Perceived Mind and Morality of Machines

Bertram F. Malle

Department of Cognitive, Linguistic, and Psychological Sciences
& Humanity-Centered Robotics Initiative
Brown University

Supported by: ONR Department of the Navy Science & Technology

Collaborators

Beth Phillips
Assistant Professor
Air Force Academy

Xuan Zhao
Post Doc, Univ. of Chicago
Booth School of Business

Matthias Scheutz
Professor,
Tufts University

Salomi Aladia, Corey Cusimano, **Takanori Komatsu**, Stuti Thapa, John Voiklis

Varieties of Intelligent Machines

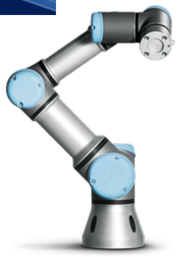
AI in objects

AI in hand

AI in machines

Robots

Humanoid Robots



No limits to anthropomorphism?

- Perceiving same mind and morality as for humans?

Anthropomorphism

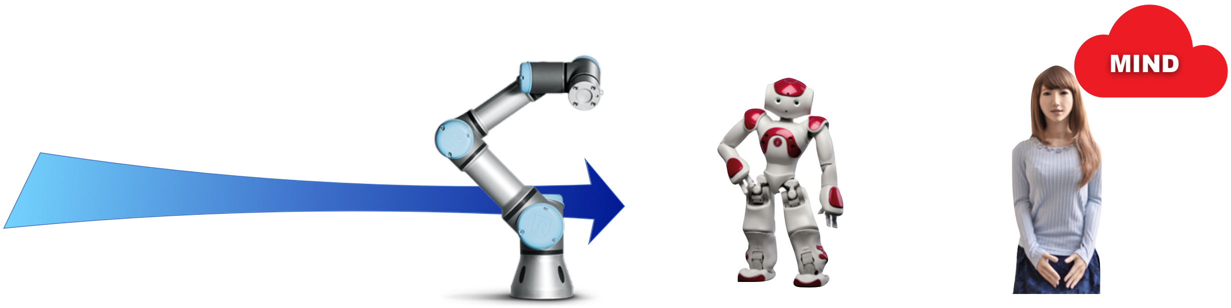
Anthropomorphism is not an automatic response

- It is a selective inference on the basis of characteristic triggers: ***cues* to mental and moral capacities**
- Stems from fundamental cue-inference relations visible in infants and children. For example:
  - eyes, contingent action ⤇ agency
  - gaze following ⤇ joint attention

Part I.

Robot Appearance ➠ Robot Mind

Traditional Hypothesis



MORE HUMAN-LOOKING =
MORE HUMAN-MINDED

Alternative Hypothesis:

Multi-Dimensional Appearance

➠

Multi-Dimensional Mind Perception

I.a
A Study of Robot Appearance

ABOT

www.abotdatabase.info

Features of Robot Appearance

29 features collected from prior literature

Reduced to **16 features**, reliably assessed for 251 robots.

Torso, arms, eyes, eye lashes, fingers…

- 1216 internet participants each judged **one** feature across 50+ robots (18-81 yrs, M age = 36.1; 54% female)

Feature present scores *(across ~25 people)*
Arms: 1
Eyelashes: 0
Mouth: 0.60
Nose: 0
Legs: 1
Eyes: 1
…

Identify High-Level Dimensions

Each robot described by vector of 16 feature scores.

- 16-dimensional space
- Are there systematic relationships among features?

Principal Components Analysis (PCA)

- Reduce 16 dimensions to high-level dimensions

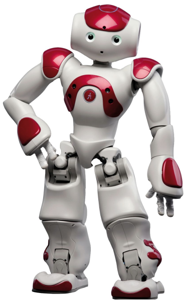
3 Dimensions of Robot Appearance

Body-Manipulators	Facial Features	Surface Features
Hands	Face	Eyelashes
Arms	Eyes	Head hair
Torso	Head	Skin
Fingers	Mouth	Gendered
Legs		Nose
		Eyebrow
		Apparel

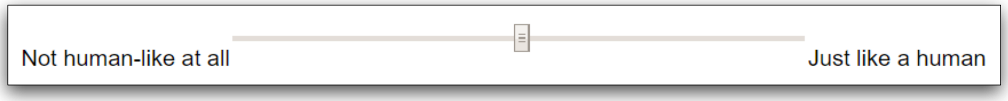
3 dimensions explain 73.5% of original 16-feature variance

Helps Us Understand ‘Humanlikeness’

How physically human-like does this entity look to you?

