



Brown University

Lab Web Page:

bit.ly/scsrl



Could Generative Language Models Have Moral Trustworthiness?

Bertram F. Malle

Cognitive, Linguistic, and Psychological Sciences

Brown University

Support:



Questions

1. Does it even make sense to treat LLMs as “agents” that “have trustworthiness”?
2. If we do query LLMs’ trustworthiness, what specific moral attributes of trustworthiness would we want them to exhibit?
3. What would it take to design LLMs that actually have those attributes of moral trustworthiness?

Previous Studies: Some Findings

People ascribe a variety of capacities to robots

- ▶ Little to no *experience*, moderate *moral and social cognition*, substantial *reality interaction* Malle (2019). How many dimensions of mind perception... *CogSci Proceedings*
- ▶ Data on AI not yet available but

People explain robot behavior using

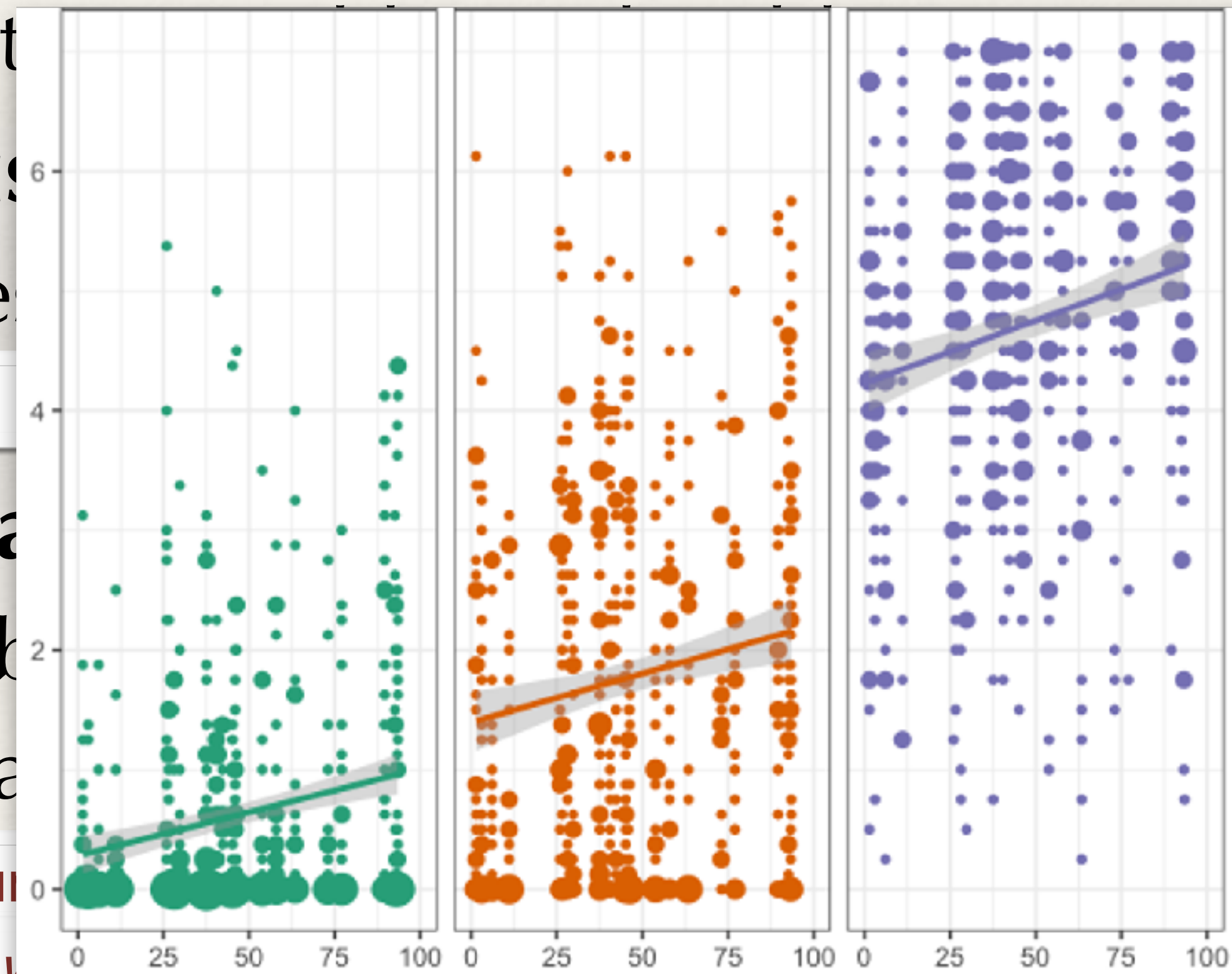
- ▶ For robots, prefer belief over desire mental state ascriptions de Graaf & Malle (2019). *CogSci Proceedings*

70% of people consider robots as

- ▶ Blame robot more for inaction than action
- ▶ Possible explanation for both cases

