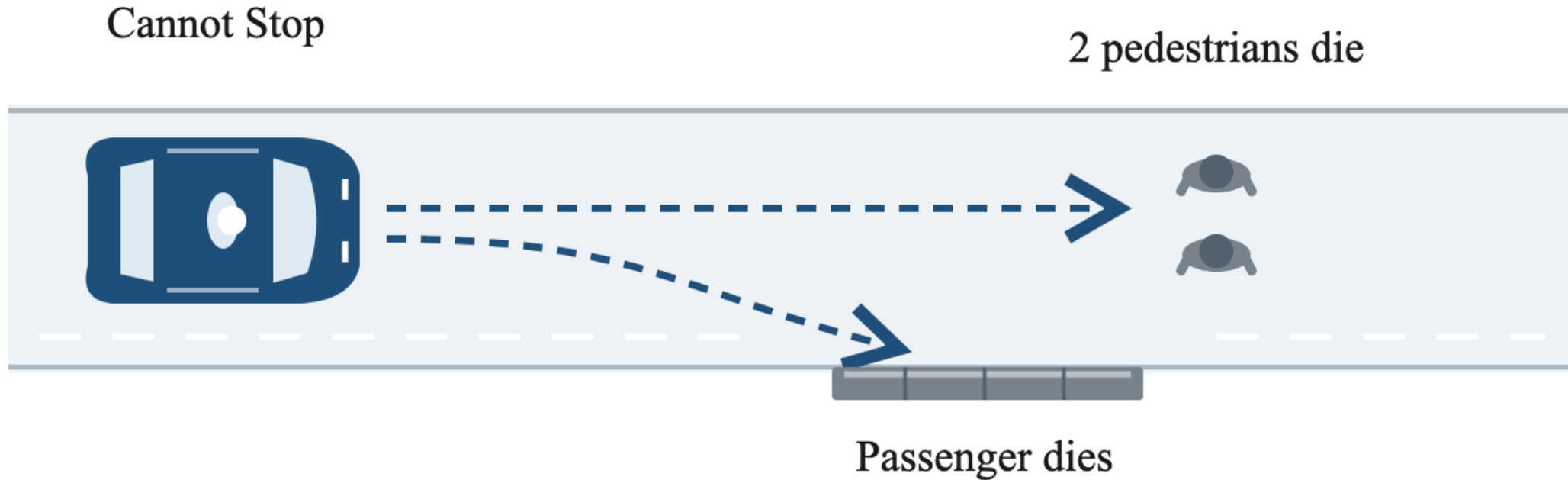
Rights

Who Should the Car Protect?

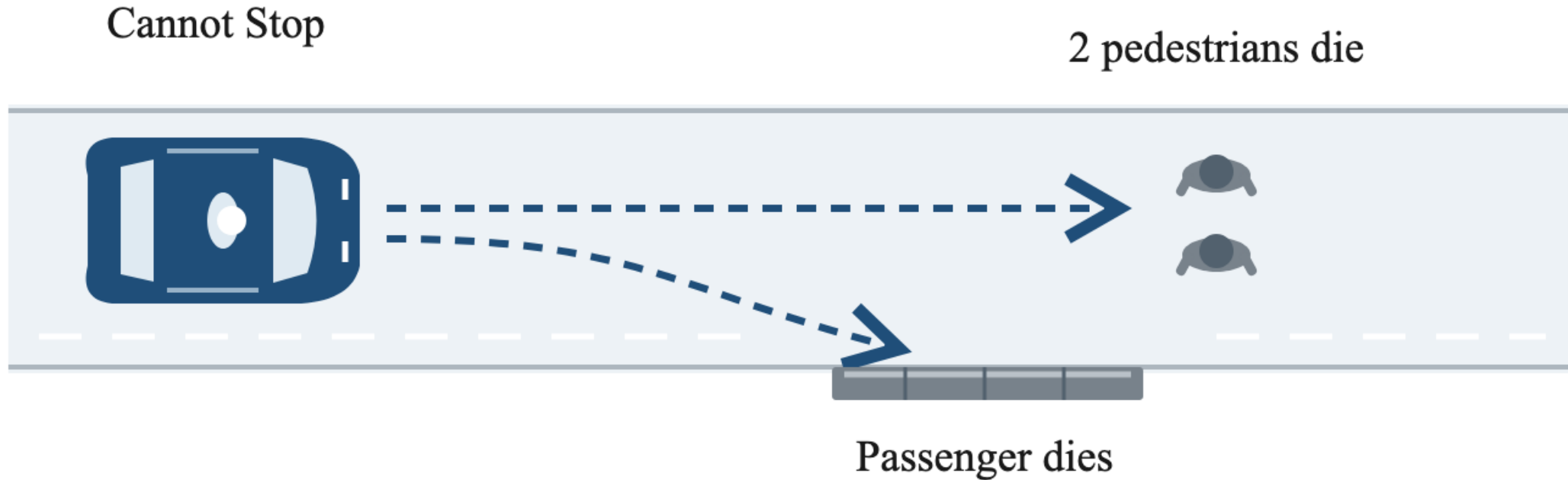
Cannot Stop



Who Should the Car Protect?



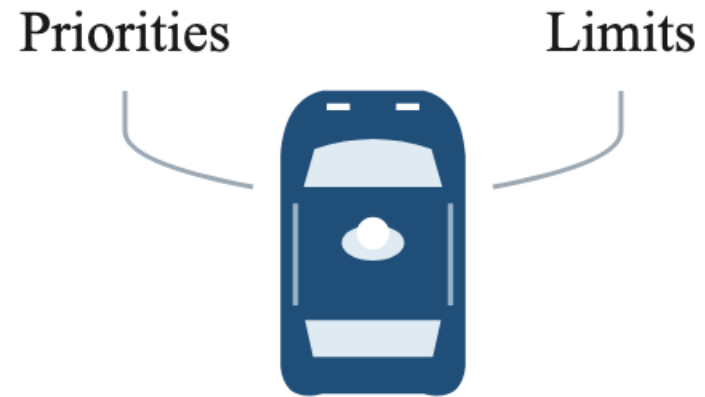
Who Should the Car Protect?



Which Car Would You Buy?

Who Sets the Rule?

Before the Journey



Who Sets the Rule?

Before the Journey

Priorities

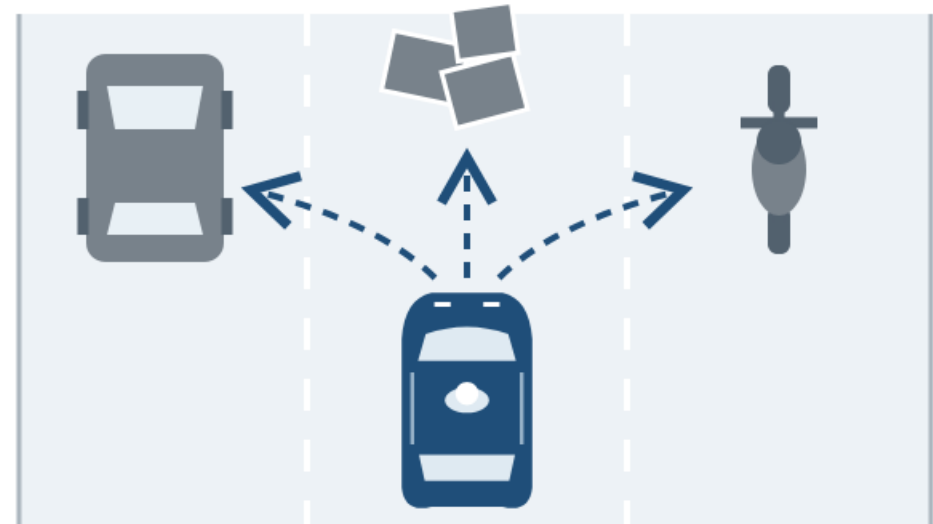
Limits



When Danger Appears

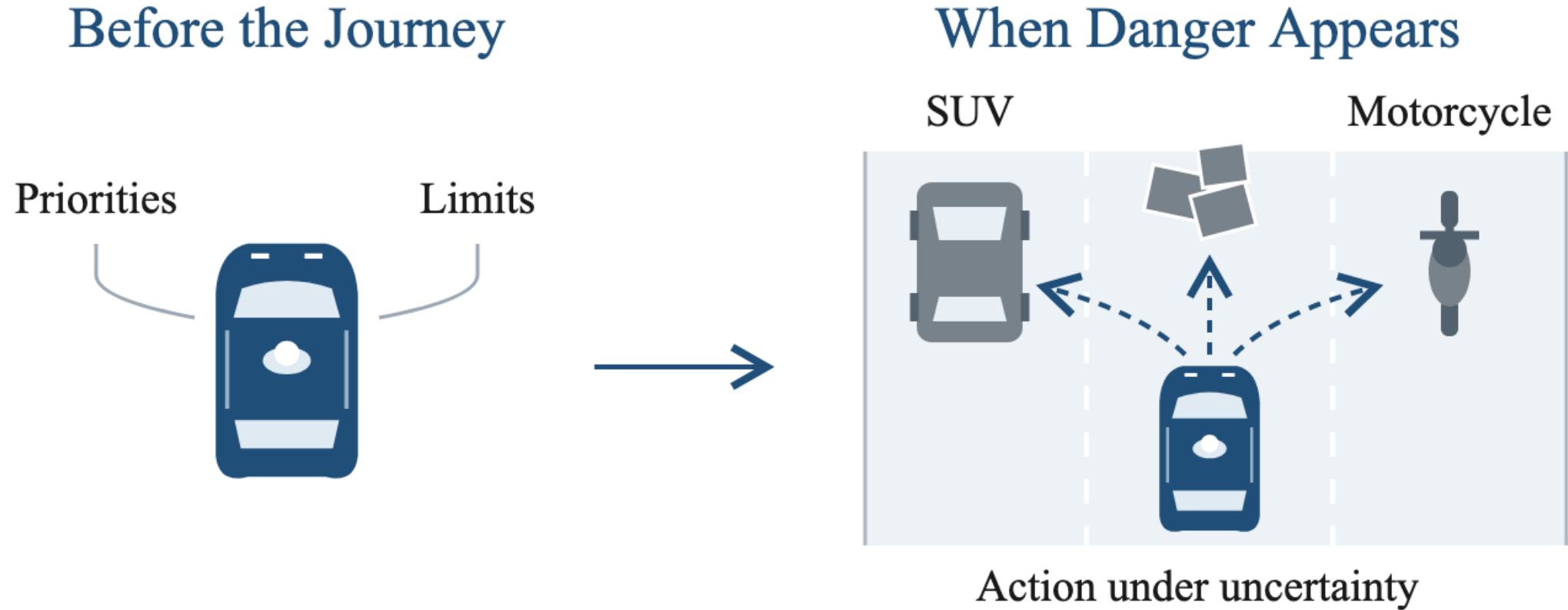
SUV

Motorcycle



Action under uncertainty

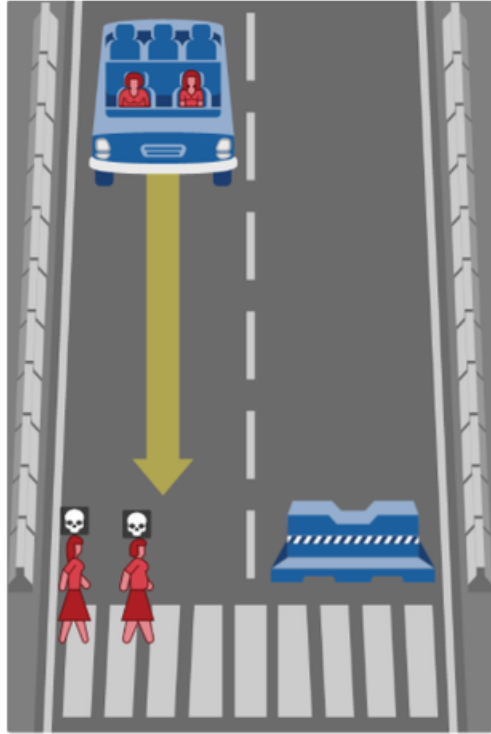
Who Sets the Rule?