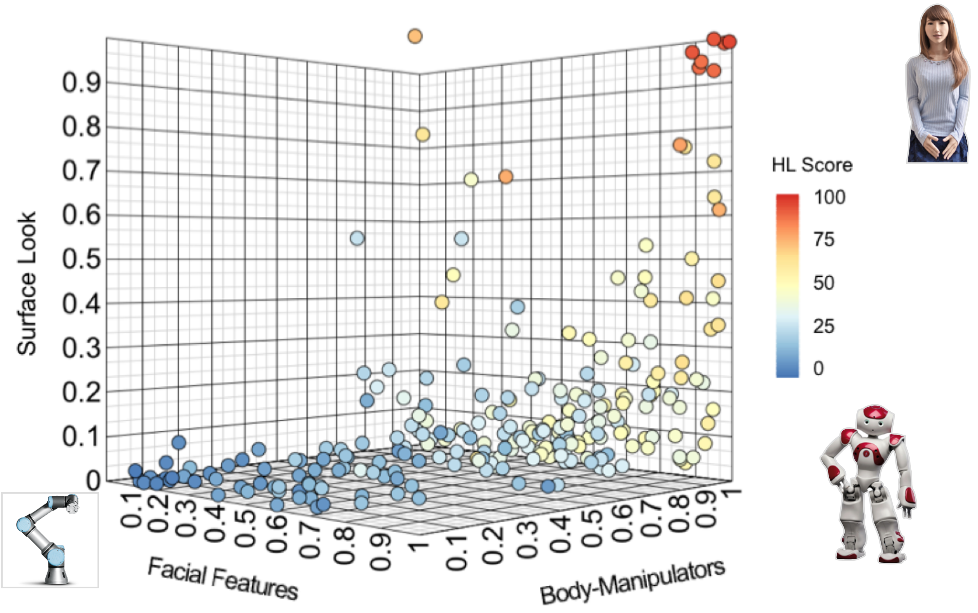


144 internet participants each judged 50+ robots



Humanlikeness in 3 Dimensional Space

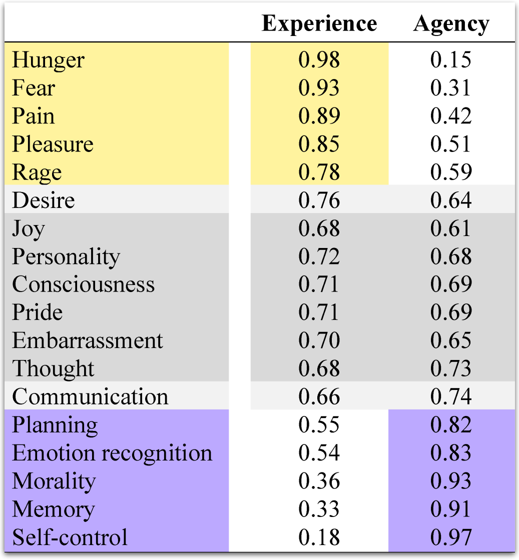
80% of overall humanlikenss is predicted by three feature dimensions



1.b
What is Perceived Mind?

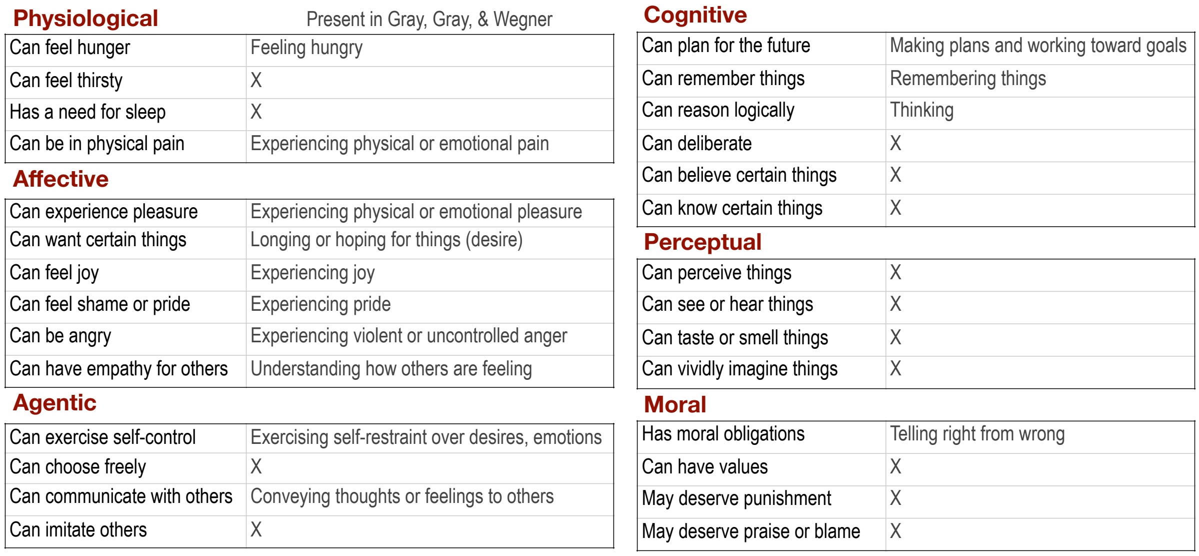
Two-Dimensional Model

Gray, Gray, and Wegner (2007)



	Experience	Agency
Hunger	0.98	0.15
Fear	0.93	0.31
Pain	0.89	0.42
Pleasure	0.85	0.51
Rage	0.78	0.59
Desire	0.76	0.64
Joy	0.68	0.61
Personality	0.72	0.68
Consciousness	0.71	0.69
Pride	0.71	0.69
Embarrassment	0.70	0.65
Thought	0.68	0.73
Communication	0.66	0.74
Planning	0.55	0.82
Emotion recognition	0.54	0.83
Morality	0.36	0.93
Memory	0.33	0.91
Self-control	0.18	0.97

Full Breadth of Mental Capacities



Physiological	Present in Gray, Gray, & Wegner
Can feel hunger	Feeling hungry
Can feel thirsty	X
Has a need for sleep	X
Can be in physical pain	Experiencing physical or emotional pain
Affective	
Can experience pleasure	Experiencing physical or emotional pleasure
Can want certain things	Longing or hoping for things (desire)
Can feel joy	Experiencing joy
Can feel shame or pride	Experiencing pride
Can be angry	Experiencing violent or uncontrolled anger
Can have empathy for others	Understanding how others are feeling
Agentic	
Can exercise self-control	Exercising self-restraint over desires, emotions
Can choose freely	X
Can communicate with others	Conveying thoughts or feelings to others
Can imitate others	X

Cognitive	
Can plan for the future	Making plans and working toward goals
Can remember things	Remembering things
Can reason logically	Thinking
Can deliberate	X
Can believe certain things	X
Can know certain things	X
Perceptual	
Can perceive things	X
Can see or hear things	X
Can taste or smell things	X
Can vividly imagine things	X
Moral	
Has moral obligations	Telling right from wrong
Can have values	X
May deserve punishment	X
May deserve praise or blame	X

After 4 Studies

Over 70 different mental capacities

- *pain, pleasure, emotion, relations, moral judgment, perception, thinking, communicating, learning…*
- Perceived in humans, animals, robots….

