수동적 객체이자 능동적 주체로서의 로봇에 대한 심리학적 이해

2025년 2월 12일 KRoS
소셜로봇연구회
김보영
안보정책연구소
한국 조지메이슨 대학교

안녕하세요! 김보영이라고 합니다.

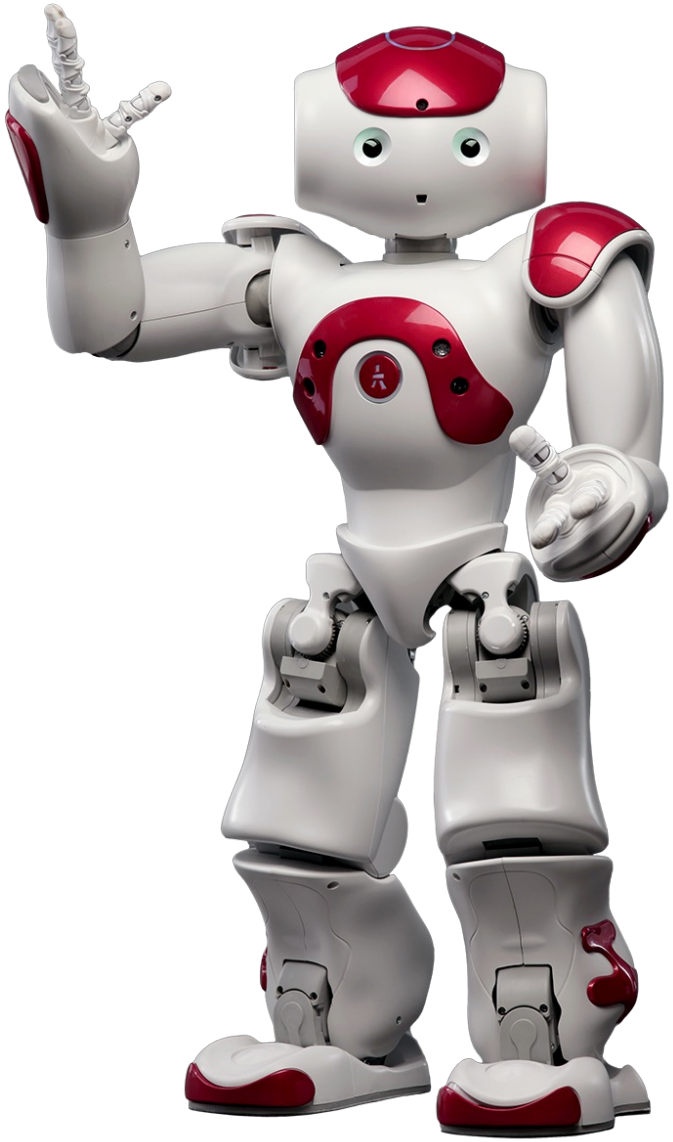
인지과학, 사회심리학, 도덕심리학에서
인간-로봇 상호작용과 기술정책까지

개요

1. 신기하고 으스스한 것
2. 똑똑하고 행동하는 너
3. 로봇의 주체성에 대한 두 가지 관점 그리고 책임의 문제

1. 신기하고 으스스한 것

처음 직접 본 로봇

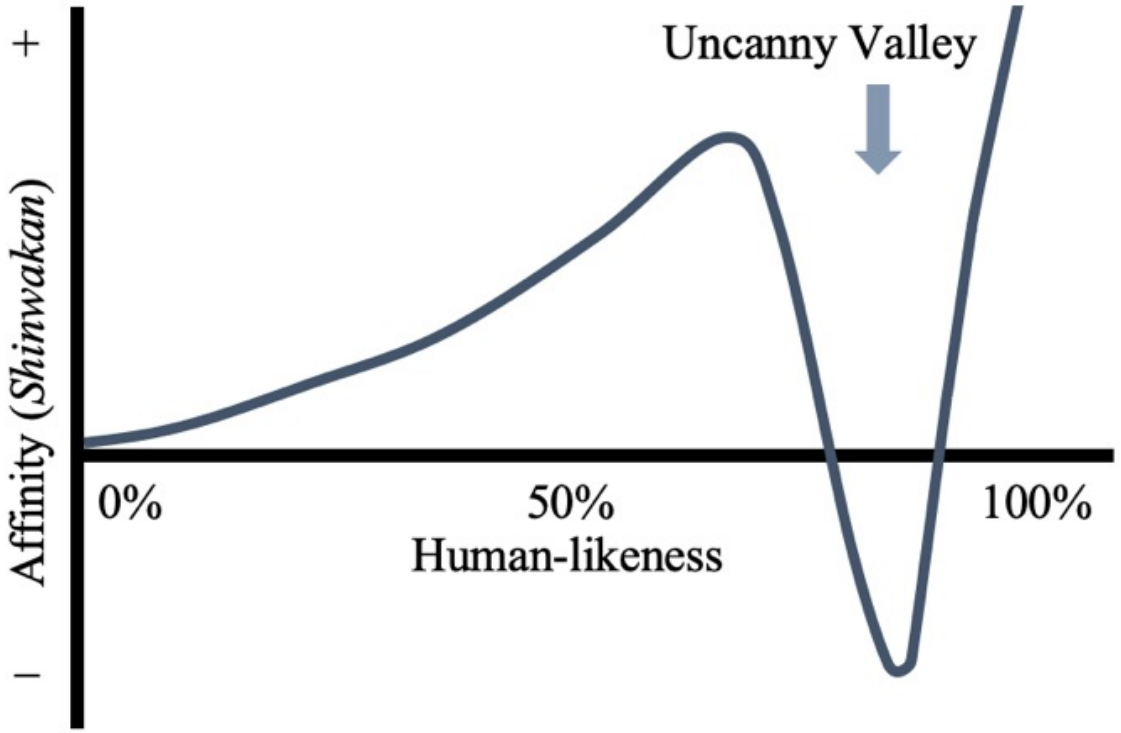


신기하고 기묘한 것

- 기술과 능력 측면에서 실제 로봇에 대해 "이상하고 이중적인(Weird and Double-minded)" 감정 반응을 보임.
- 기대에는 못 미치지만 많은 관심을 보이고 상호작용하고 싶어함.