Malle et al. (2025). People's judgments of human-like robots

Malle et al. (2019). AI in the sky. *Robots and Human Interaction*

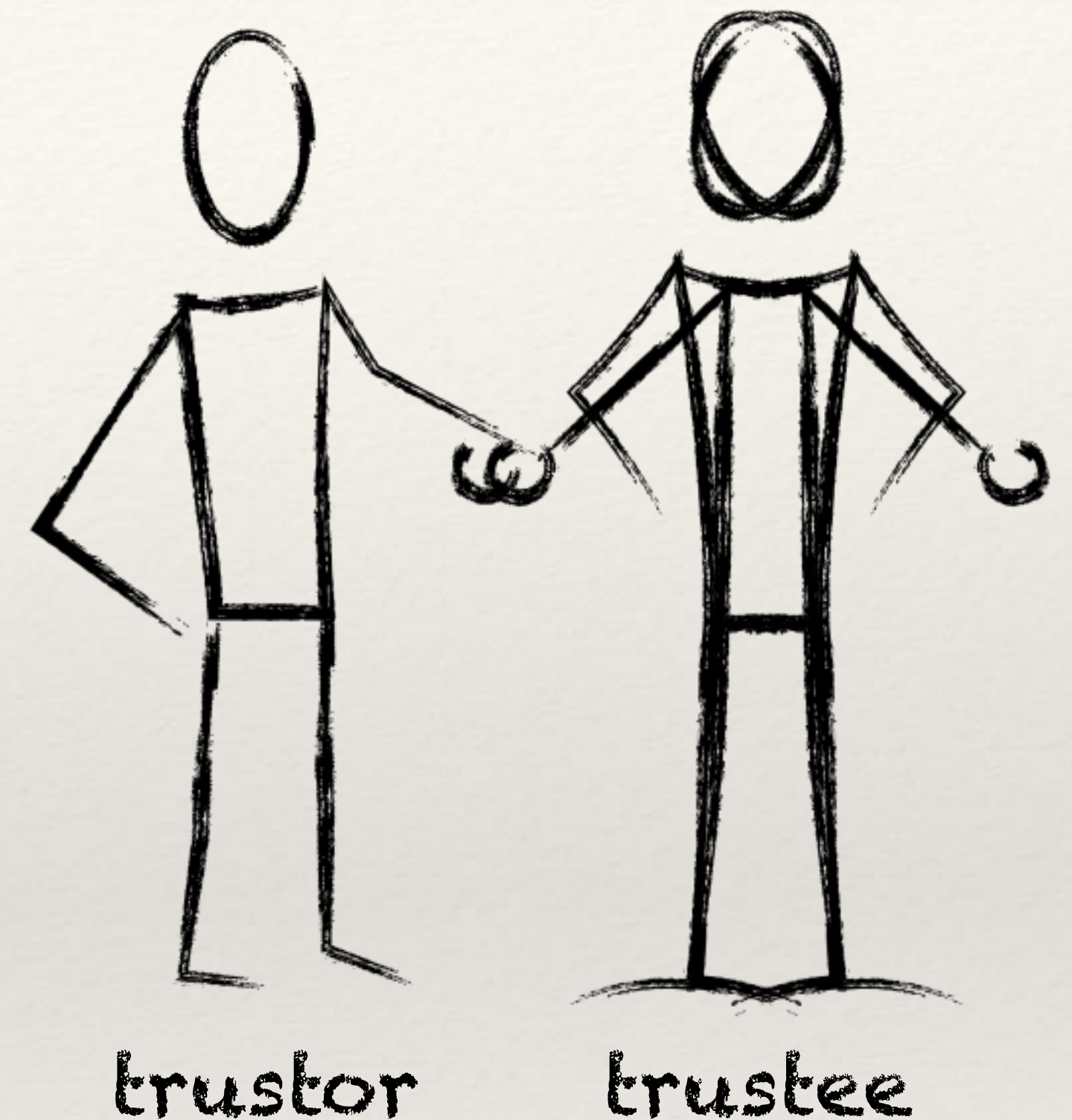


Trust

What is Trust?

Starting Point: “I trust you to do something”

- ▶ the trustee's future action has some benefit but also puts the trustor at risk
- ▶ the trustor accepts this risk because
- ▶ they expect that the trustee will minimize the risk by virtue of certain relevant attributes of trustworthiness.



Trust $\approx$ An expectation
of trustworthiness

What is Trustworthiness?

	Performance		Moral		
	Competence	Reliability	Integrity	Sincerity	Benevolence
Rotter (1967)				✓	
Chun and Campbell (1974)			✓	✓	✓
Gabarro (1978)	✓	✓	✓	✓ ^a	✓ ^b
Parsons (1969)	✓		✓		
Cook and Wall (1980)	✓	✓			✓

Trust(worthiness) is multi-dimensional

Mayer et al. (1995)	✓		✓		✓
Slovic et al. (1993)	✓			✓	✓
Carnevale (1995)	✓	✓	✓		✓
McKnight et al. (1998)	✓	✓	✓ ^e	✓ ^e	✓
Caldwell and Clapham (2003)	✓		✓		✓
Analysis of definitions in Burke et al. (2007)	✓	✓	✓		✓
Kim et al. (2009)	✓		✓		

Evidence for Multiple Dimensions

Previous work (ratings and sorting): **4D**

New Study (free sorting):

- ▶ specific effort to include **benevolence**-related
- ▶ 41 trust-related words or phrases
- ▶ open ended number of bins, unlabeled

Once you have finished the sorting task, we would like you to pick the word from each group that best represents the entire group. You do not need to label the "OTHER" group.

Box 1	
Box 2	
Box 3	
Box 4	
Box 5	
Box 6	

Chaffin, D. (2021). *Developing a multi-dimensional model and measure of human-robot trust*. Ph.D. thesis, Brown University.