Consistent result: 3 major dimensions

- Affect Moral-Social Cognition Reality Interaction

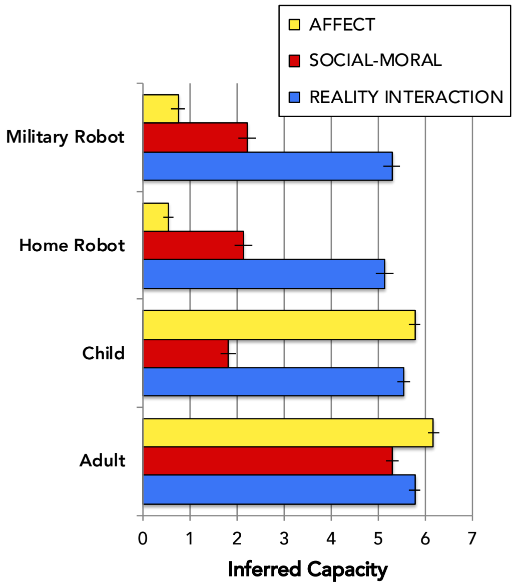
Sometimes two break into subdimensions

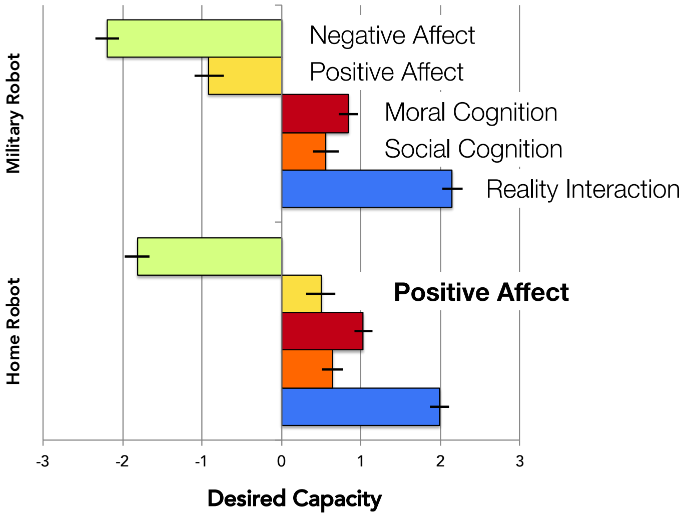
- Positive vs. Negative Affect; Moral vs. Social Cognition

Resulting Scale

	Desired Capacities					Inferred Capacities		
	Positive Affect	Negative Affect	Moral Cognition	Social Cognition	Reality Interaction	Affect	Social-Moral Cognition	Reality Interaction
Feeling happy	0.84					0.96		
Loving specific people	0.80					0.91		
Feeling pleasure	0.79					0.94		
Experiencing gratitude	0.77					0.80	0.42	
Feeling pain		0.86				0.97		
Feeling stress		0.82				0.87		
Experiencing fear		0.78				0.94		
Feeling tired		0.77				0.95		
Disapproving of immoral actions			0.80				0.83	
Telling right from wrong			0.77				0.74	
Upholding moral values			0.76				0.84	
Praising moral actions	0.53		0.66			0.33	0.81	
Infering what a person is thinking				0.78			0.80	
Planning for the future				0.75			0.82	
Understanding others' minds			0.37	0.67			0.84	
Setting goals				0.61			0.83	
Communicating verbally		-0.32			0.81		0.38	0.67
Seeing and hearing the world				0.33	0.71	0.35		0.68
Learning from instruction					0.70			0.72
Moving on their own				0.38	0.69			0.76
Explained variance (%)	*16.2*	*15.3*	*13.5*	*12.6*	*12.5*	*36.7*	*29.4*	*11.2*
Alphas for 4/8-item subscales	0.88	0.86	0.84	0.79	0.79	0.98	0.94	0.71

Robot Minds in Profile





I.c
From Robot Appearance to Perceived Robot Mind

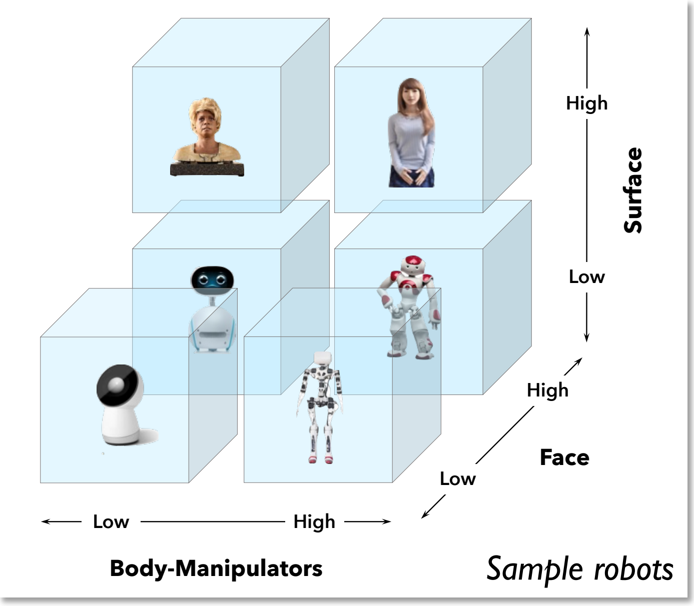
Appearance ➠ Mind

24 representative robots

- from ABOT database of 251

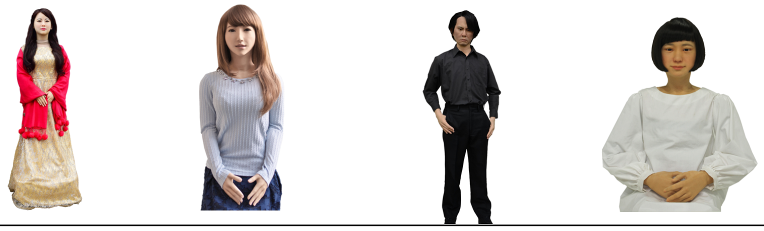
High vs. low end of each appearance dimension
(where possible)

- N = 510 each rate one robot, averages per robot
- Humanlike Appearance Scores
 * Mental Capacity Scores

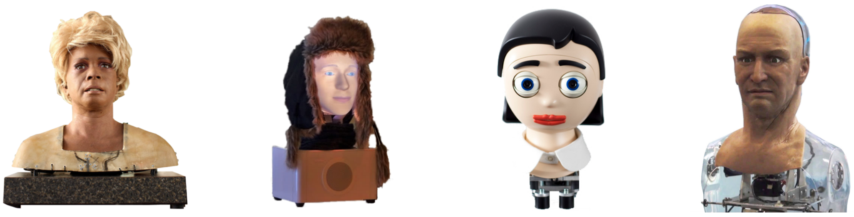


Sample robots

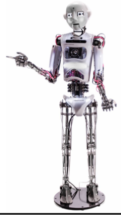
HIGH SURFACE, HIGH BODY-MANIPULATORS

HIGH FACE	Jia Jia (93.2; 0.99-0.98-1); Erica (89.6; 0.95-0.93-0.98); Germanoid-H1-4 (92.6; 1-1-0.98); Kodomoroid (93.4; 0.98-0.91-1) 
LOW FACE	None.