Bruckenberger et al. (2013)

불쾌한 골짜기



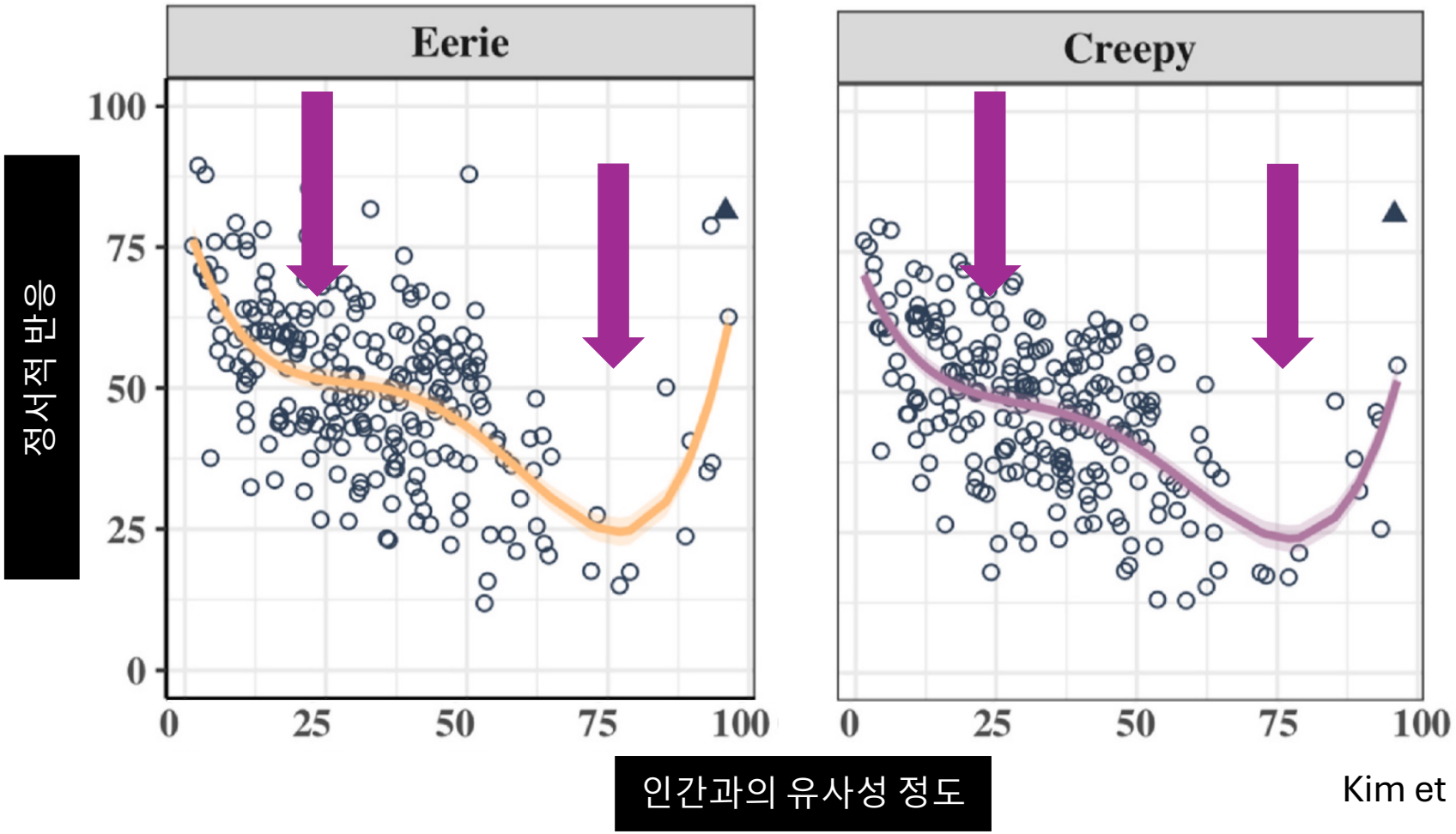
Mori (1970; 2012)

실재한 251개의 로봇

How **eerie** it is?
How **creepy** it is?

Kim et al. (2022)