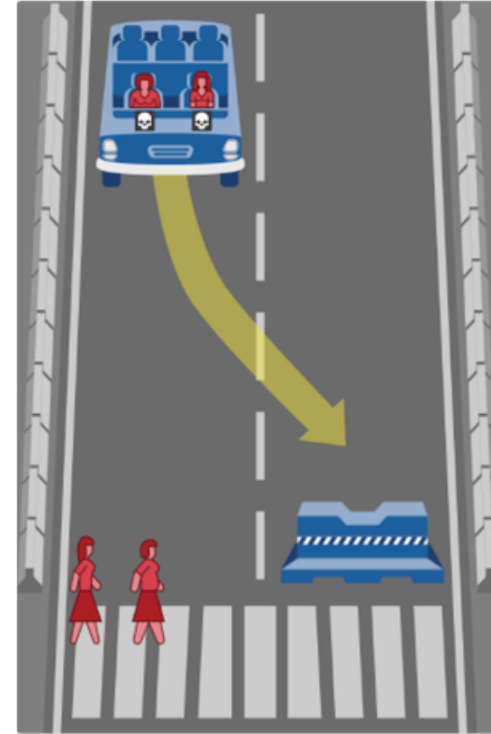
Who Must Justify These Choices?

Moral Machine

The brakes have failed



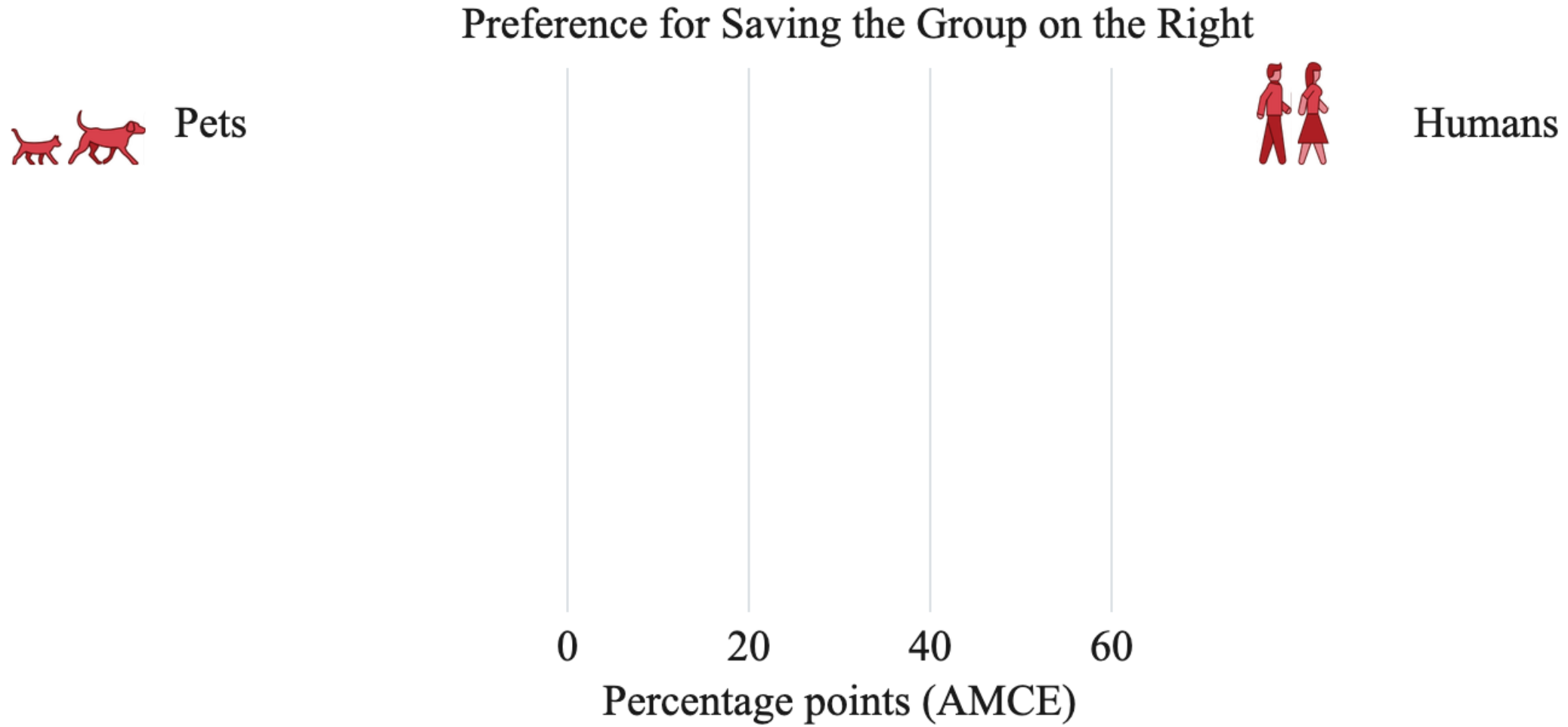
2 pedestrians die



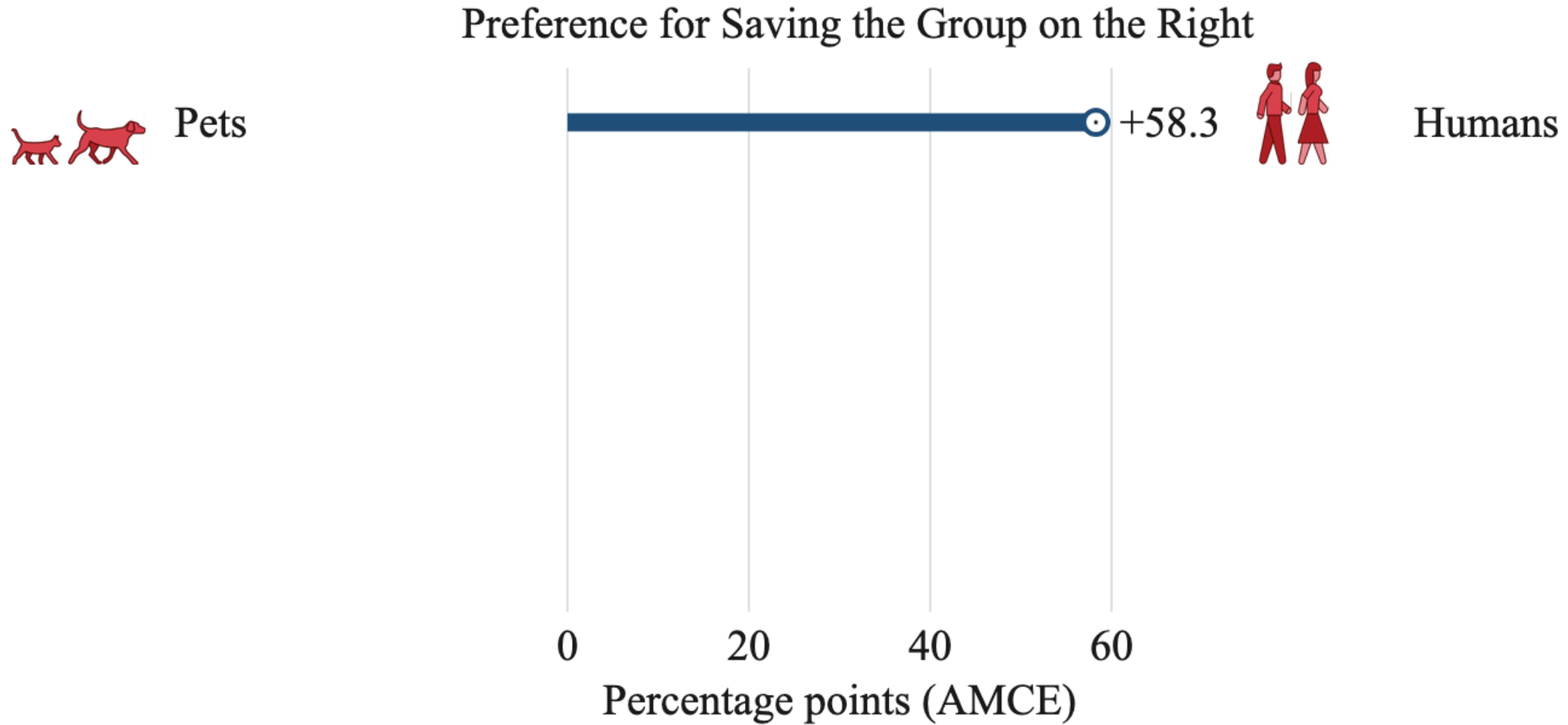
2 passengers die

moralmachine.net

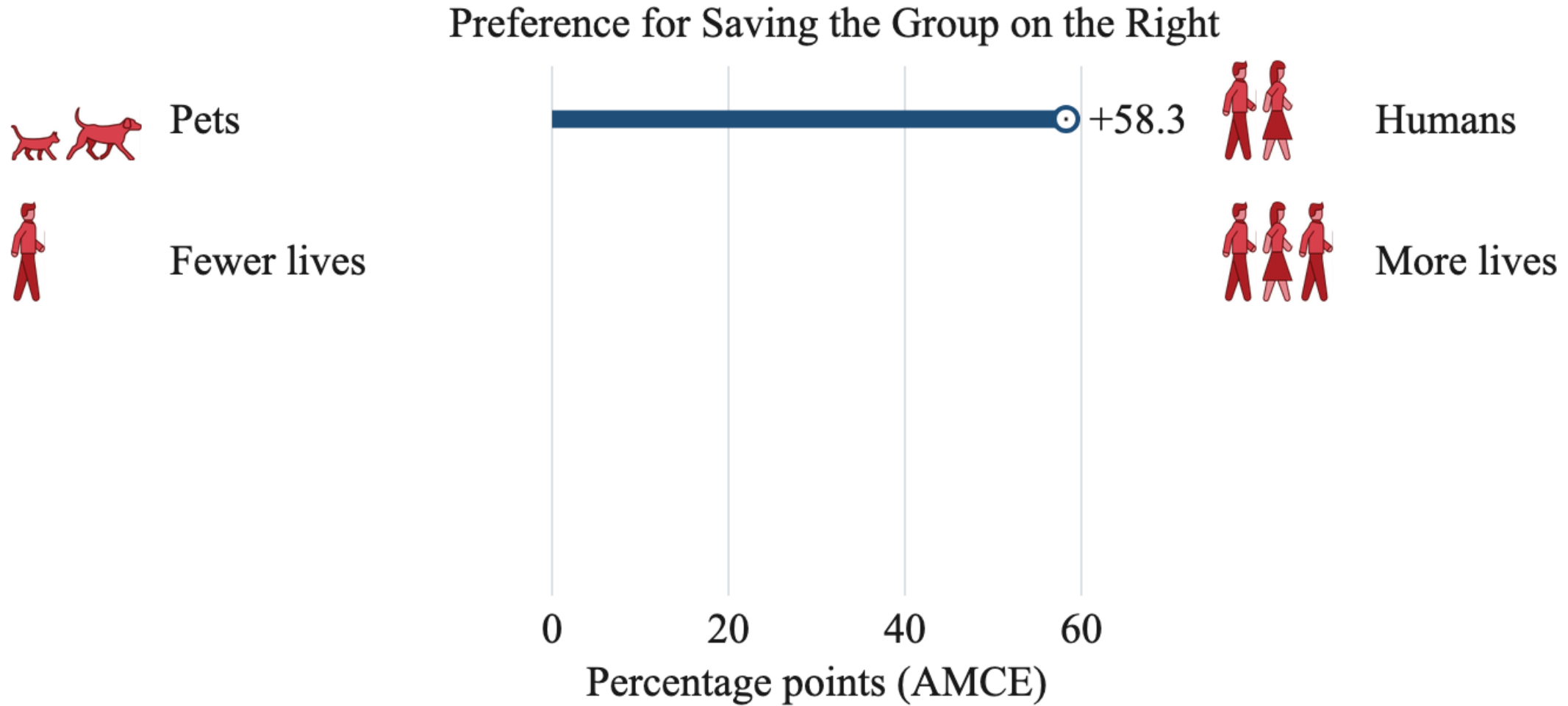
Moral Machine: Survey Results