Items

Someone you can have faith in

Someone you can count on

Responsible

Reliable

Truthful

Sincere

Diligent

Authentic

Honest

Believable

Steadfast

Loyal

Skilled

Benevolent

Shows goodwill

Accurate

Considerate

Predictable

Kind

Open

Trustworthy

Meticulous

Genuine

Acts with others in mind

Frank

1

2

3

4

5

6

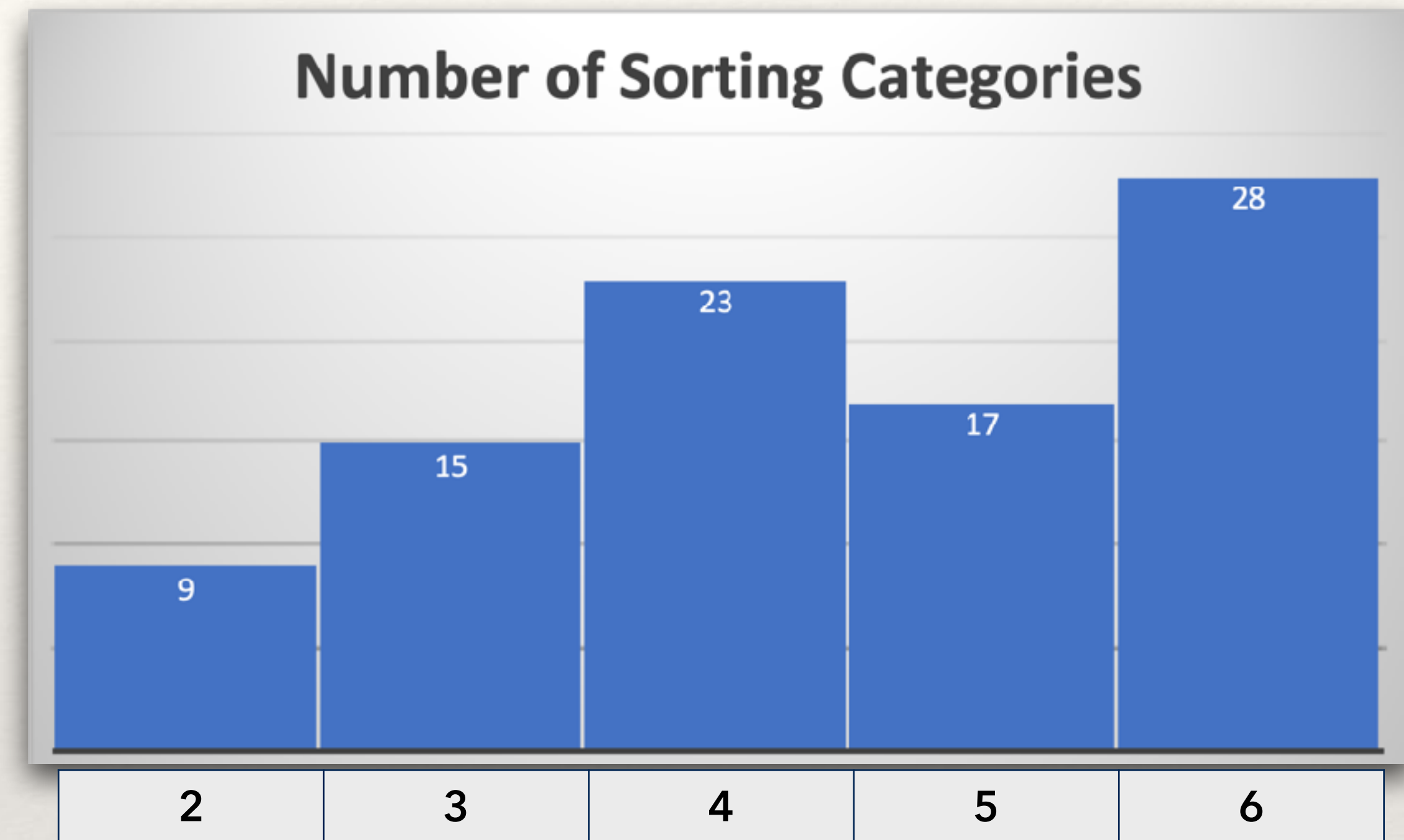