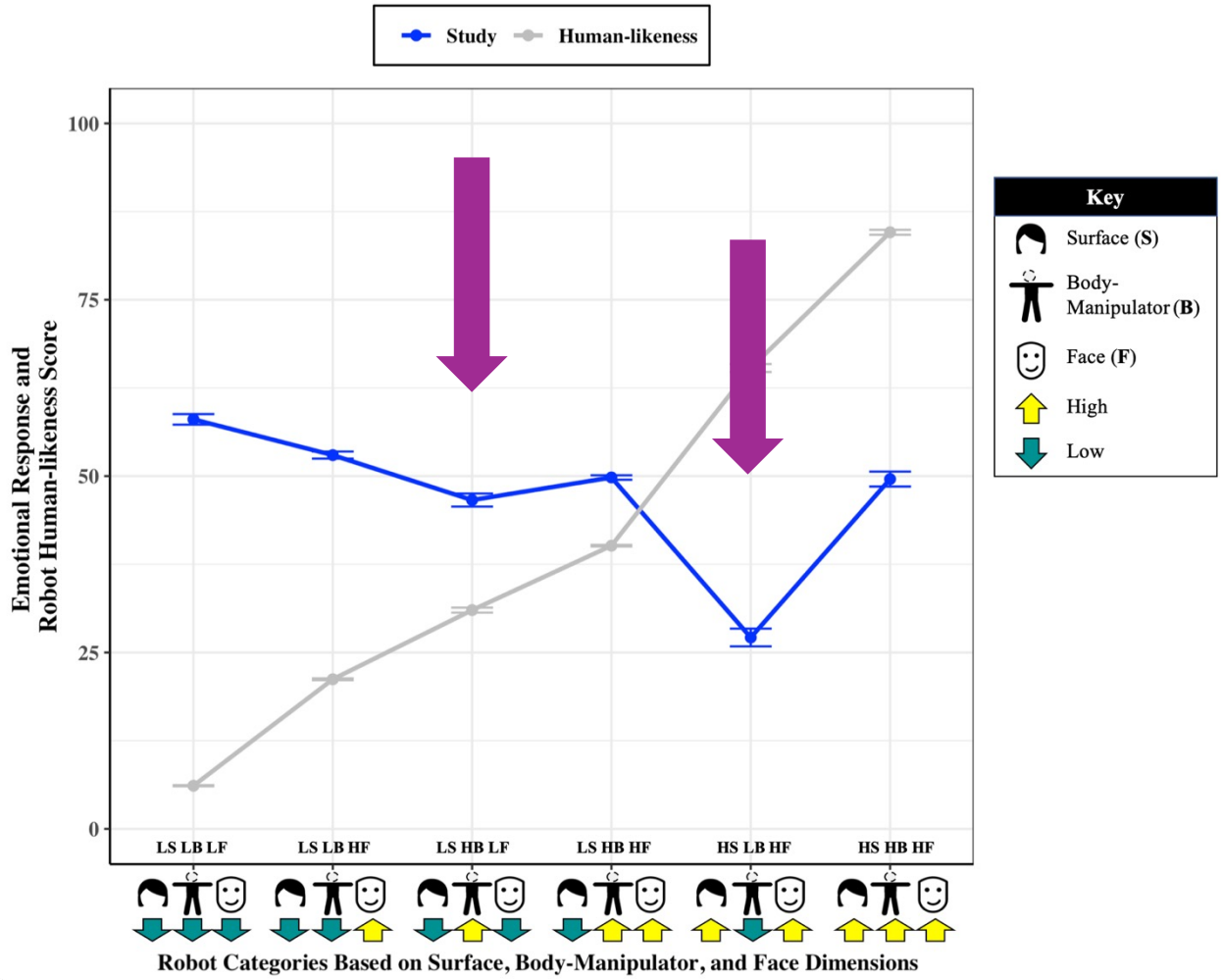
두 개의 불쾌한 골짜기



Kim et al. (2022)

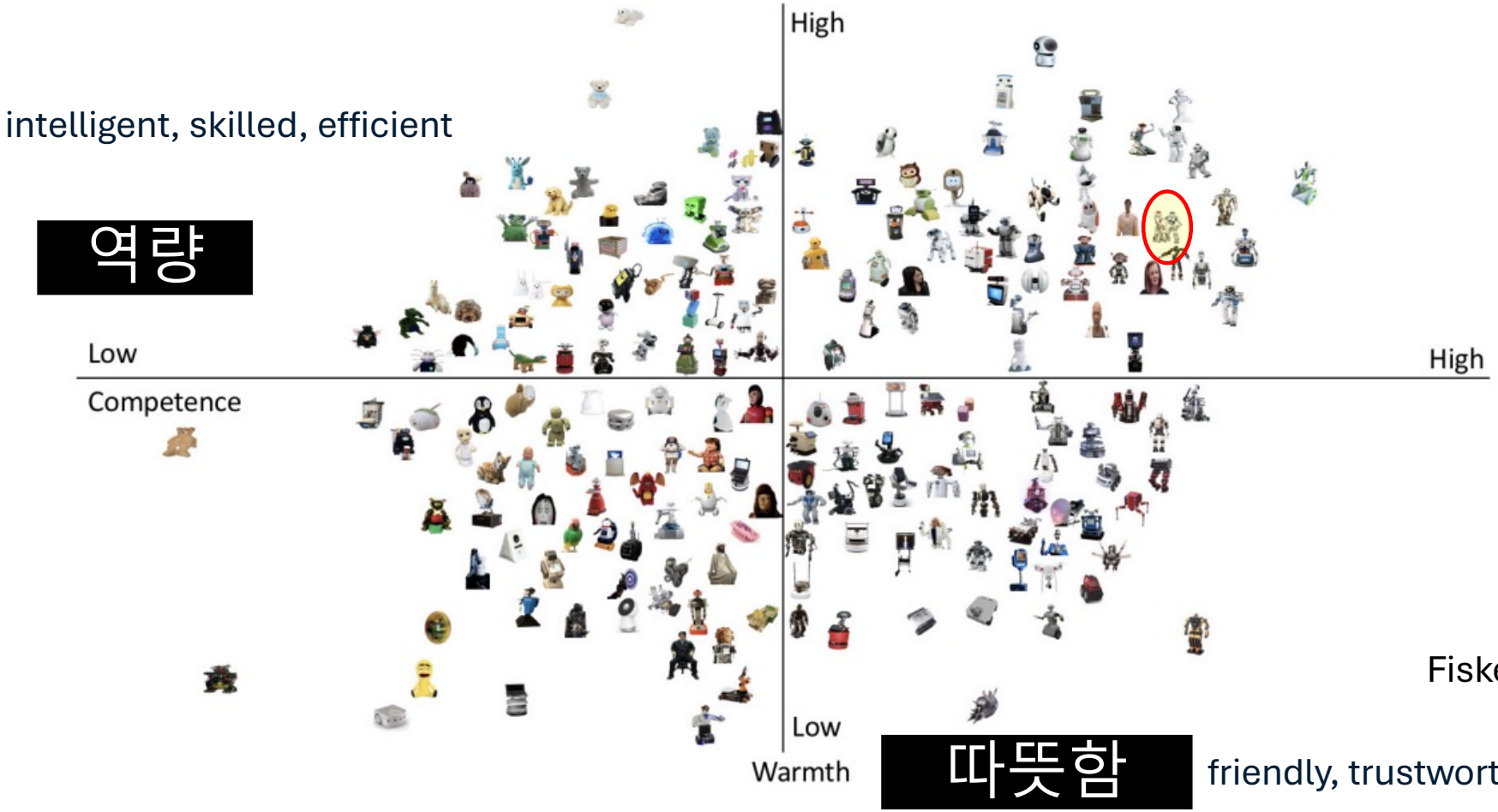
이러한 사람들의 반응을 설명할 수 있는 심리학적 기제는 무엇일까요?