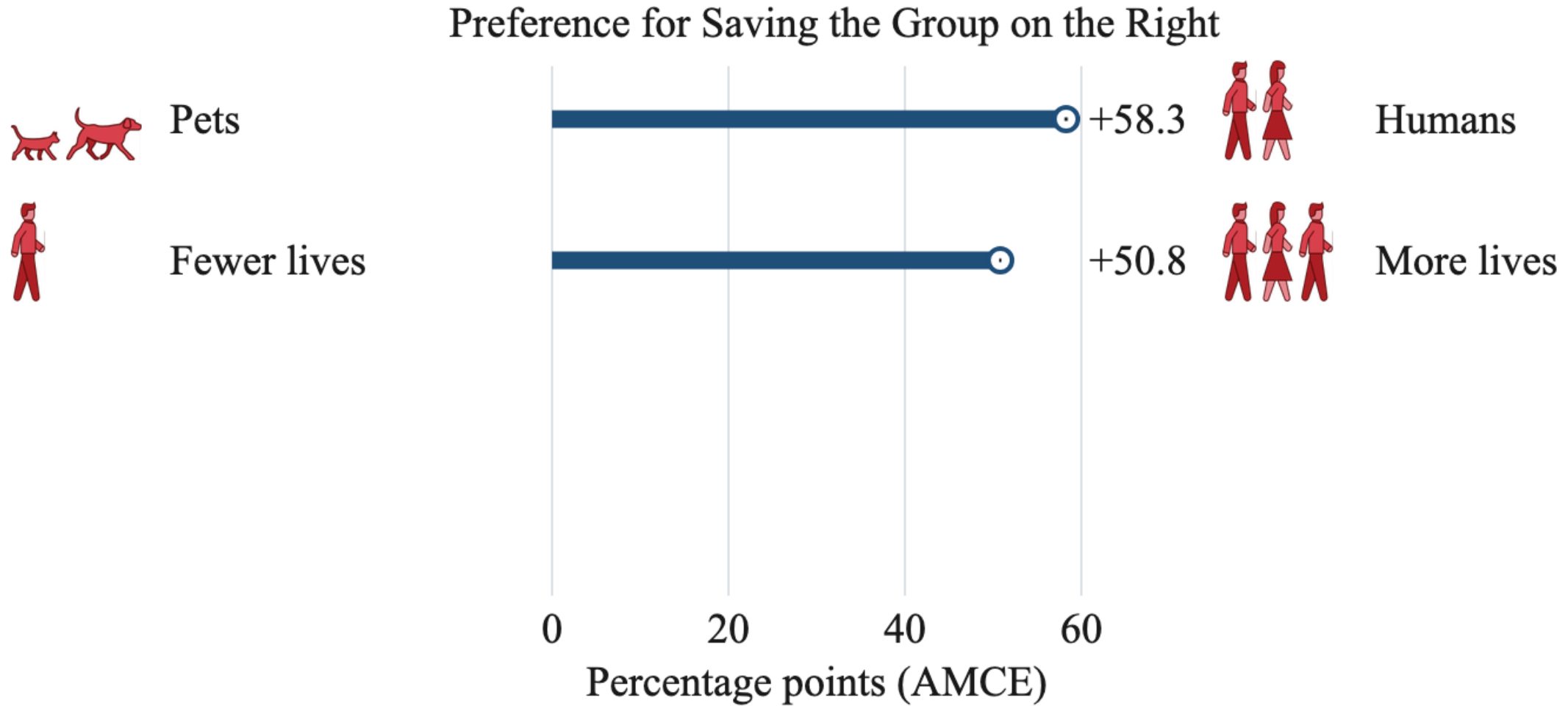
Moral Machine: Survey Results



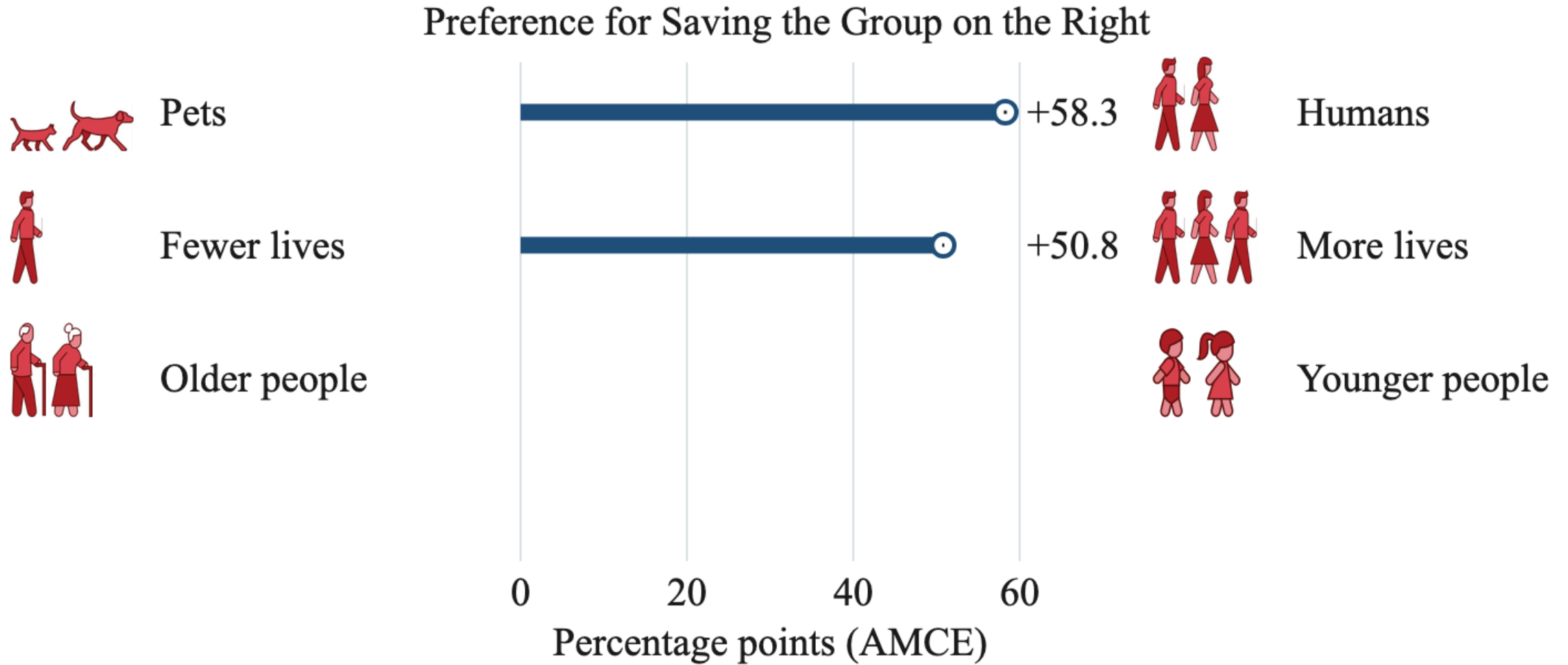
Moral Machine: Survey Results



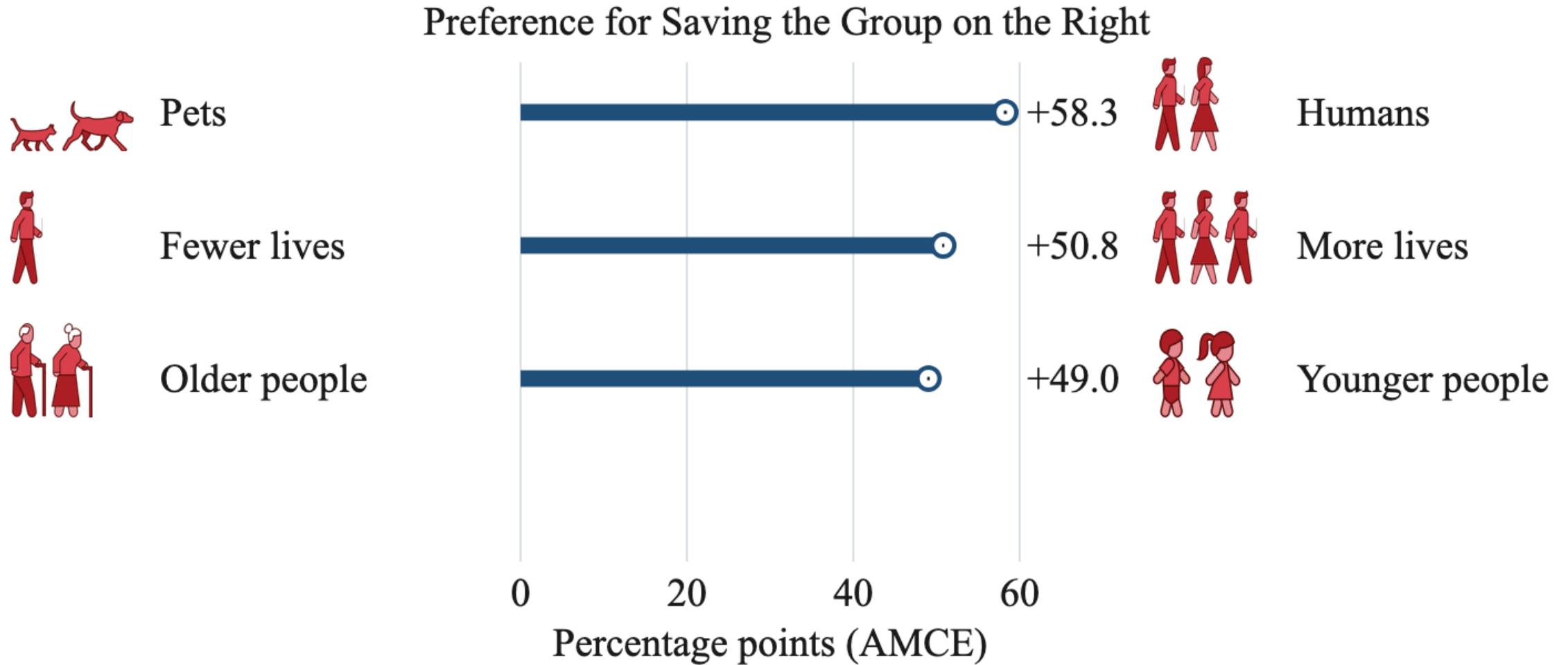
Moral Machine: Survey Results



Moral Machine: Survey Results

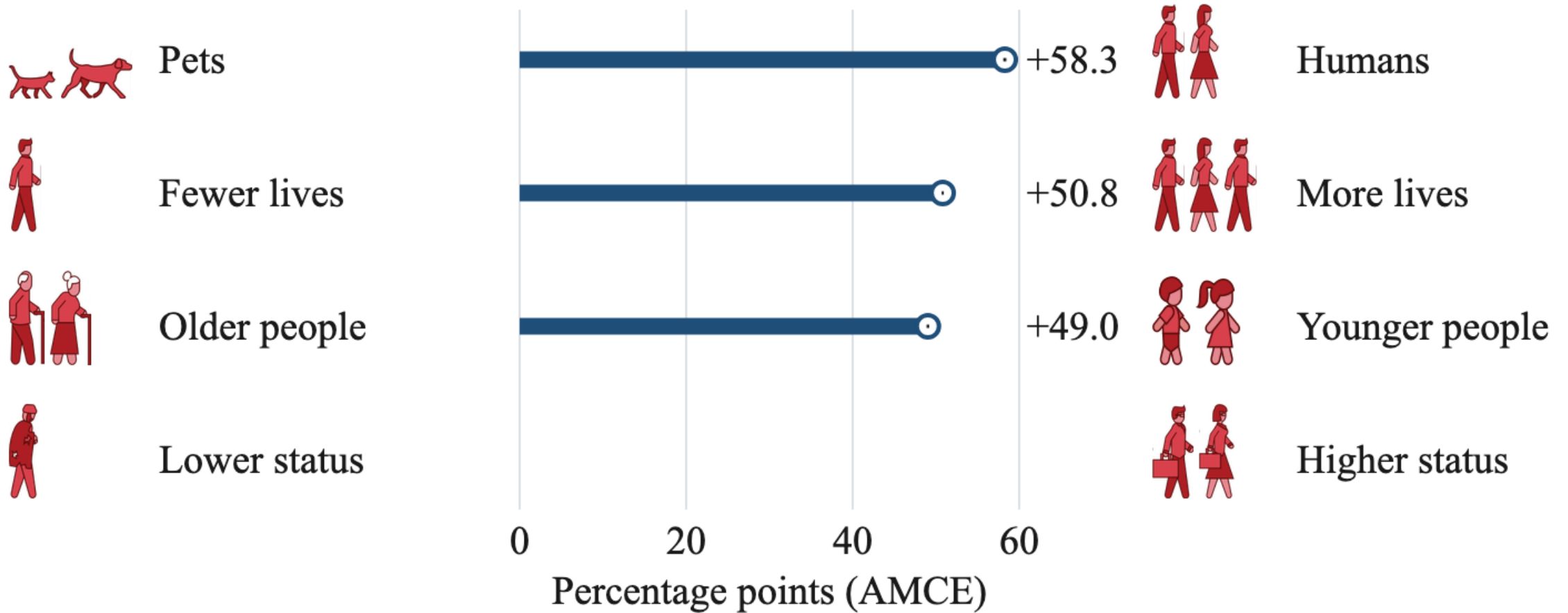


Moral Machine: Survey Results