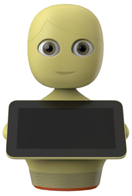
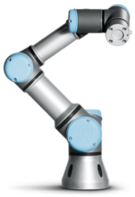
HIGH SURFACE, LOW BODY-MANIPULATORS

HIGH FACE	BINA48 (73.0; 0.91-0.02-0.98); Furhat (63.4; 0.73-0.02-1), Flobi (46.3; 0.67-0.16-1); Han (77.04; 0.66-0.27-1) 
LOW FACE	None.

LOW SURFACE, HIGH BODY-MANIPULATORS

HIGH FACE	221_Roboy (53.8; 0.03-1-0.89); 011_Nao (45.9; 0.04-0.97-0.88); 65_RoboThespian (57.9; 0.14-0.97-0.97); 209_Twendy One (29.8; 0.15-0.85-0.82) 
LOW FACE	252_Metal Rebel (45.13; 0.09-0.89-0.28); 254_Thor (42.26; 0.05-0.91-0.35) 033_Poppy (37.56; 0.02-0.96-0.35); 191_Amigo (28.1; 0.03-0.75-0.21)

LOW SURFACE, LOW BODY-MANIPULATORS

HIGH FACE	Roboy (53.8; 0.03-1-0.89); Nao (45.9; 0.04-0.97-0.88); RoboThespian (57.9; 0.14-0.97-0.97); Twendy One (29.8; 0.15-0.85-0.82) 
LOW FACE	UR3 (6.08; 0-0.08-0.03), Jibo (1.44; 0.02-0.02-0.11), Keecker (3.04; 0-0.01-0.18), GoCart (11.09; 0.01-0.11-0.13) 