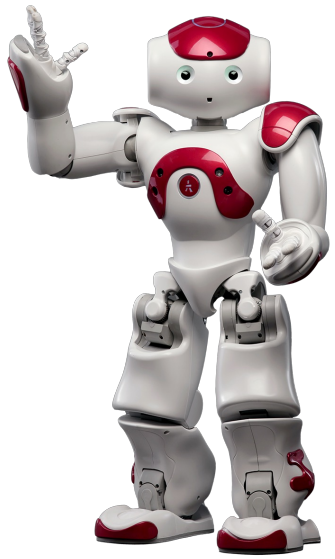
얼굴과 몸의 인간유사도 불일치



Kim et al. (2022)

처음 보는 사람/로봇에 대한 반응
(사회적/존재론적 집단)





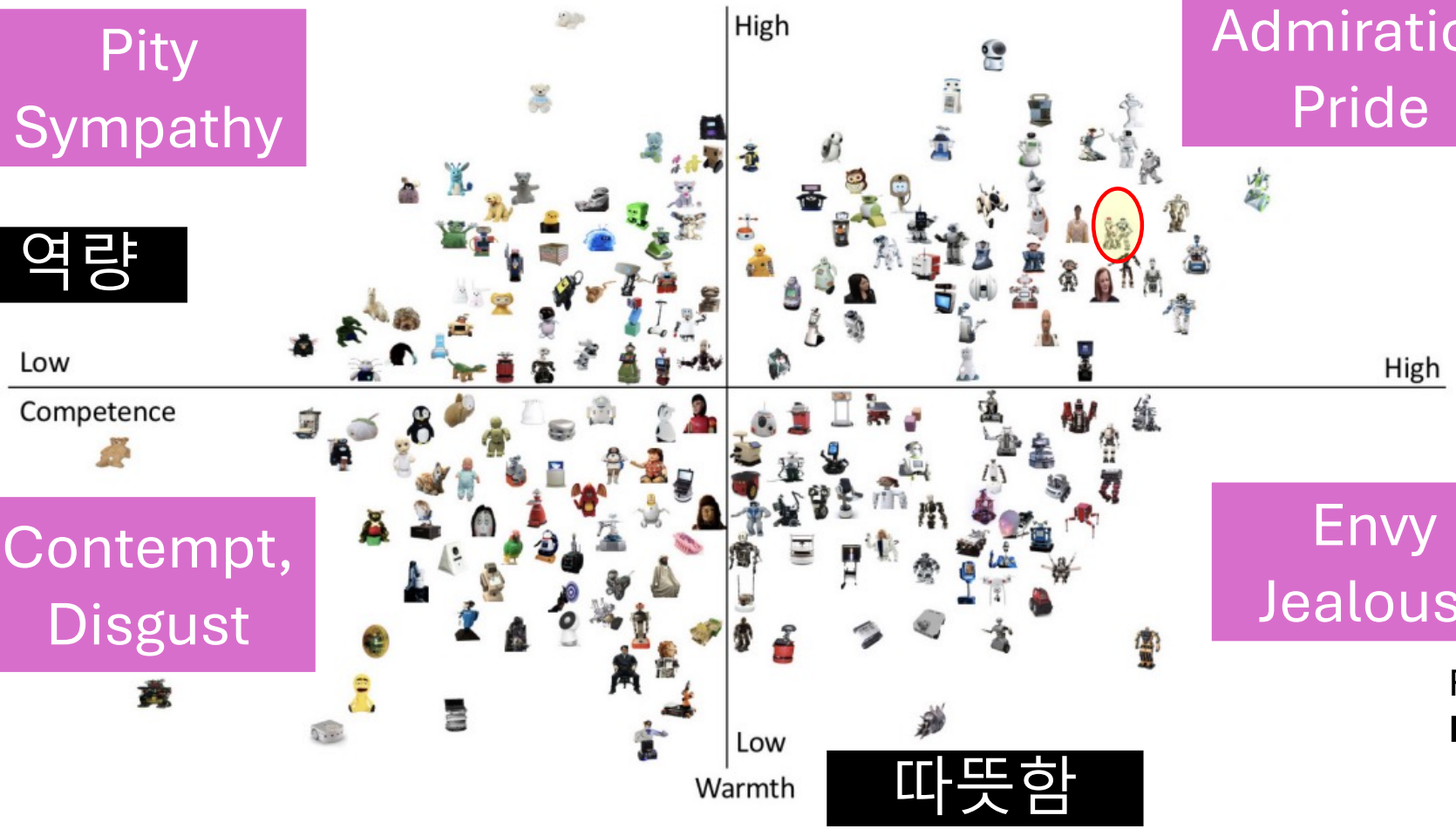
Fiske et al. (2002; 2007)

처음 보는 사람/로봇에 대한 반응 (사회적/존재론적 집단)