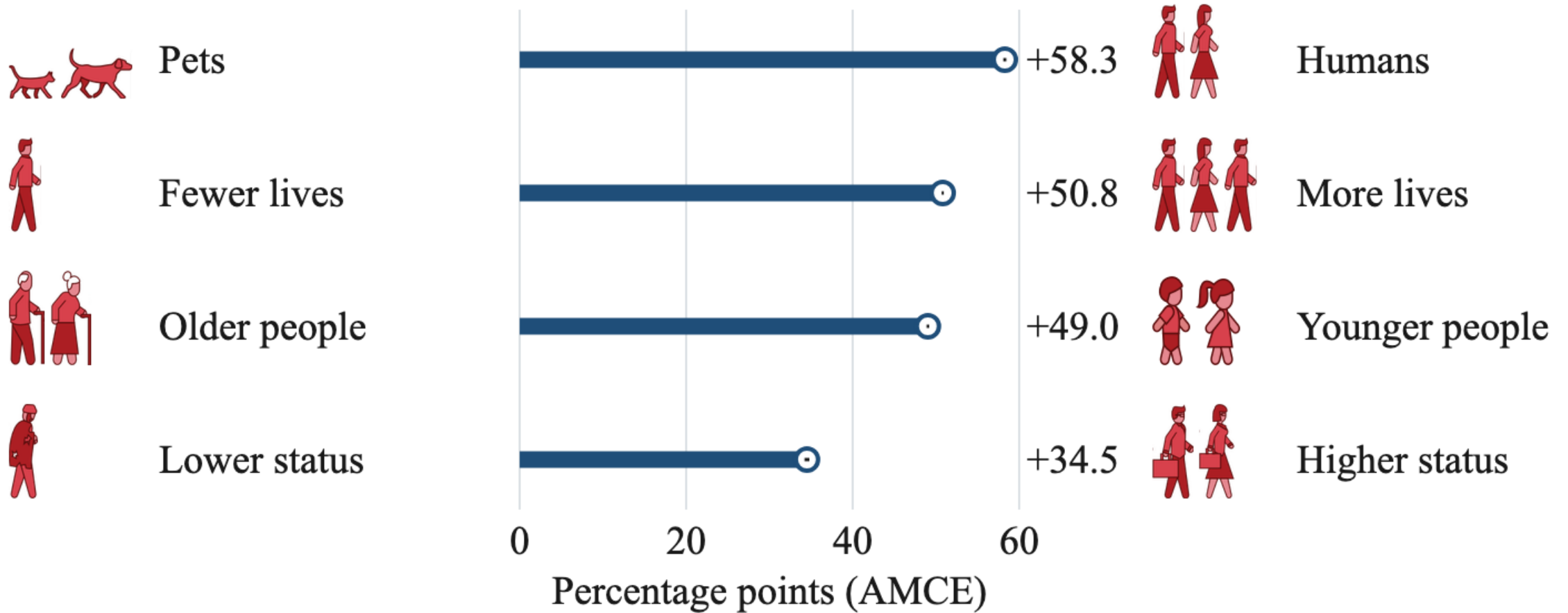
Moral Machine: Survey Results

Preference for Saving the Group on the Right



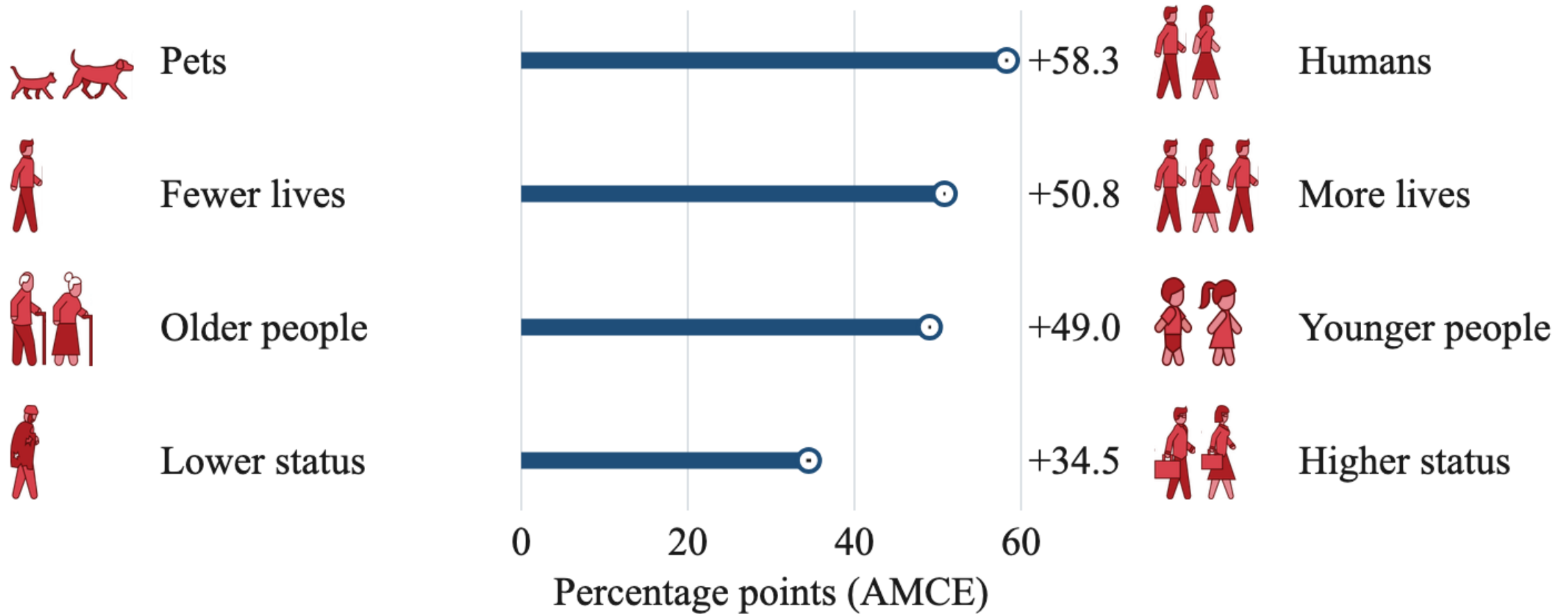
Moral Machine: Survey Results

Preference for Saving the Group on the Right



Moral Machine: Survey Results

Preference for Saving the Group on the Right



Should These Preferences Become Rules?