OTHER

Results

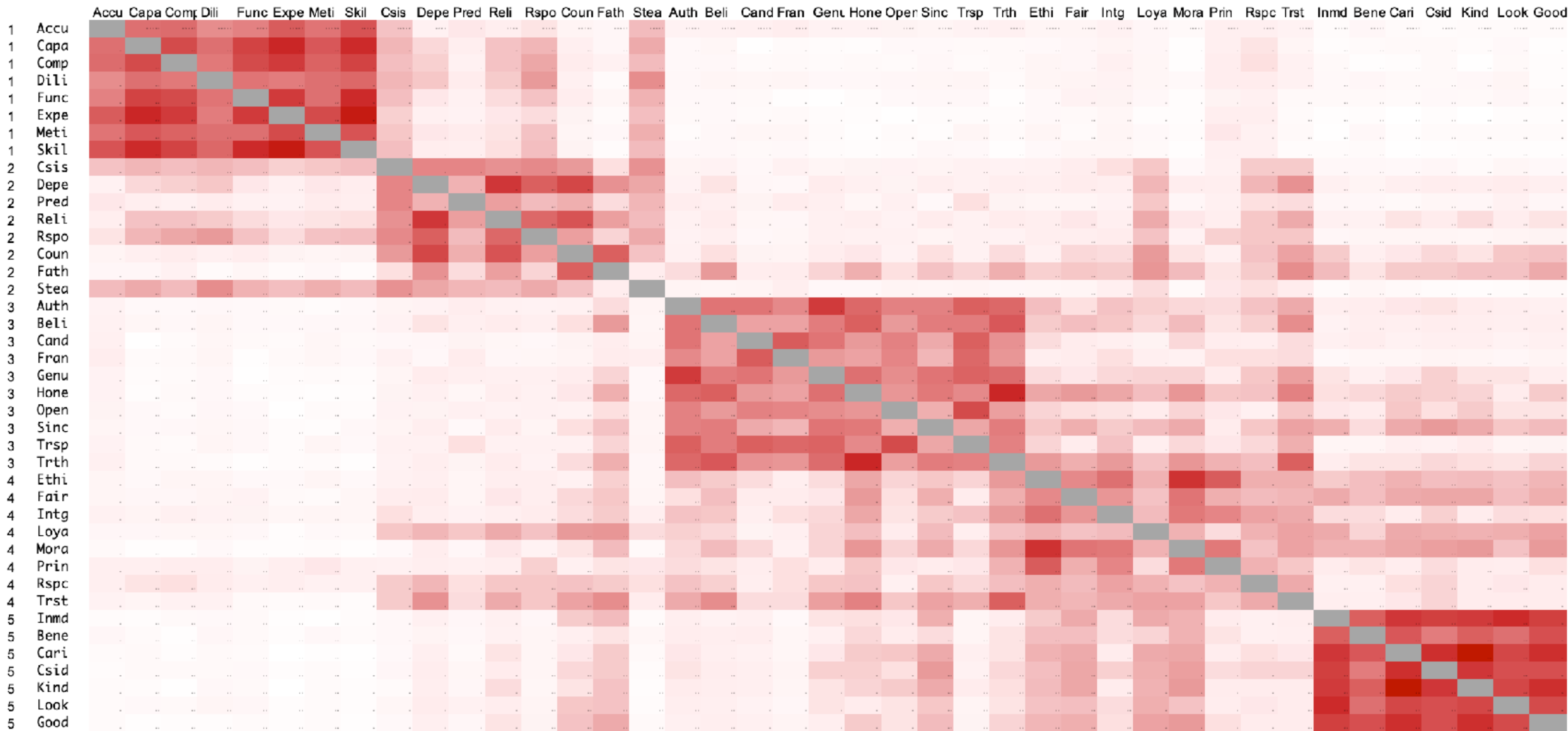
74% created 4 or more categories

People's own labels reflected the hypothesized categories

<i>First 5</i>		<i>Next 5</i>		<i>Next 5</i>	
Competent	29	Skilled	17	Capable	14
Dependable	23	Reliable	14	Responsible	14
Moral	30	Loyal	15	Ethical	11
Honest	28	Authentic	15	Candid	13
Caring	21	Kind	20	Benevolent	13