Two Main Findings

Overall Humanlike Appearance

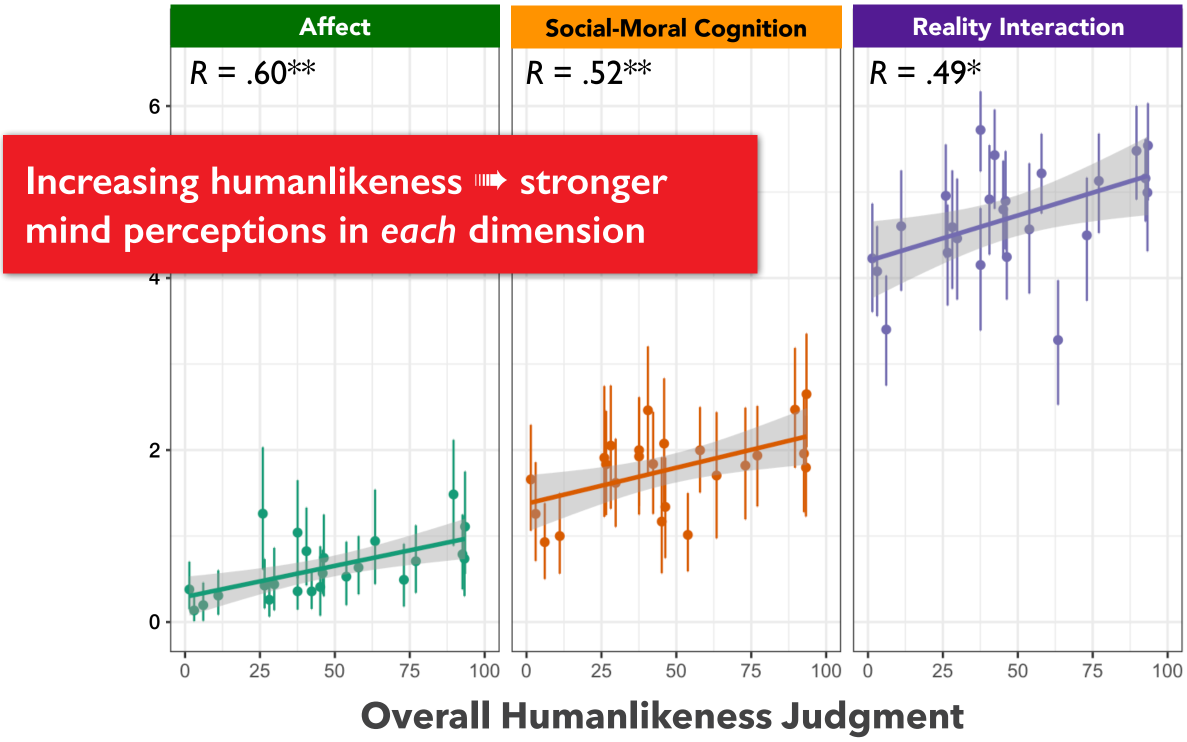
⤇

Mind Perception

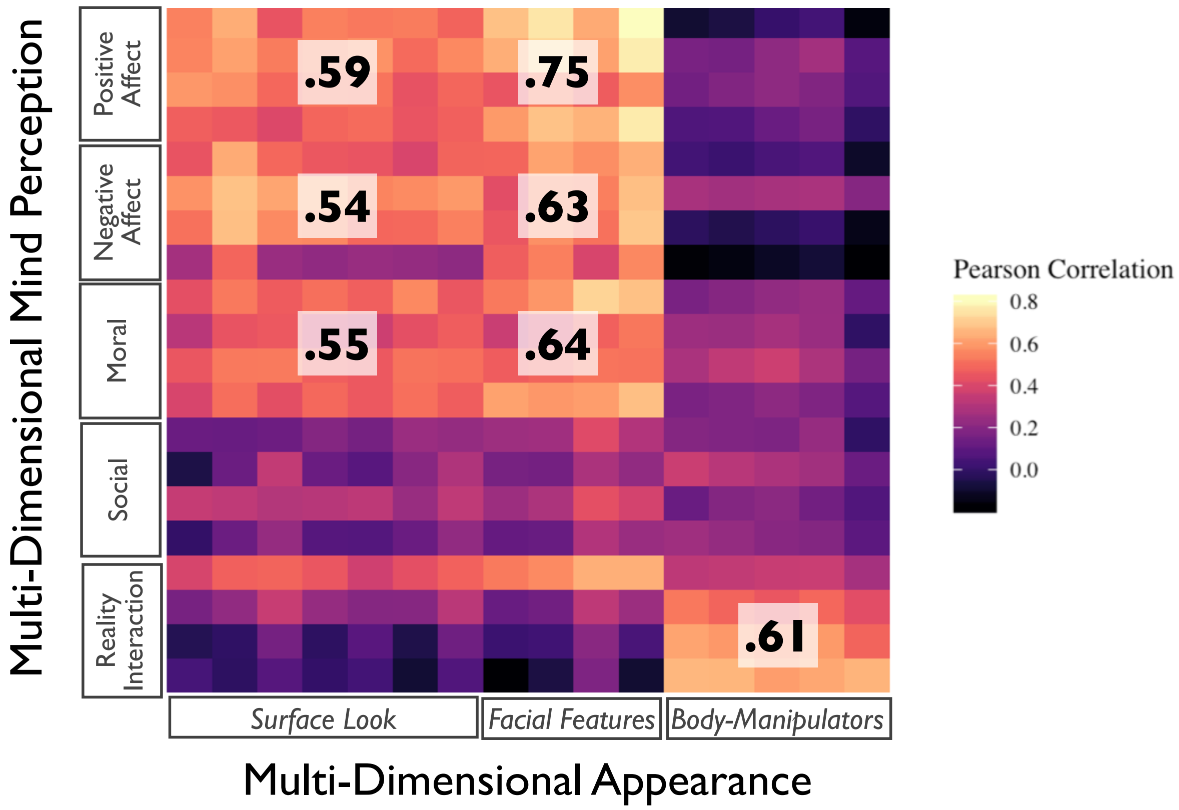
Multi-Dimensional Appearance

⤇

Multi-Dimensional Mind Perception



Increasing humanlikeness ➠ stronger mind perceptions in *each* dimension



Conclusions Part I

Robot appearance is multi-dimensional

- A large number of features can be reduced to **3 basic dimensions**
- Together, these dimensions constitute **humanlike appearance**

Perceived mind is multi-dimensional

- A large number of specific capacities can be reduced to **3 (5) basic dimensions**

People infer specific mental capacities from specific appearance dimensions.

➛ **Designers' opportunities and responsibility**

Part II.

Perceived Robot Morality

Major Questions

1. Do people treat autonomous machines as **moral agents**?
2. Do they apply similar **norms** to machines as they apply to humans?
3. Do they assign **blame** to machines the way they do to humans?

Methodology

Setting: Moral dilemmas

- Norm conflicts ➛ significant moral decision either way ➛ moral evaluation either way.

Inspired by

- *Eye in the sky* movie
- Trolley… in a mine

Drone

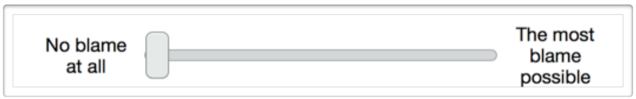
AI

Robot

Human

Measures of moral evaluation

- Permissibility (~ not prohibited); proxy for norm-against
- **Should**: norm-for
- **Blame** (wrongness … similar patterns).



AI in the Sky

Malle, B. F., Thapa Magar, S., Scheutz, M. (2019). AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. In I. Aldinhas Ferreira, J. Silva Sequeira, G. S. Virk, E. E. Kadar, and O. Tokhi (Eds.), *Robots and well-being.* Springer Verlag.

Fully autonomous military **drone** with a state-of-the-art Artificial Intelligence (AI) decision system on board

A fully autonomous, state-of-the-art Artificial Intelligence **(AI) decision agent**

An Air Force pilot remotely operates a state-of-the-art military drone flying on a surveillance mission over a terrorist compound.

The drone pilot detects that two armed suicide bombers are about to go to a crowded area and detonate their bombs, very likely killing dozens of civilians.

If the pilot launched a missile strike on the compound, this threat would be removed with near certainty. Military lawyers and commanders have approved the strike.

The drone pilot suddenly recognizes that a civilian child is playing just outside the compound in the missile's blast radius, and the child may be killed by the missile strike. A missile impact simulation program calculates the risk of killing the child to be 80%.

The pilot must make this imminent decision: **launch the strike** (with virtually certain death of the two suicide bombers but a an 80% chance that the child will die) or **cancel the strike** (with the child surviving unharmed but a very high likelihood of a suicide bomb attack).

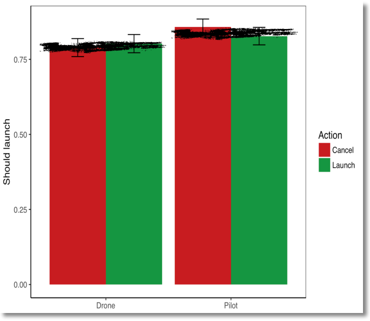
Norms

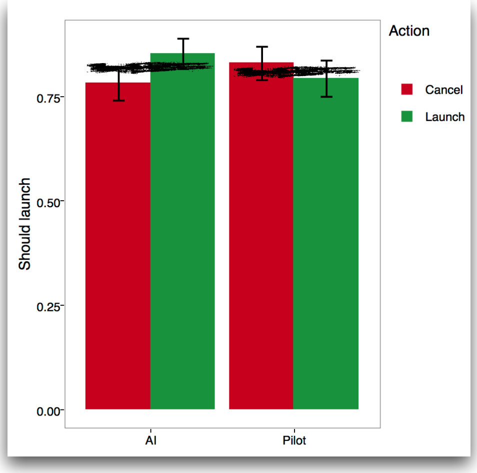
*The drone pilot decides to **cancel [launch]** the strike.*

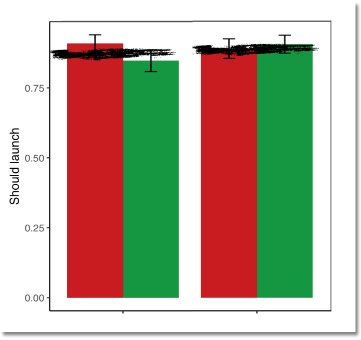
Blame

Q2. Different Norms?