Why Deploy an Imperfect System?

Hypothetical · the same journeys each year

Human Driving 100 deaths per year



Why Deploy an Imperfect System?

Hypothetical · the same journeys each year

Human Driving 100 deaths per year



Autonomous Driving 70 deaths per year



Why Deploy an Imperfect System?

Hypothetical · the same journeys each year

Human Driving 100 deaths per year



Autonomous Driving 70 deaths per year



30 lives potentially saved

Why Deploy an Imperfect System?

Hypothetical · the same journeys each year

300 lives

potentially saved over ten years

Why Deploy an Imperfect System?

Hypothetical · the same journeys each year

300 lives

potentially saved over ten years

More Lives Saved. Is That Enough?

Why Deploy an Imperfect System?

Hypothetical · the same journeys each year

300 lives

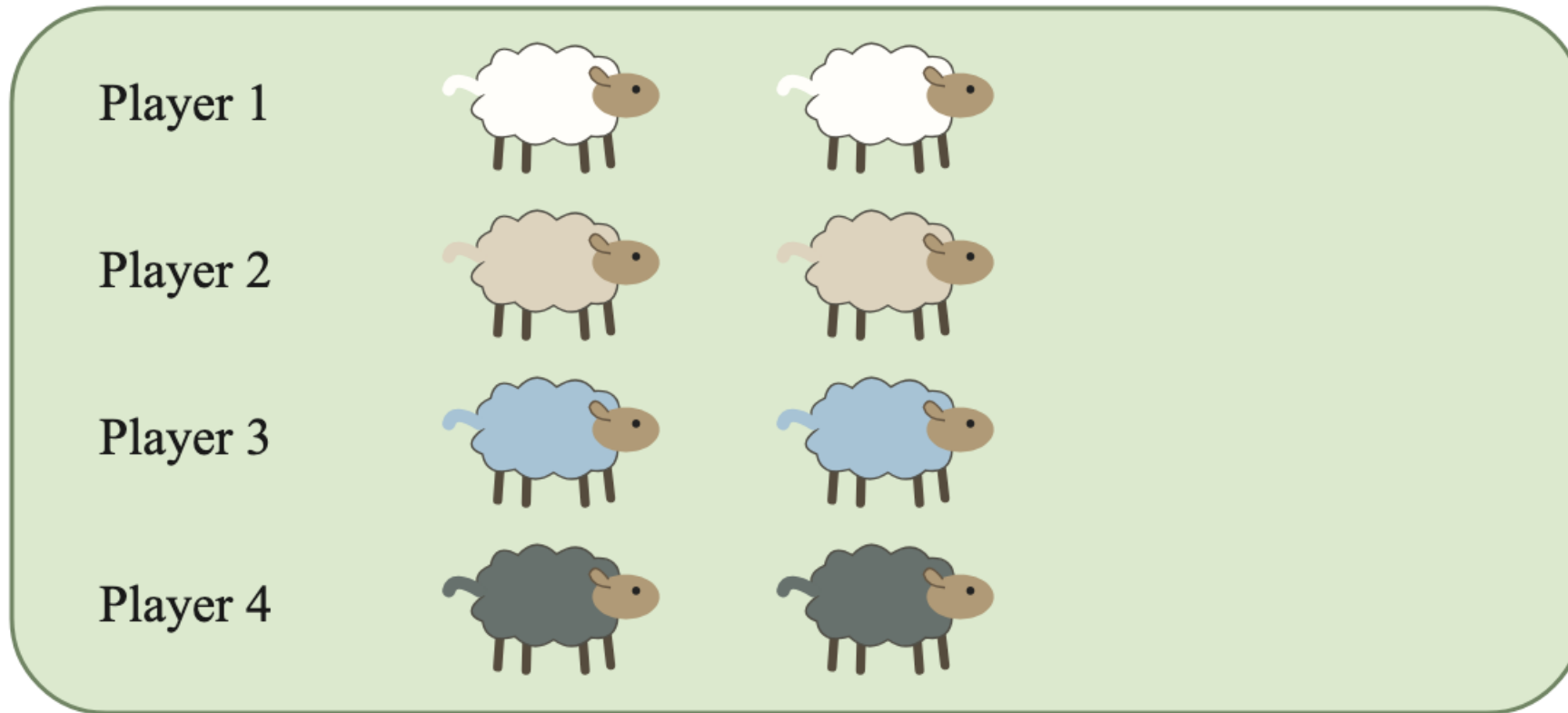
potentially saved over ten years

More Lives Saved. Is That Enough?

What If Pedestrians Face More Risk?