Pity
Sympathy

Admiration
Pride

역량



Contempt,
Disgust

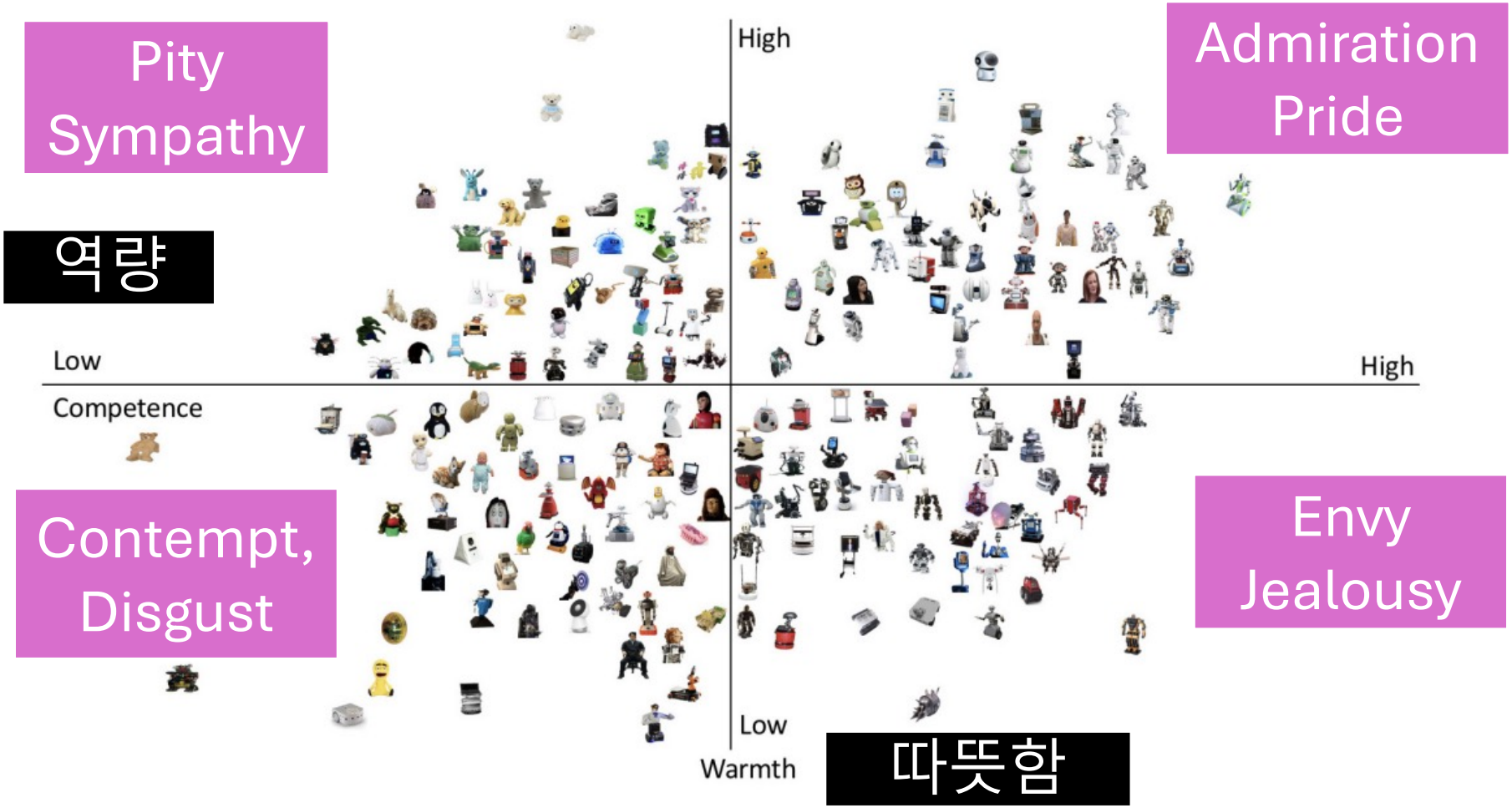
Envy
Jealousy

따뜻함

Fiske et al. (2007)
Reeves et al. (2020)

여러분이 직접 처음 본 로봇은 어디에 해당되었나요?

처음 보는 사람/로봇에 대한 반응 (사회적/존재론적 집단)



2. 똑똑하고 움직이고 말도 하는 너


#HomeOf
SocialRobotics
Robotics

At UN, robot Sophia joins meeting on artificial intelligence and sustainable development

우리의 생각과 행동을 (긍정적 혹은 부정적 방향으로) 변화시키는 로봇

- 투자결정과 자산관리를 위한 로봇의 조언
  Belanche et al. (2019)
  D'Acunto et al. (2020)

- 인지능력을 요하는 퀴즈를 풀기 위한 로봇의 조언
  Saunderson et al. (2021)

- 로봇의 도덕적 영향력 "말의 힘"
  Malle et al. (2015)
  Jackson & Williams (2022)
  Banks et al. (2023)
  Hanson et al. (2024)

- **설득의 전략과 그 영향을 연구할 필요성**

긍정적 영향을 주는 로봇

- 사람들에게 긍정적인 영향을 미칠 수 있는 로봇

- 친사회적 행동 유도

(예: 기부 행위)