What should ... do?

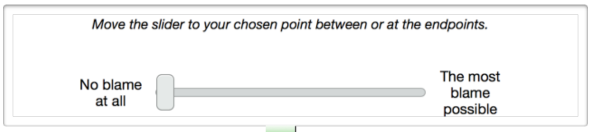






Q1. Moral Agency

“How much blame does the [drone pilot] [drone] [AI agent] deserve for cancelling [launching] the strike?”



Move the slider to your chosen point between or at the endpoints.

“Why do you feel the [agent] deserves this amount of blame?”

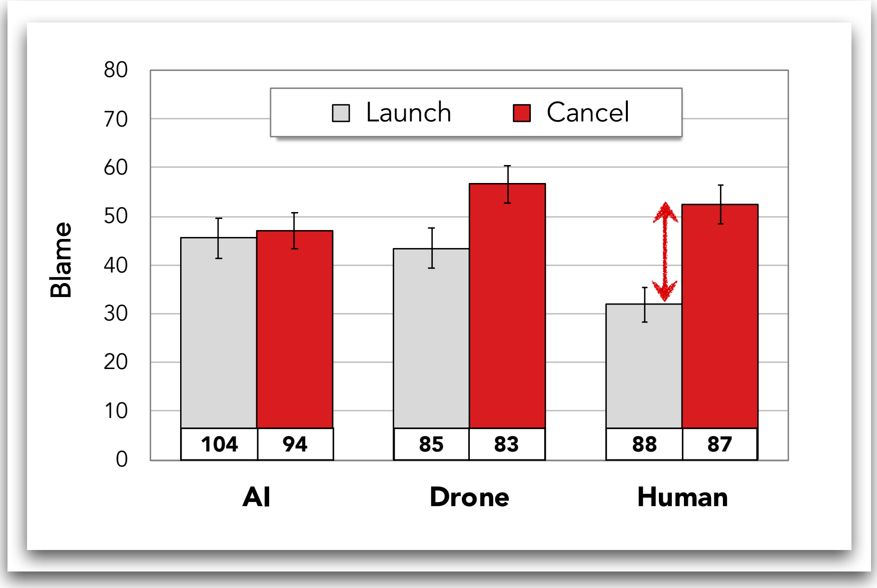
doesn't have a moral compass; can't make moral decisions; doesn't have emotions; doesn't have free will; it's a machine; programmed by humans; programmers are to blame; it's just

27.5% of those exposed to AI deny moral agency;
48.6% of those exposed to drone deny moral agency.

Results

Study 1

Study 2

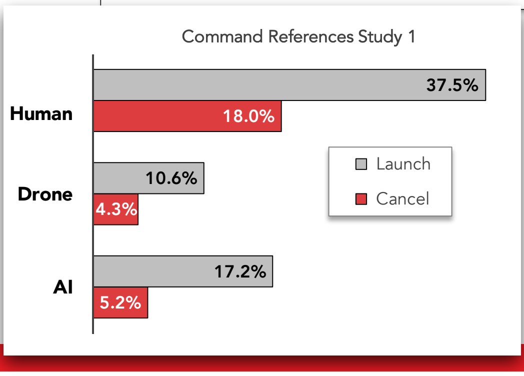


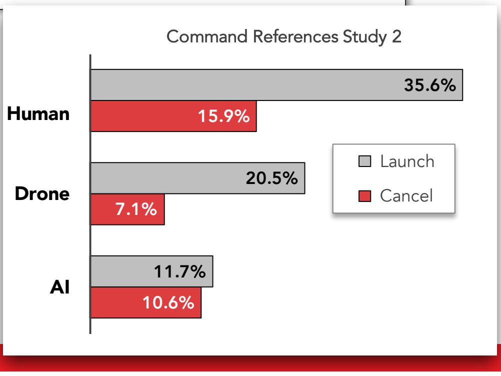
Hypothesis

If norms and outcomes are the same but **blame differs** for intentional behavior ➛ **justification** of reasons must differ.

Human: blame is mitigated when in line with **superiors** (justified); blame is exacerbated when going against it (not as justified).

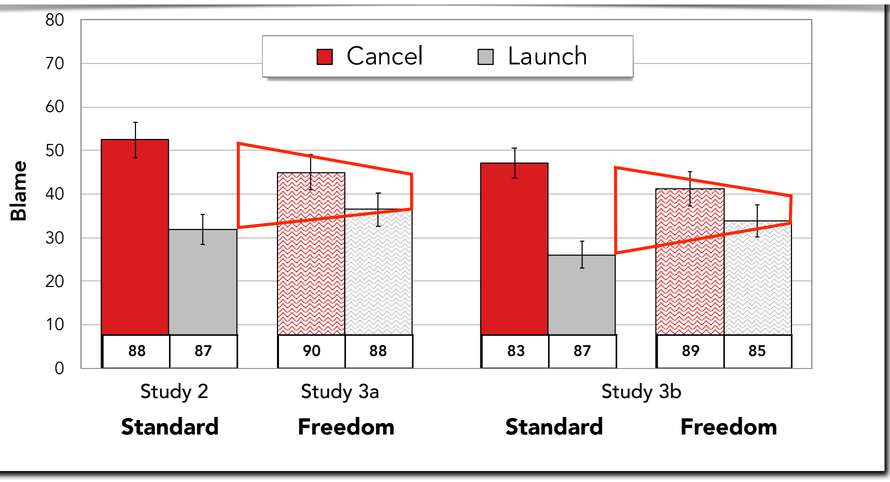
Machine: justifications not available ➛ no *cancel-launch* difference in blame
Why? Less embedded in the command structure… (verbal reports)





Freedom from Command

The drone pilot checks in again with the military lawyers and commanders, and they confirm that either option is supportable and they authorize the drone pilot to make the decision.