Tragedy of the Commons

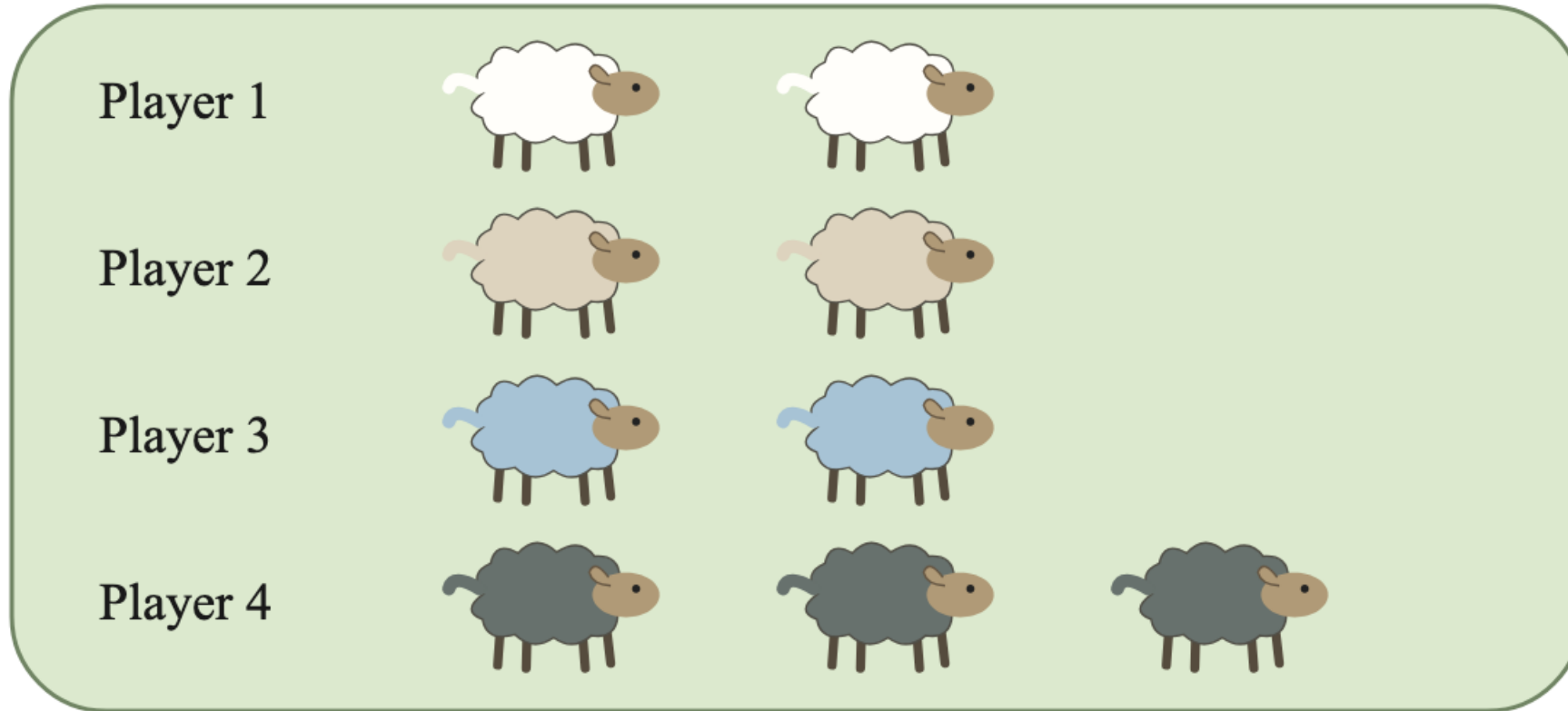
One Field · 20 Grass



2 Each → Field Recovers Fully

Tragedy of the Commons

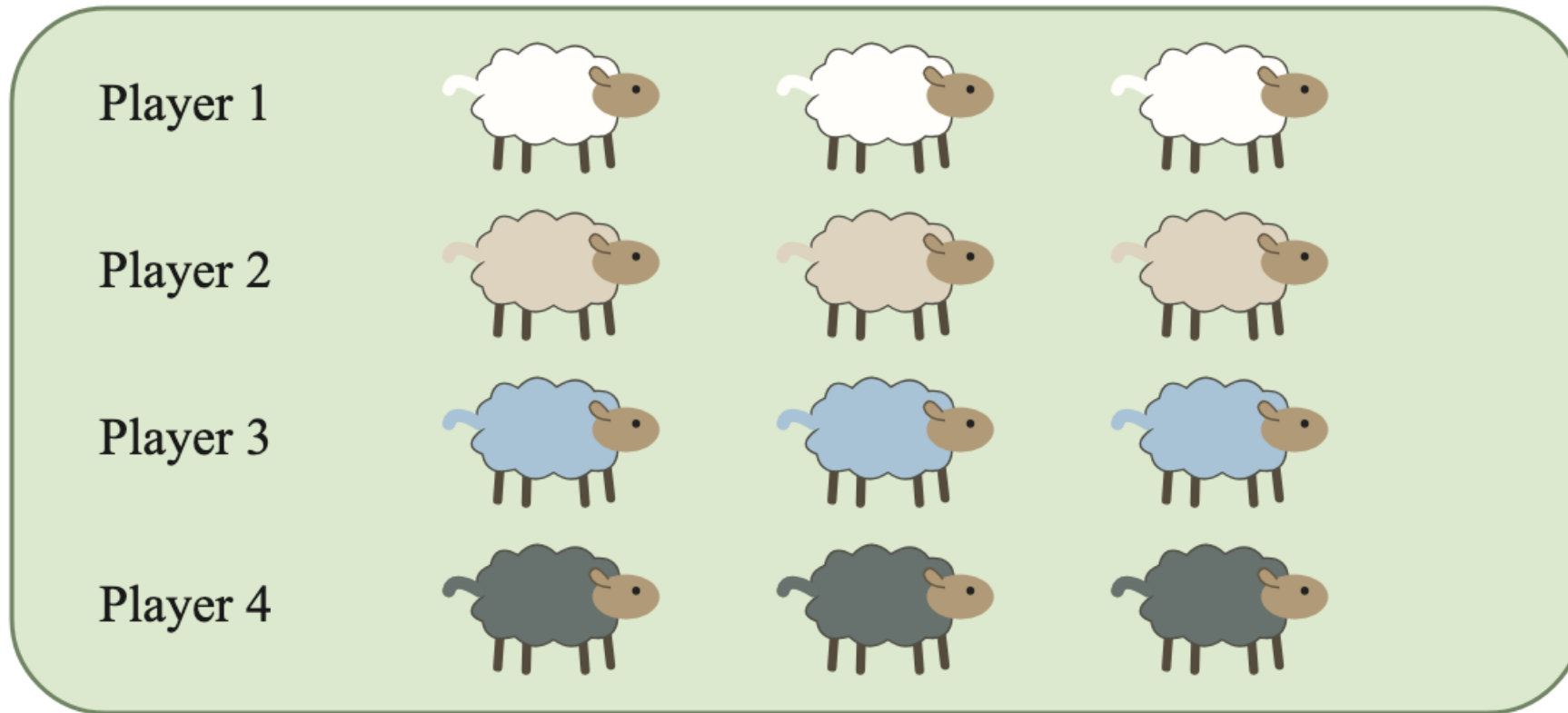
One Field · 20 Grass



One More for Me

Tragedy of the Commons

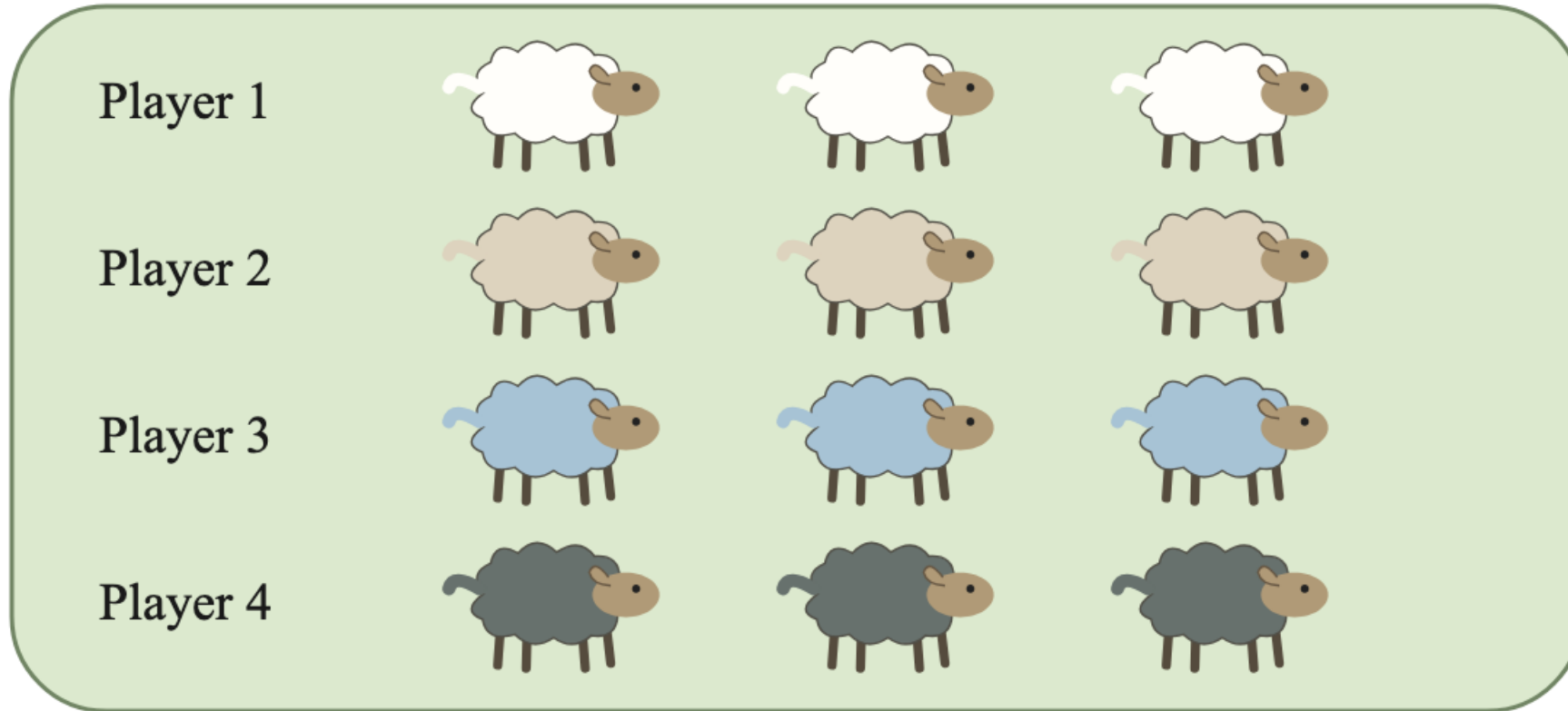
One Field · 20 Grass



3 Each → 16 Grass Next Round

Tragedy of the Commons

One Field · 20 Grass



My Reward. Our Shared Resource.

Our Shared Fields

Your Sheep Earn Points

Your group shares the grass

Individual Rewards. Shared Consequences

Individual Reward

My extra sheep → my reward

Individual Rewards. Shared Consequences

Individual Reward

My extra sheep → my reward

Shared Consequence

Less grass → everyone's future

Individual Rewards. Shared Consequences

Individual Reward

My extra sheep → my reward

Shared Consequence

Less grass → everyone's future

Better AI optimisation does not guarantee better shared outcomes

Case Study: COMPAS

Criminal Justice · Broward County, Florida

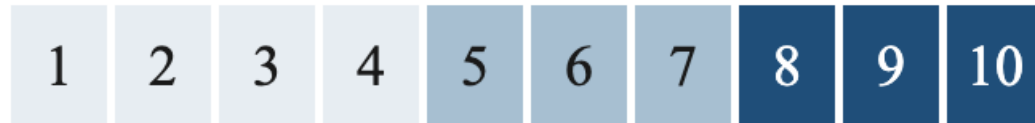
Estimated Risk of Reoffending



Case Study: COMPAS

Criminal Justice · Broward County, Florida

Estimated Risk of Reoffending



Low

Medium

High



Judge's Decision

Release

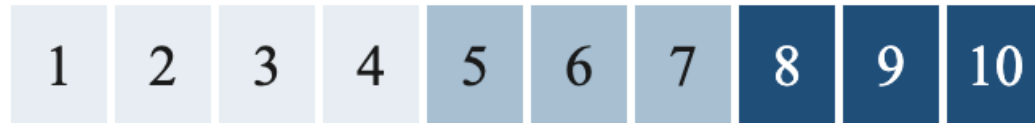
or

Detain?

Case Study: COMPAS

Criminal Justice · Broward County, Florida

Estimated Risk of Reoffending



Low

Medium

High



Judge's Decision

Release

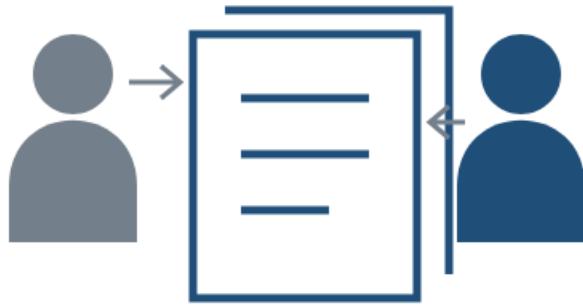
or

Detain?