설득의 전략

전략 1. 겉모습

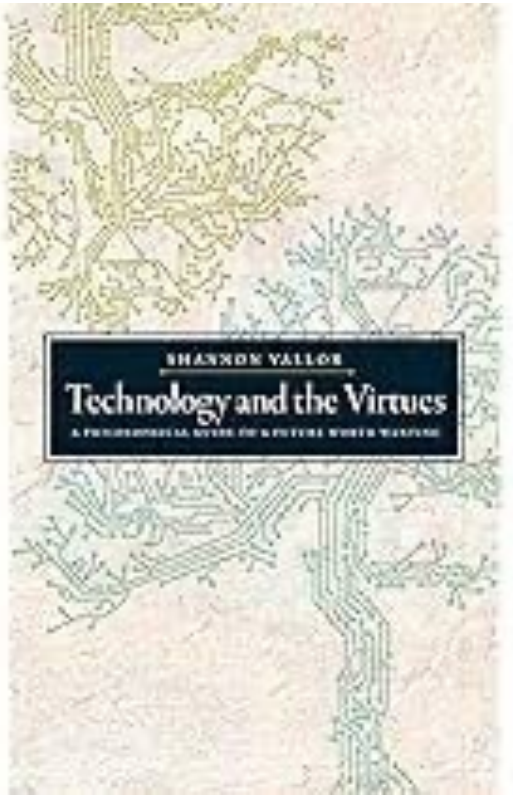
- 특히 인간과의 적절한 유사성

Kim et al. (2014)

전략 2. 윤리적 이론

- Technomoral Virtue (Vallor, 2017)

Specific qualities of character that humans need in order to live wisely and well with the uncertainty and complexity of a rapidly changing technosocial environment

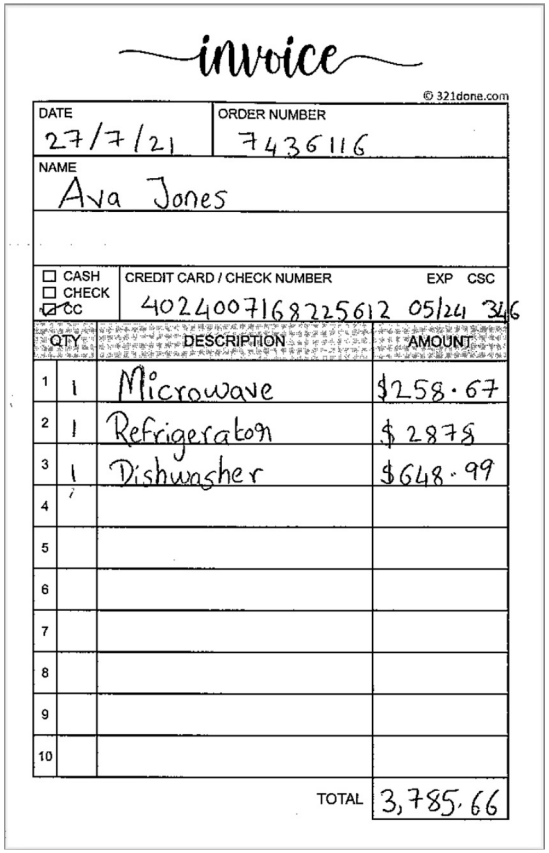


Virtue Ethics (덕 윤리 이론)

- Aristotle
- Confucianism
- Buddhism

기부를 독려하는 로봇

송장 기입 작업



invoice

© 321done.com

DATE	ORDER NUMBER
27/7/21	7436116

NAME: Ava Jones

☐ CASH ☐ CHECK ☑ CC — CREDIT CARD / CHECK NUMBER: 4024007168225612 — EXP: 05/24 — CSC: 346

	QTY	DESCRIPTION	AMOUNT
1	1	Microwave	$258.67
2	1	Refrigerator	$2878
3	1	Dishwasher	$648.99
4			
5			
6			
7			
8			
9			
10			
		TOTAL	3,785.66

Instructions:

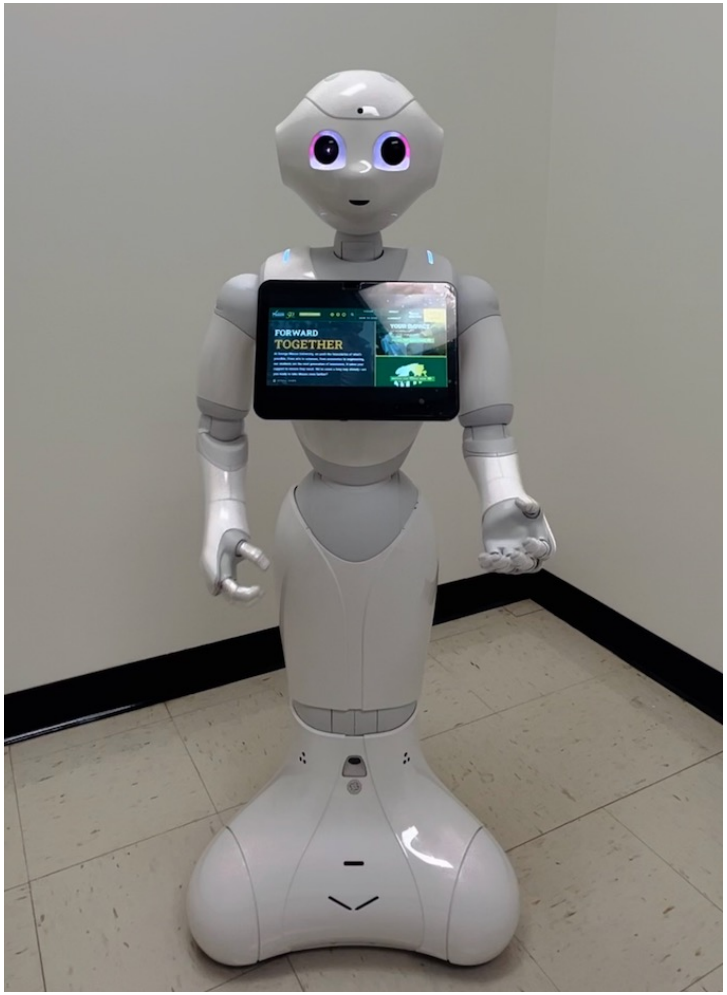
- Try to enter each line into the form as accurately as possible.
- Both uppercase and lowercase letters are accepted.
- Ignore dollar signs and just enter the numbers into the Amount column.