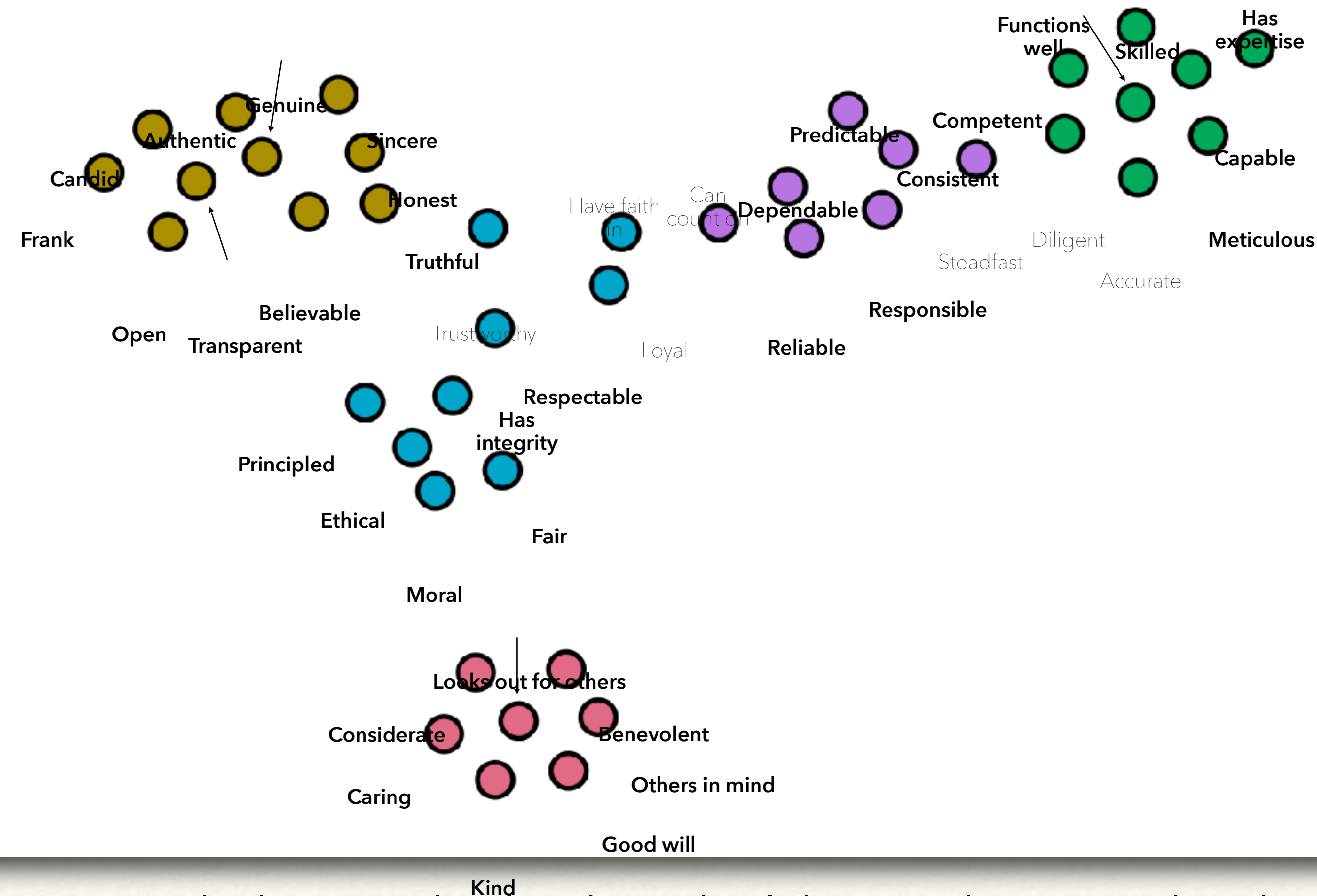


Co-Occurrence Patterns



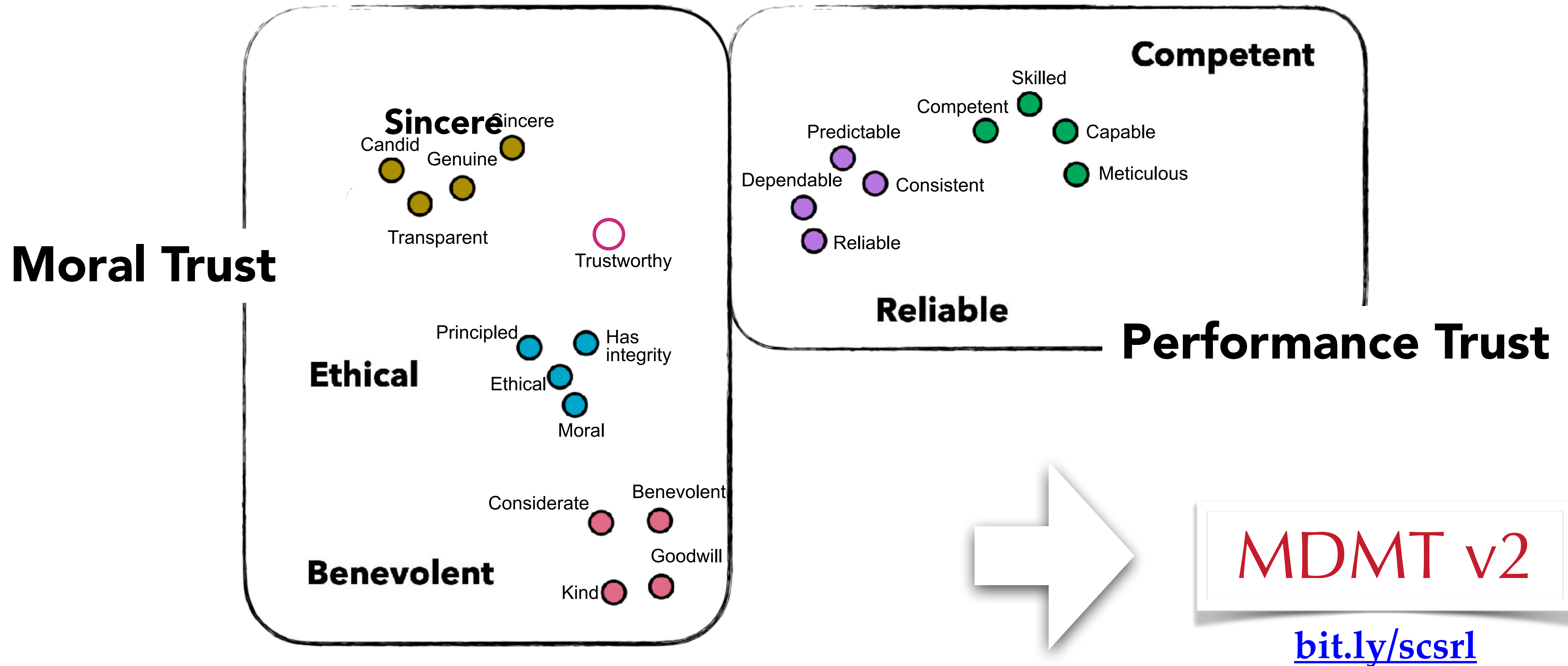
Five Clusters → Dimensions

k-Means Cluster Analysis



Replicates in multi-dimensional scaling, hierarchical cluster analysis, network analysis...