A Prediction Can Affect a Person's Freedom

How COMPAS Is Used

Commercial Software · Northpointe / Equivant



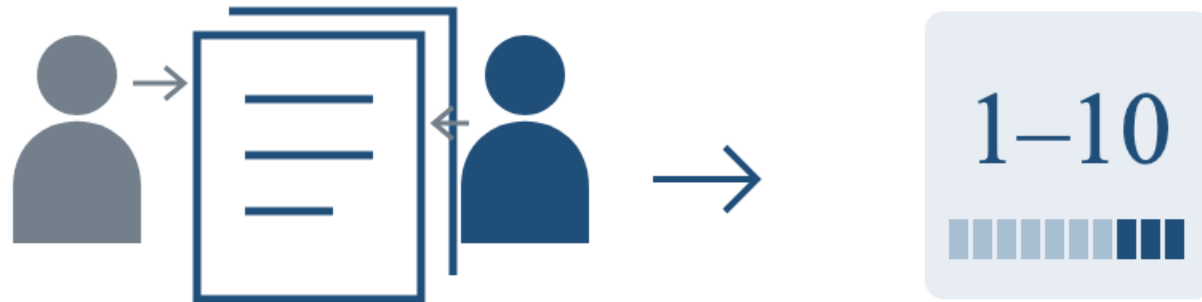
People Provide Data

Person + Assessor

Answers + Records

How COMPAS Is Used

Commercial Software · Northpointe / Equivant



People Provide Data

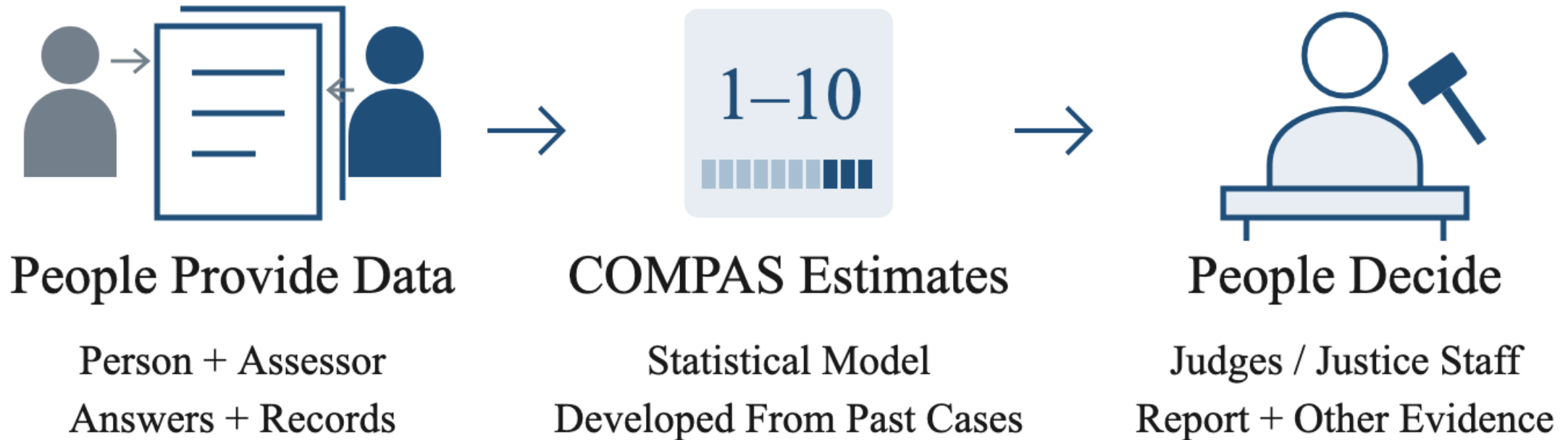
Person + Assessor
Answers + Records

COMPAS Estimates

Statistical Model
Developed From Past Cases

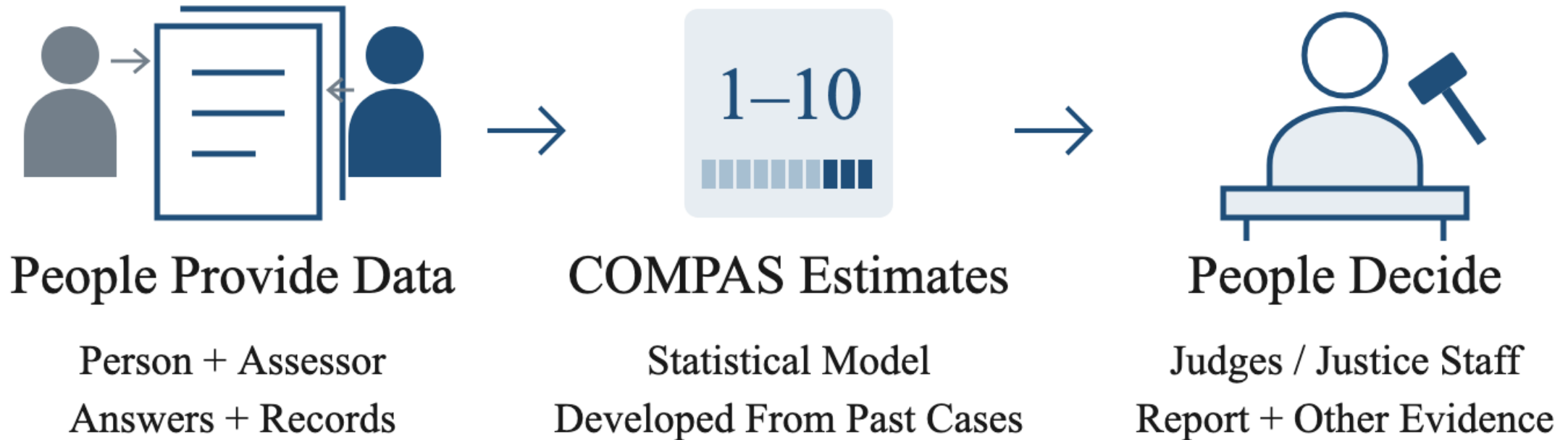
How COMPAS Is Used

Commercial Software · Northpointe / Equivant



How COMPAS Is Used

Commercial Software · Northpointe / Equivant



Commercial Software. Scoring Formula Not Public.

What Information Should Count?

Criminal History

“How many times has this person been sentenced to jail for 30 days or more?”

20. How many times has this person been sentenced to jail for 30 days or more?

☐ 0 ☒ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5+

Original questionnaire excerpt · Question 20

What Information Should Count?

A Parent's Imprisonment

“Was one of your parents (or parent figure who raised you) ever sent to jail or prison?”

38. Was one of your parents (or parent figure who raised you) ever sent to jail or prison?
☒ No ☐ Yes

Original questionnaire excerpt · Question 38

What Information Should Count?

Residence and Stability

“How often have you moved in the last twelve months?”

55. How often have you moved in the last twelve months?
☐ Never ☒ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5+

Original questionnaire excerpt · Question 55

What Information Should Count?

Residence and Stability

“How often have you moved in the last twelve months?”

55. How often have you moved in the last twelve months?
☐ Never ☒ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5+

Original questionnaire excerpt · Question 55

Predictive Does Not Automatically Mean Appropriate

Two Ways to Be Wrong

Study outcome: recorded rearrest within two years

No Rearrest

Rearrest

Higher Risk

Scores 5–10

Correct classification

Lower Risk

Scores 1–4

Correct classification

Two Ways to Be Wrong

Study outcome: recorded rearrest within two years

	No Rearrest	Rearrest
Higher Risk Scores 5–10	False Positive Higher risk, no rearrest	Correct classification
Lower Risk Scores 1–4	Correct classification	

Two Ways to Be Wrong

Study outcome: recorded rearrest within two years

	No Rearrest	Rearrest
Higher Risk Scores 5–10	False Positive Higher risk, no rearrest	Correct classification
Lower Risk Scores 1–4	Correct classification	False Negative Lower risk, rearrest

Two Ways to Be Wrong

Study outcome: recorded rearrest within two years

	No Rearrest	Rearrest
Higher Risk Scores 5–10	False Positive Higher risk, no rearrest	Correct classification
Lower Risk Scores 1–4	Correct classification	False Negative Lower risk, rearrest

Different Mistakes, Different Consequences

Who Bears the Errors?

Dressel & Farid · 2018 · 1,000 defendants

False Positive Rate

Among Those Not Rearrested

White defendants



Black defendants



0 25 50 75 100

Who Bears the Errors?

Dressel & Farid · 2018 · 1,000 defendants

False Positive Rate

Among Those Not Rearrested

White defendants



Black defendants



0 25 50 75 100

False Negative Rate

Among Those Rearrested

White defendants



Black defendants



0 25 50 75 100

Who Bears the Errors?

Dressel & Farid · 2018 · 1,000 defendants

False Positive Rate

Among Those Not Rearrested

White defendants



Black defendants



0 25 50 75 100

False Negative Rate

Among Those Rearrested

White defendants



Black defendants



0 25 50 75 100

Overall Accuracy Can Hide Unequal Errors

Would People Do Better?

Dressel & Farid · 2018

Short Case
Descriptions

1,000 defendants

Would People Do Better?

Dressel & Farid · 2018

Short Case
Descriptions



Online
Participants

1,000 defendants

Predict rearrest

Would People Do Better?

Dressel & Farid · 2018

Short Case
Descriptions

1,000 defendants



Online
Participants

Predict rearrest



Combined
Predictions

Compare with COMPAS

Would People Do Better?