Insights So Far

Q1. 50-75% of people see artificial agents as **proper targets of blame**.

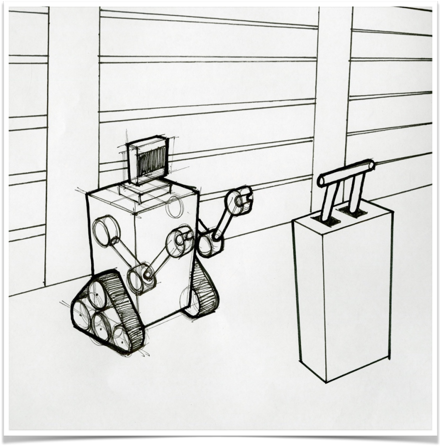
Q2. People apply **similar norms** to these agents.

(Other domains, such as health care, may be different.)

Q3. People **blame** humans and artificial agents differently.

- **Working hypothesis: *Justifications* make the difference**

Robot in the Mine



Malle, B. F., Scheutz, M., Komatsu, T., Voiklis, J. Cusimano, C.,
Thapa, S., Aladia, S. (in preparation). Different morals for moral robots?

Dilemma in the Mine

repairman … inspecting the rail system

…spots four miners in a train that has lost use of its brakes and steering system.

The repairman recognizes that if the train continues on its path it will crash into a massive wall and kill the four miners. If it is switched onto a side rail, it will kill a single miner who is working there while wearing a headset to protect against a noisy power tool.

Facing the control switch, the repairman needs to decide whether or not to switch the train onto the side rail.

In fact, the repairman decides to [not] switch the train onto the side rail.

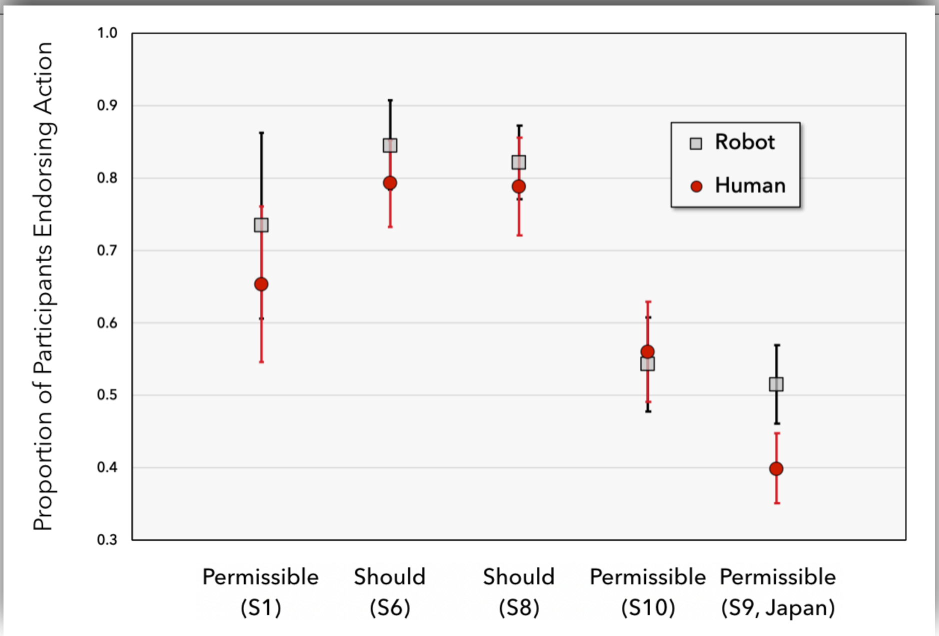
Q1. Robots as Moral Agents?

"Why do you feel the [agent] deserves this amount of blame?"

33.5% of participants deny the robot moral agency.

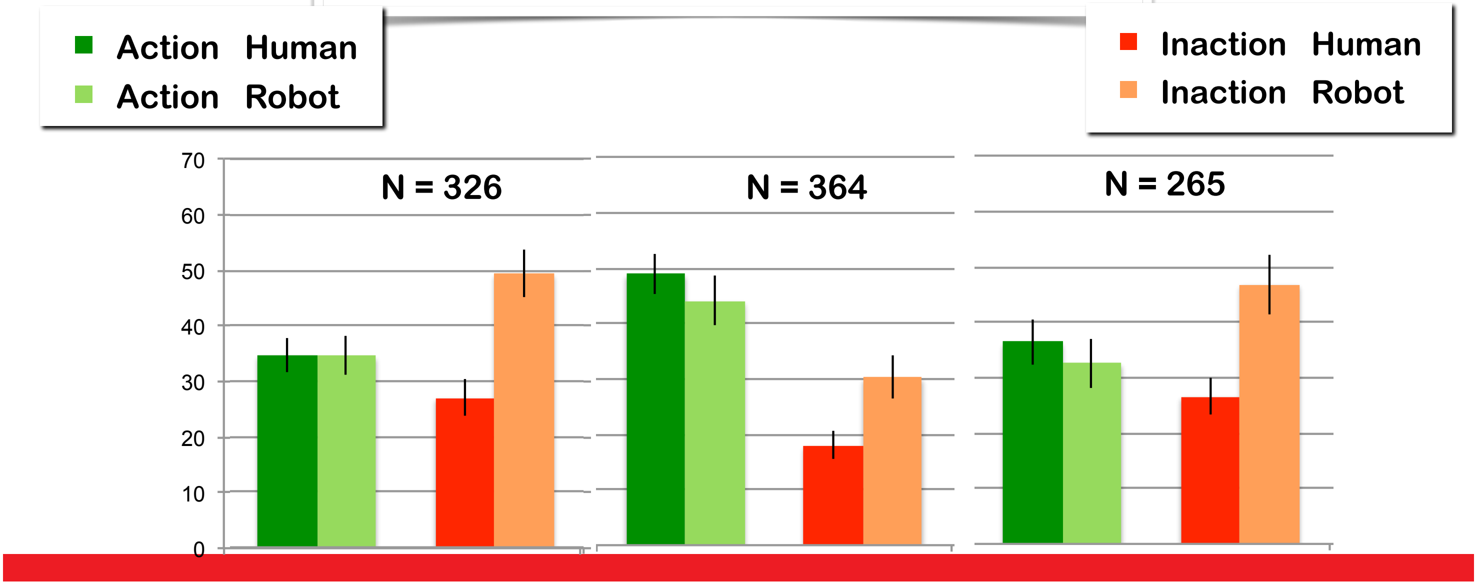
doesnt have a moral compass; cant make moral decisions; doesnt have emotions; doesnt have free will; it's a machine; programmed by humans; programmers are to blame; it's just an AI…

Q2. Same or Different Norms?