Dimensions of Perceived Trustworthiness



MDMT v2

https://research.clps.brown.edu/SocCogSci/Measures/MDMT_v2.pdf

MDMT v2 (2020-09-01)

© 2020 Daniel Ullman & Bertram F. Malle

MDMT: Multi-Dimensional Measure of Trust

Daniel Ullman & Bertram F. Malle

OVERVIEW

The Multi-Dimensional Measure of Trust (MDMT) is designed to be an intuitive and comprehensive measure of trust that is simple to administer in person or online. The MDMT was created to address the pressing need for valid measurement tools in the domain of human-robot trust but can also be used in human-human trust situations. The MDMT captures the different dimensions of trust in an agent. The first version MDMT v1 (2019-04-01) has been updated based on subsequent research to MDMT v2 (2020-09-01). Both versions are shared below to clearly exhibit the adjustments and the overall continuity.

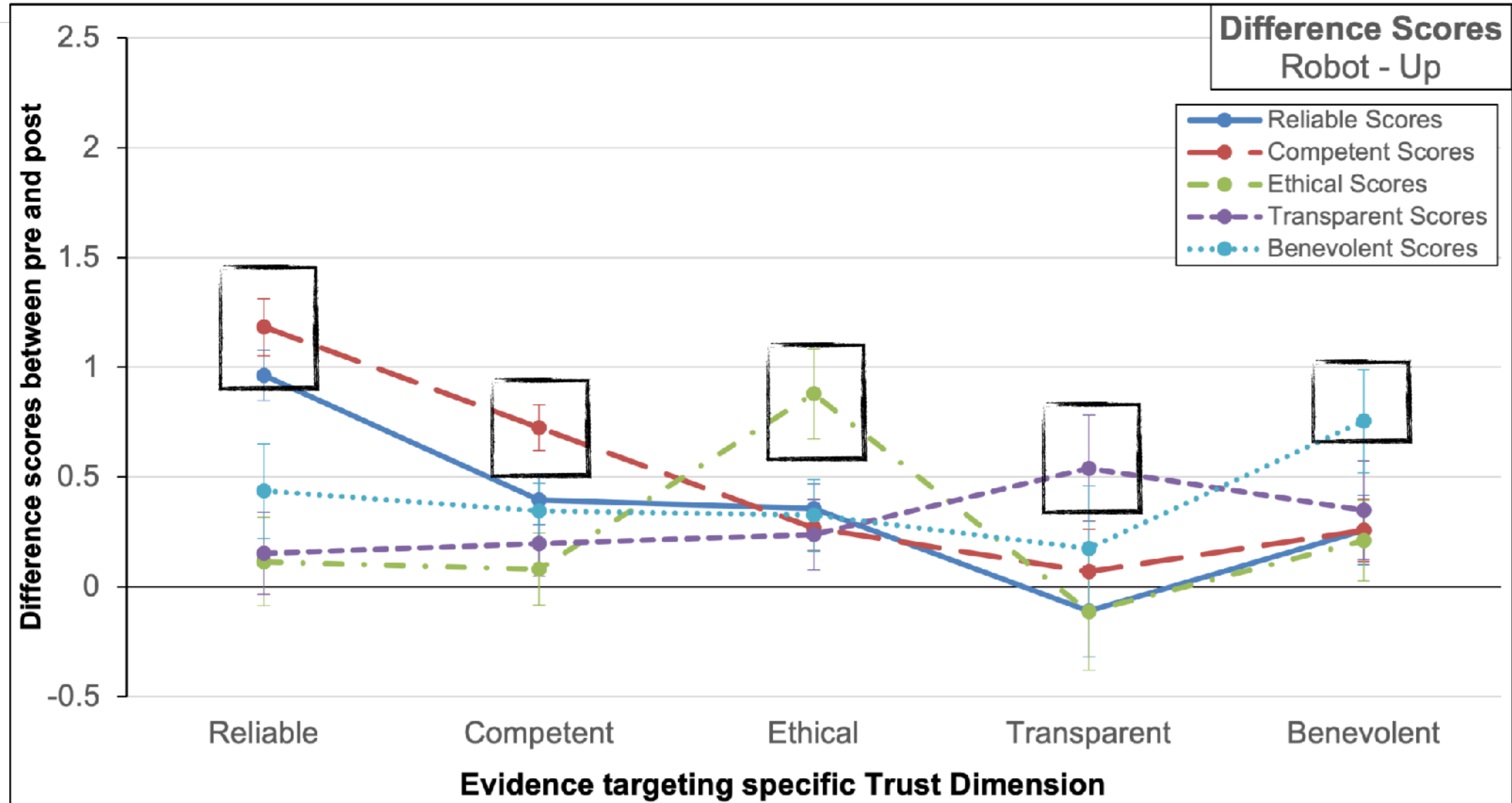
MDMT v2 (2020-09-01) [UPDATED]					
MDMT v2 (2020-09-01) [UPDATED]	Performance Trust		Moral Trust		
	Competent	Reliable	Ethical	Transparent	Benevolent
	competent	reliable	ethical	transparent	benevolent
	skilled	predictable	principled	genuine	kind
PERFORMANCE TRUST	capable	dependable	moral	sincere	considerate
	skilled	predictable	principled	genuine	kind
MORAL TRUST	capable	dependable	moral	sincere	considerate
	skilled	predictable	principled	genuine	kind