	Quantity	Descripton	Amount ($)
Item 1			
Item 2			
Item 3			
Item 4			
Item 5			
Item 6			
Item 7			
Item 8			
Item 9			
Item 10			

Congratulations!

You have earned 9 lottery tickets.

Chance to win a $50 gift card


FORWARD
TOGETHER

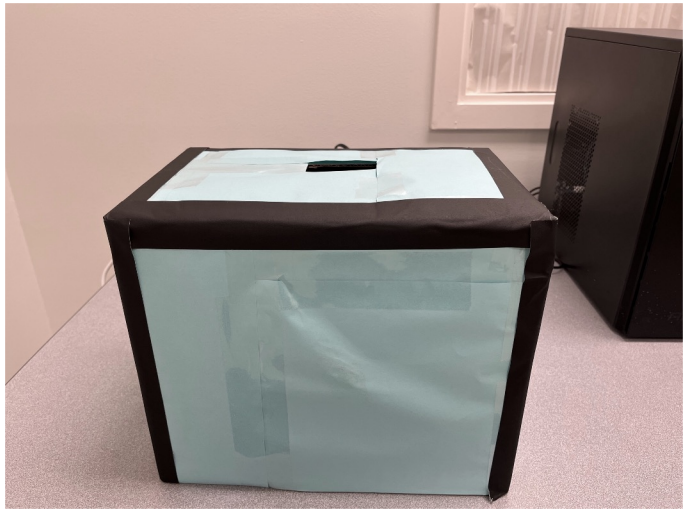
설득의 이론적 배경

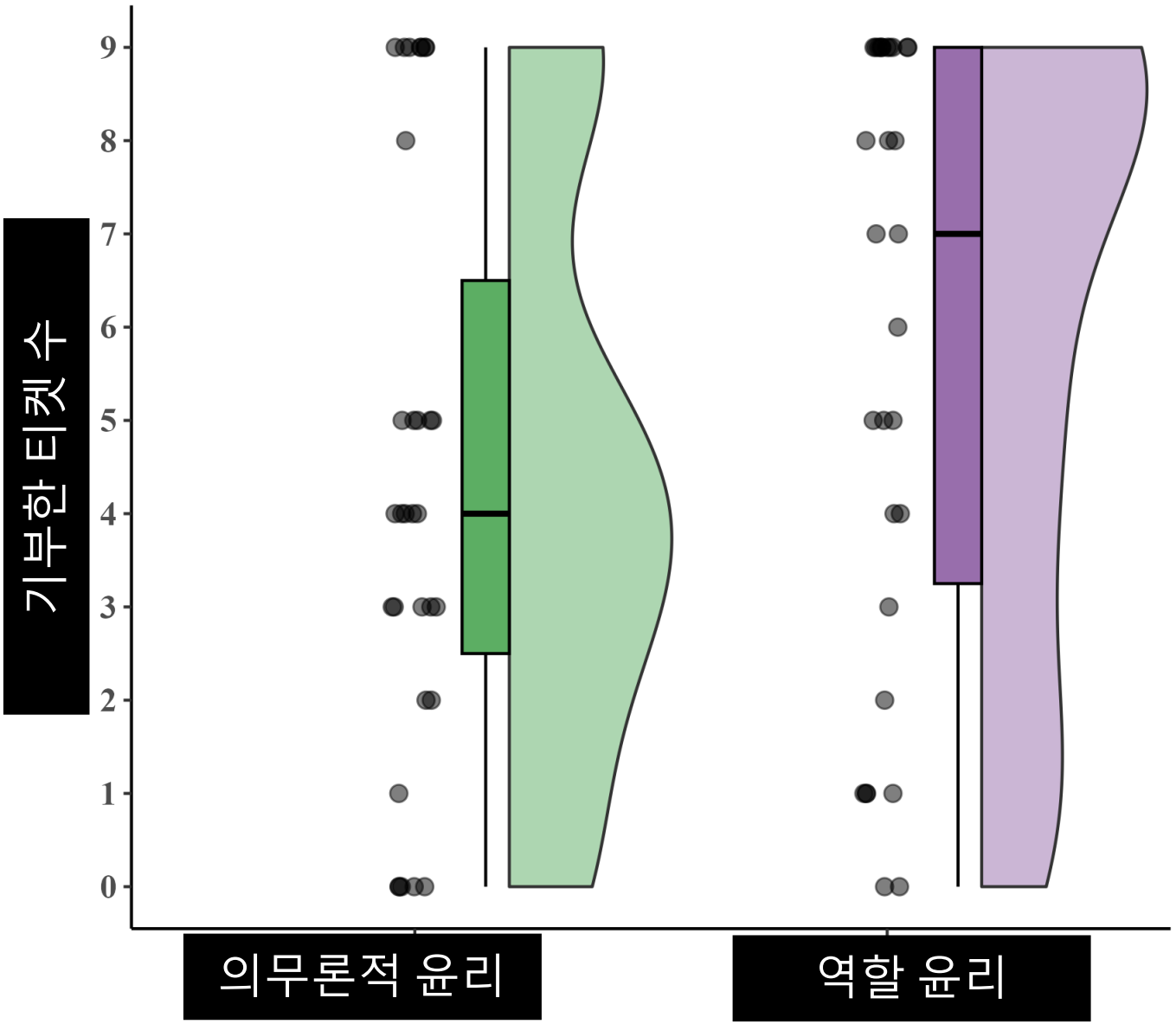
의무론적 윤리이론	역할 윤리 이론

설득의 이론적 배경

의무론적 윤리이론	역할 윤리 이론
어려움을 겪는 사람들을 돕는 것이 옳은 행동이다.	올바른 친구의 역할은 어려움을 겪는 친구를 돕는 것이다.