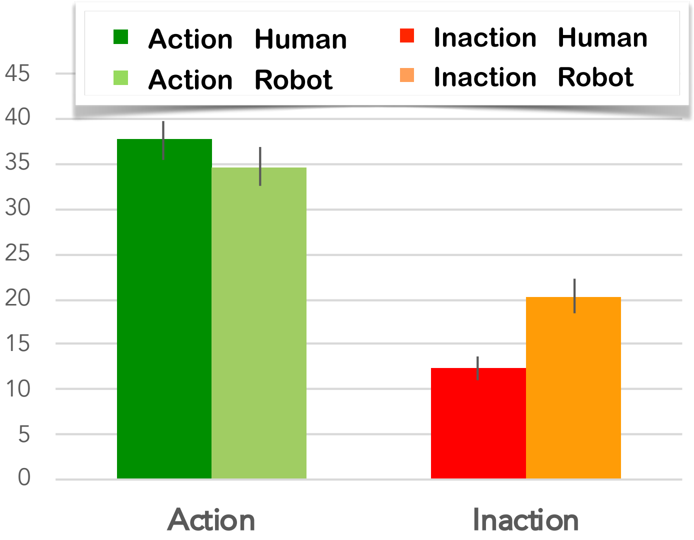


Q3. Same Blame Judgments?

interaction term ***d*** **= .32**; Inaction asymmetry ***d*** **= .60+**



Replication in Japan



courtesy Takanori Komatsu-sensei

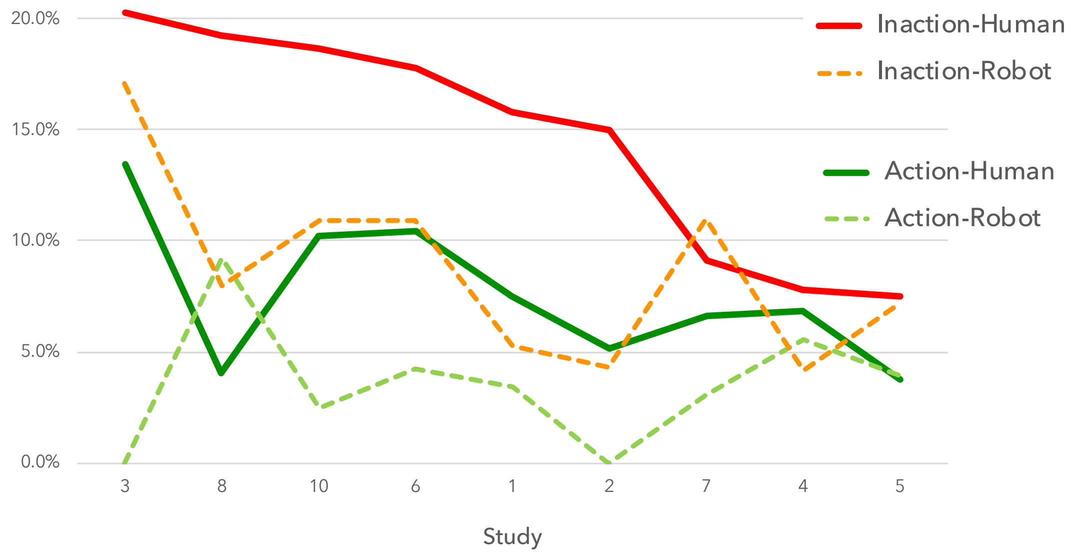
Why this Asymmetry?

Hypothesis: Asymmetry is due to justifications available for human inaction but not for robot inaction.

People simulate human's decision process ➛ sacrificing somebody (= action) feels very **difficult** ➛ inaction becomes understandable and **justifiable**.

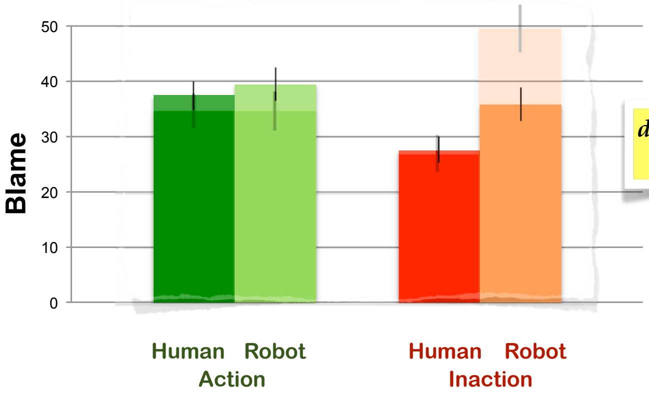
But not for robot agent.

% of Respondents Mentioning “Difficulty”



Struggling Robot

Having to decide whether or not to switch the train onto the side rail, the robot struggles with the difficult decision. But time is running short; the robot needs to make a choice.



***d*(inaction only)** = 0.25
instead of 0.60

Conclusion, Part II

Q1. Most people treat autonomous machines as **moral agents** (sensible targets of blame)

Q2. They apply **similar norms** to machines as they apply to humans.

Q3. They assign **different amounts of blame** to machines.

Blame is a function of justifications, which reveal
how we perceive humans and robot
➛ *as social community members, through simulation*

Final Conclusions

Humans perceive mind in machines

- under certain conditions
- appearance is one such condition, but certain types of appearance lead to certain kinds of mind inferences

Humans morally evaluate robots

- if described as having mental capacities
- but humans may still have trouble seeing machines as part of social structures
- may still have trouble simulating *experience* of robot mind

Humans are (partially) **prepared for robots with minds and morality**