Trust Studies: Major Results

Trust responds to evidence specific to 2 factors and 5 dimensions

- ▶ for humans, robots, AI

Dimension-Specific Trust Gain



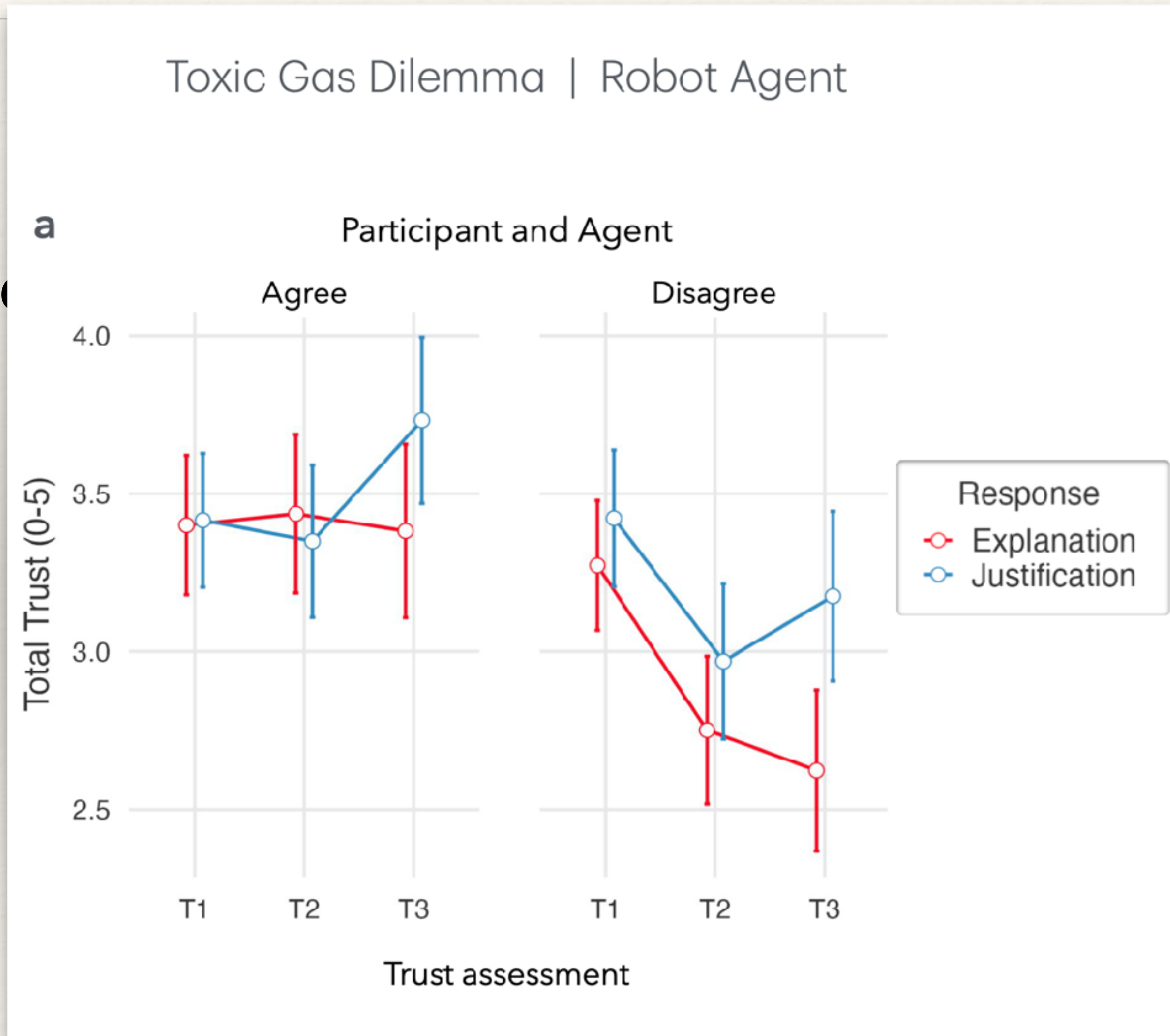
Trust Studies: Major Results

Trust responds to evidence

- ▶ for humans, robots, AI

Trust responds to justification

- ▶ for humans, robots



Trust Studies: Major Results

Trust responds to evidence

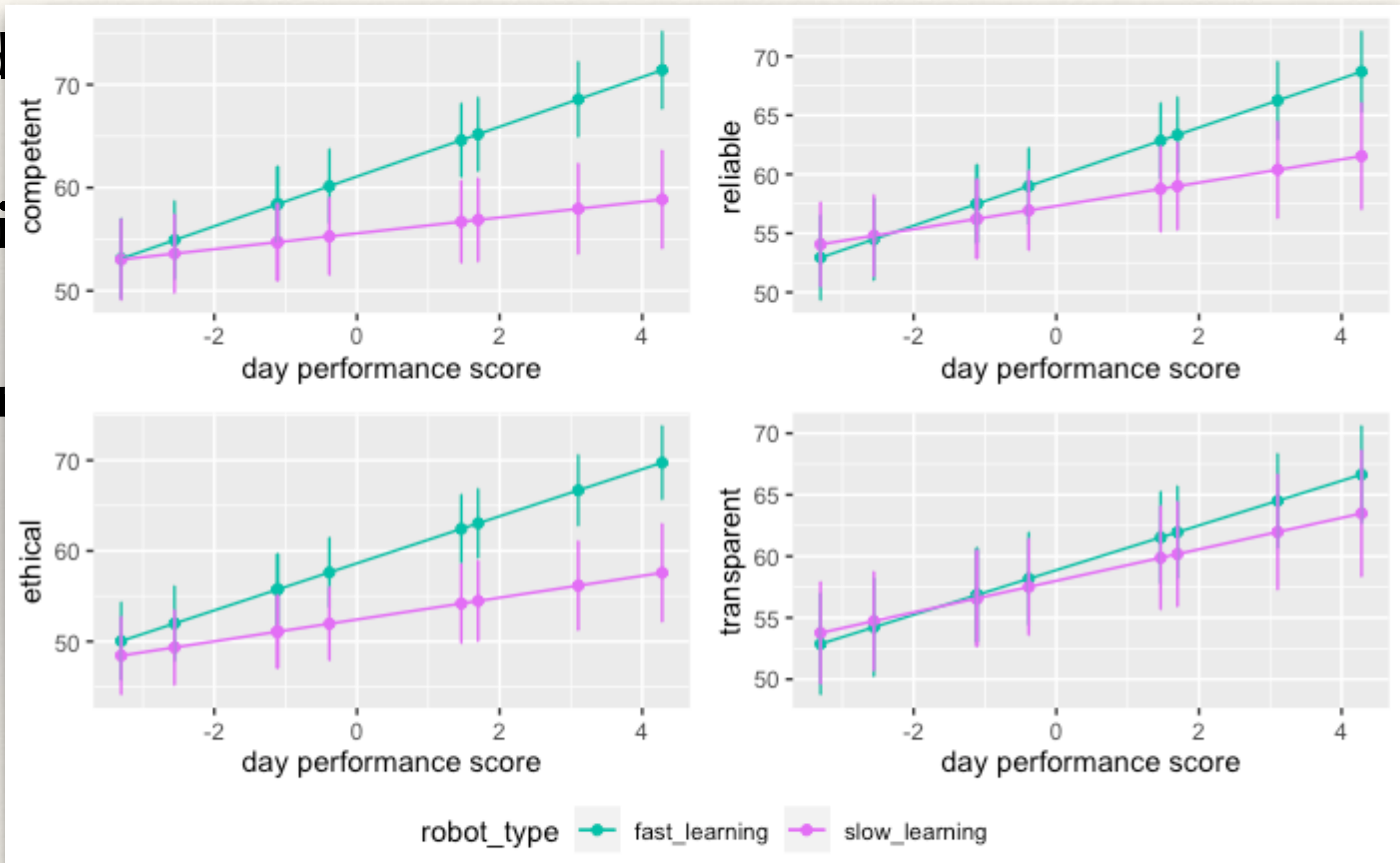
- for humans, robots,

Trust responds to justification

- for humans, robots

Trust tracks teaching

- very well calibrated



Trust Studies: Major Results

Trust responds to evidence specific to 2 factors and 5 dimensions

- ▶ for humans, robots, AI

Trust responds to justifications even under moral disagreement

- ▶ for humans, robots

Trust tracks teaching robots

- ▶ very well calibrated to local and cumulative performance, task difficulty

20% decline to judge robots on moral trustworthiness

- ▶ the simpler they look (Chita-Tegmark et al., 2021)
- ▶ the simpler they are

Questions

1. Does it even make sense to treat LLMs as “agents” that “have trustworthiness”?
2. If we do query LLMs’ trustworthiness, what specific moral attributes of trustworthiness would we want them to exhibit?
3. What would it take to design LLMs that actually have those attributes of moral trustworthiness?

Questions

1. Does it even make sense to treat LLMs as “agents” that “have trustworthiness”?

Critical point: Whether people do

Program of research: Conditions under which they do

Examples: Communication modality and style, empathic language, training history, owner, social value or purpose, interaction history

Questions

Ethical	Transparent	Benevolent
---------	-------------	------------