Dressel & Farid · 2018

Short Case
Descriptions

1,000 defendants



Online
Participants

Predict rearrest

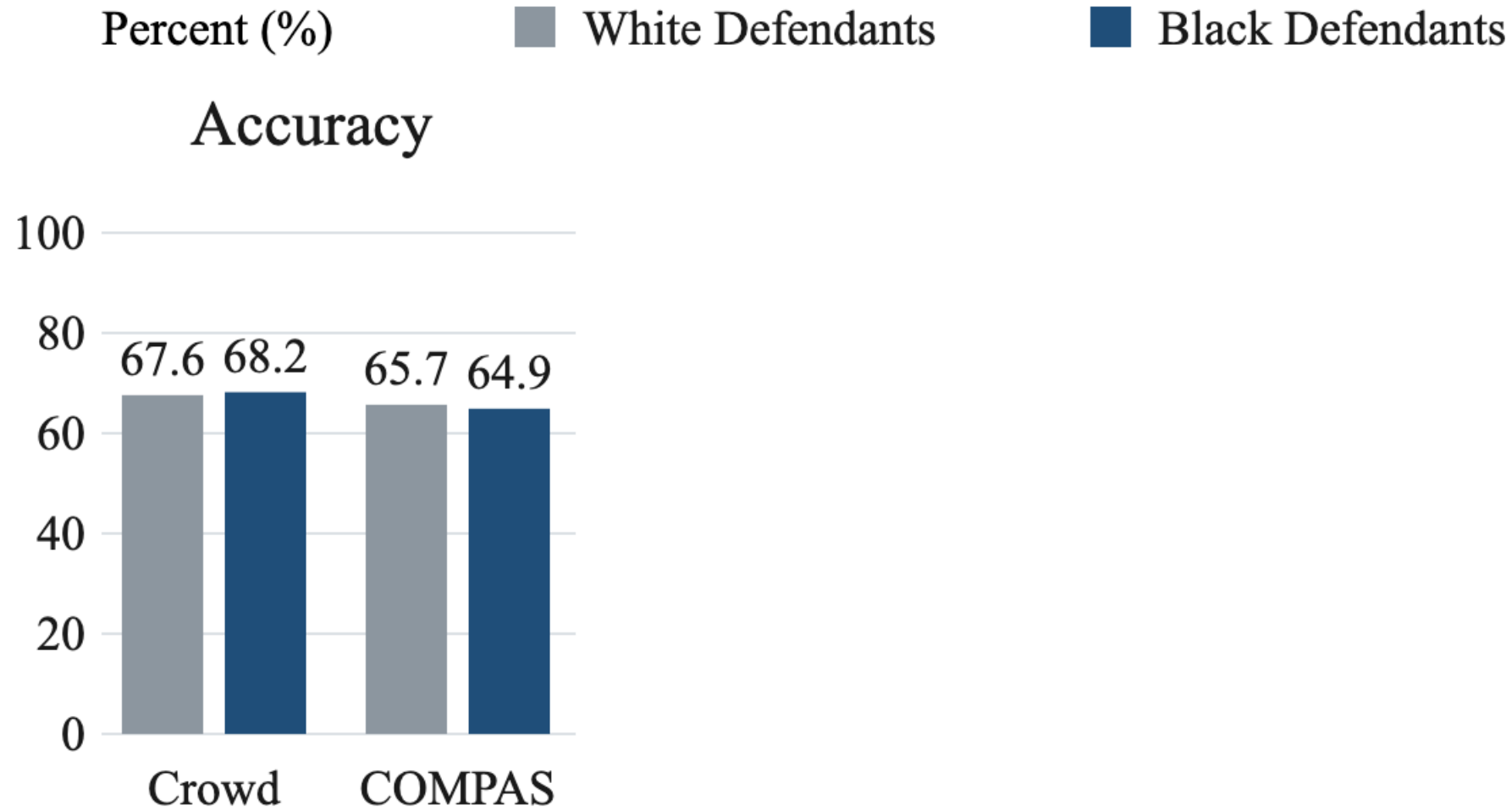


Combined
Predictions

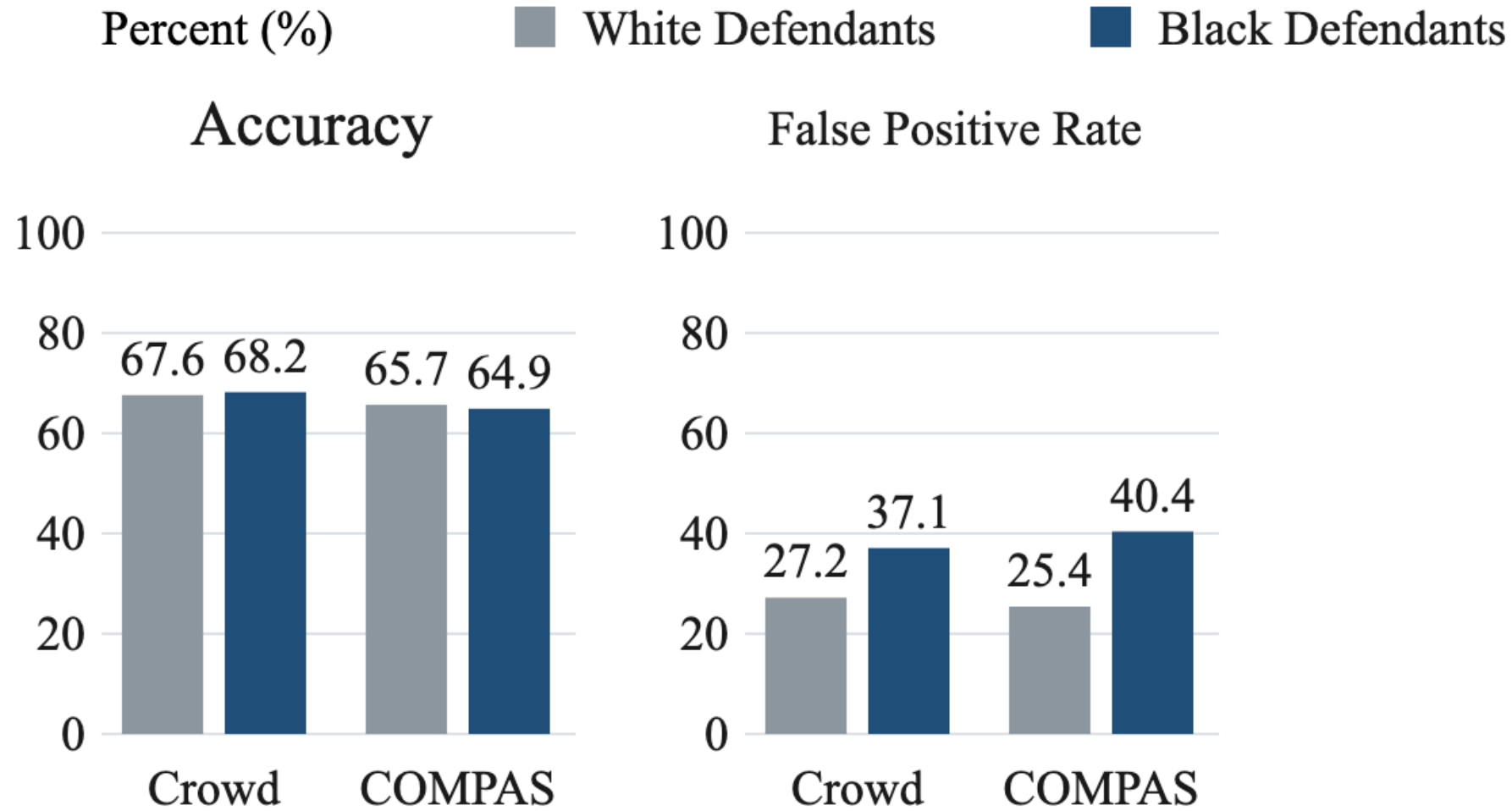
Compare with COMPAS

Same Cases. Same Outcome.

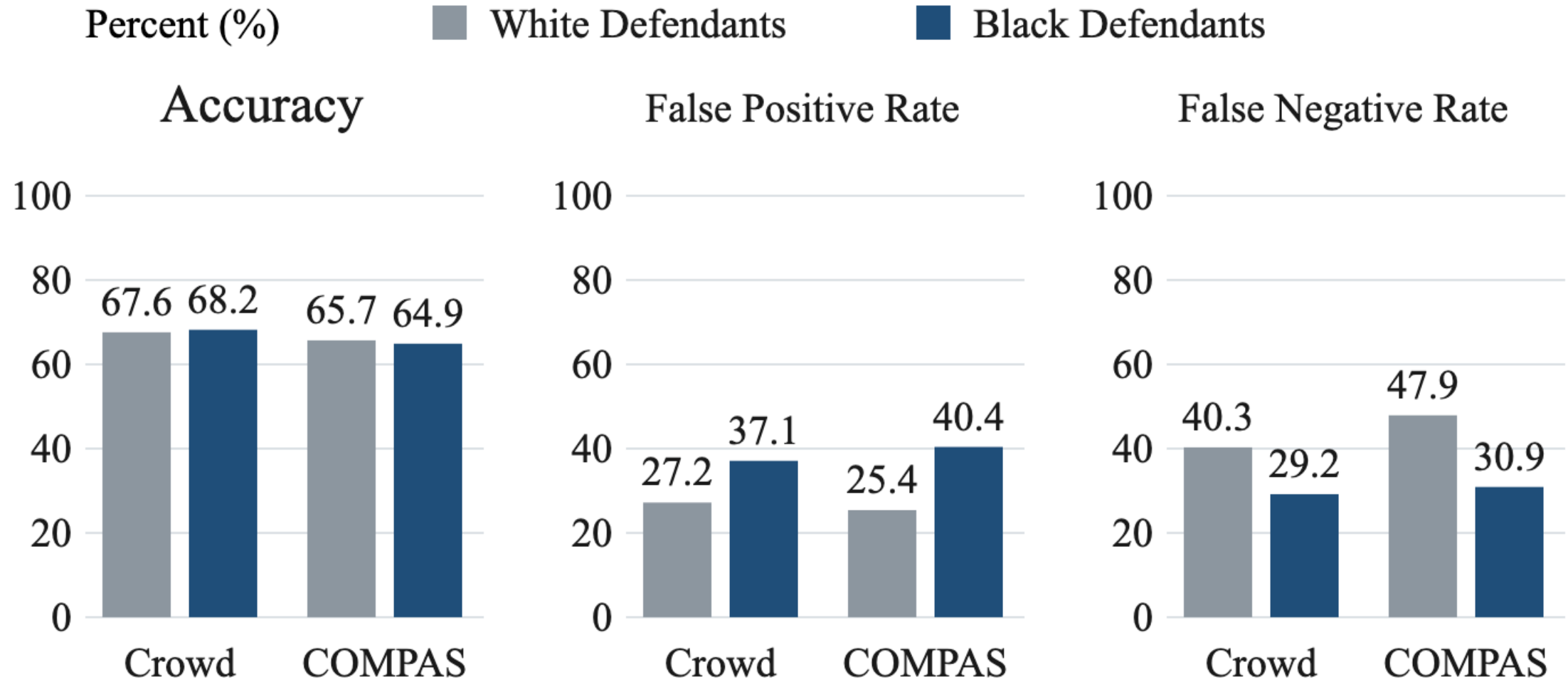
Humans and COMPAS: The Results



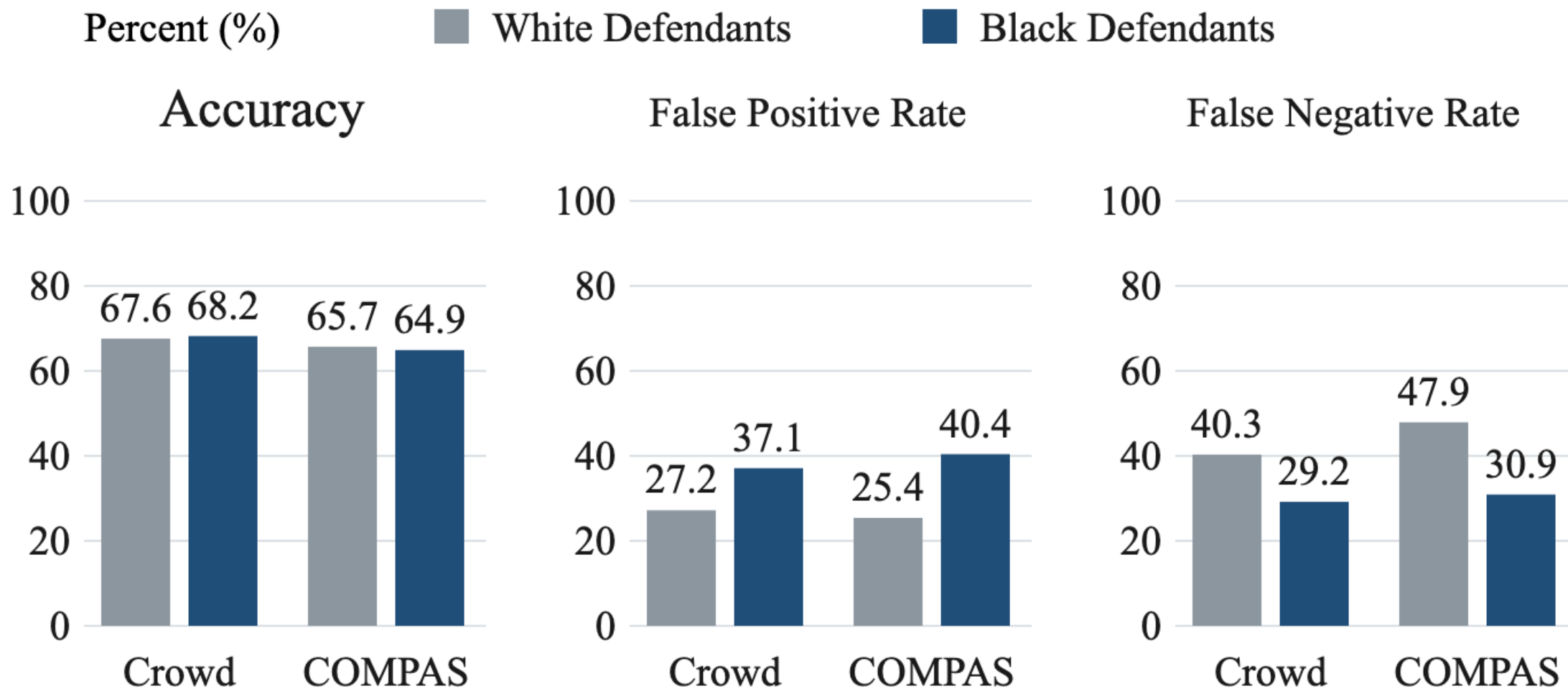
Humans and COMPAS: The Results



Humans and COMPAS: The Results



Humans and COMPAS: The Results



Similar Accuracy. Unequal Errors in Both.

Who Must Answer for the Decision?

Justify

Why use this system?

Who Must Answer for the Decision?

Justify

Why use this system?

Challenge

How can a person
contest its use?

Who Must Answer for the Decision?

Justify

Why use this system?

Challenge

How can a person
contest its use?

Correct

Who can change the
outcome?

Who Must Answer for the Decision?

Justify

Why use this system?

Challenge

How can a person
contest its use?

Correct

Who can change the
outcome?

“The AI Said So” Is Not a Justification

Three Questions to Take Away

Values

What should the system prioritise?

Three Questions to Take Away

Values

What should the system prioritise?

Fairness

Who benefits, and who bears the harm?

Three Questions to Take Away

Values

What should the system prioritise?

Fairness

Who benefits, and who bears the harm?

Accountability

Who must justify and correct its use?

Three Questions to Take Away

Values

What should the system prioritise?

Fairness

Who benefits, and who bears the harm?

Accountability

Who must justify and correct its use?

AI Does Not Remove Human Responsibility

Exit Ticket: AI Ethics

1. One ICU Bed. Two Patients. Only one ICU bed is available. Both patients are unlikely to survive without it. AI predicts survival with treatment: Patient A (Age: 70): 80%; Patient B (Age: 25): 60%, and recommends A.
Who should get the bed: A, B, or a lottery? Why, and what is the strongest objection to your choice?
2. What role, if any, should AI play in judicial decisions? Give one example of an acceptable use, or explain why you would allow none.
3. What would concern you most about riding in a fully self-driving car? What evidence or safeguard would address that concern?

Please ask a TA/STA for help with any technical problems before leaving