2. If we do query LLMs' trustworthiness, what specific moral **Proposal:** The 3 moral dimensions attributes of trustworthiness would we want them to exhibit? (possibly also social-moral virtues like patience, modesty)

Program of research: Conditions that “require” each dimension

Examples: User's goals, attitudes, assumptions; user's vulnerabilities; role and context for both; system capacities

Questions

3. V
a

Factors Dimensions attributes	Performance		Moral		
	Reliable	Competent	Ethical	Transparent	Benevolent
	reliable dependable consistent predictable	competent skilled capable meticulous	ethical principled moral has integrity	transparent candid genuine sincere	benevolent has goodwill kind considerate

Factors Dimensions attributes	Performance		Moral		
	Reliable	Competent	Ethical	Transparent	Benevolent
	reliable dependable consistent predictable	competent skilled capable meticulous	ethical principled moral has integrity	transparent candid genuine sincere	benevolent has goodwill kind considerate

Ethical dimension: Norm competence

- represent community-specific, context-specific, graded norms
- update norms from ongoing feedback

Transparent dimension:

- Self-monitoring, explaining, truthfulness

Benevolent dimension:

- Social manners, goals, perspective taking, empathy

LLMs???

Thank you

For references, see bit.ly/scsrl
or email **bfmalle@brown.edu**

Questions

1. Does it even make sense to treat LLMs as “agents” that “have trustworthiness”?
2. If we do query LLMs’ trustworthiness, what specific moral attributes of trustworthiness would we want them to exhibit?
3. What would it take to design LLMs that actually have those attributes of moral trustworthiness?

Computational Framework

Be

On