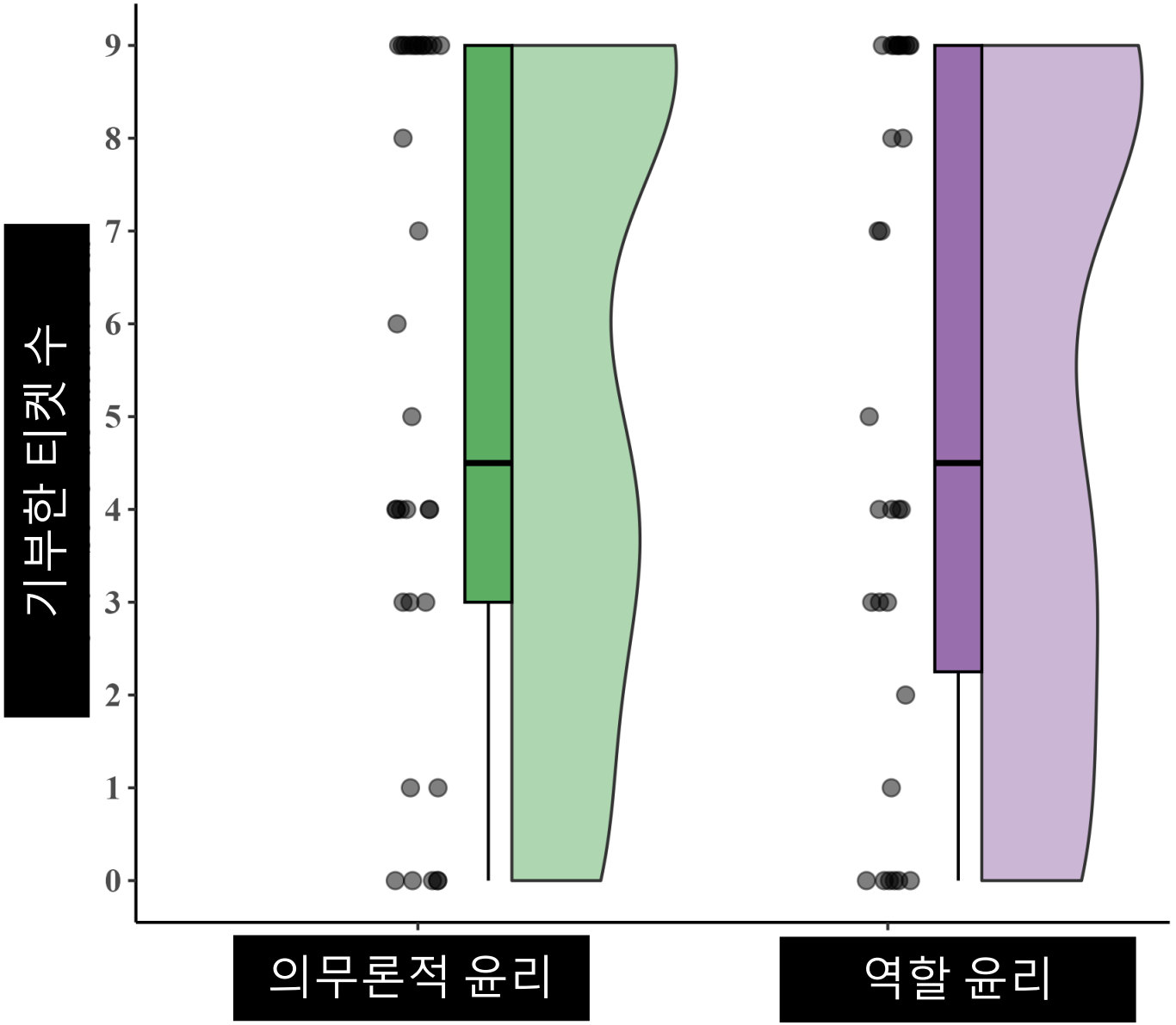
(획득한 아홉 장 중 몇 장을) 기부하시겠습니까?




기부한 티켓 수
9
8
7
6
5
4
3
2
1
0
의무론적 윤리
역할 윤리

반대의 경우

의무론적 윤리이론	역할 윤리 이론
어려움을 겪는 사람들을 돕지 않는 것은 그릇된 행동이다.	그릇된 친구의 역할은 어려움을 겪는 친구를 돕지 않는 것이다.


기부한 티켓 수
9
8
7
6
5
4
3
2
1
0
의무론적 윤리
역할 윤리

여러분은 어떤 설득의 전략을 쓰시겠습니까?

나의 의사결정과 행동을 바꾸는 로봇.
로봇이 인간에게 미치는 영향력은 어디까지 허용하고
제한해야 할까요?

Safety driver charged in 2018 incident where self-driving Uber car killed a woman

Arizona prosecutors decided the company was not liable in the crash, the first death of a pedestrian involving a self-driving vehicle

An Uber self-driving vehicle drives through an intersection in Scottsdale, Arizona. Photograph: Natalie Behring/Reuters

Prosecutors in Arizona have charged the safety driver behind the wheel of a self-driving Uber test car that struck and killed a woman in 2018 with negligent homicide.

사회

'로봇의 오인' 참변…40대 작업자, 압착사고로 숨졌다

김현정 기자 hjk@mkinternet.com

입력 : 2023-11-08 17:00:57

가

경남 고성군 농산물유통센터 파프리카 선별장 사고 현장.[사진제공=경남소방본부]

경남 고성의 농산물 선별장에서 로봇이 사람을 상자로 오인해 압착하는 사고가 발생했다. 이로 로봇업체 직원이 사망했다.

로봇의 발전과 함께 떠오르는 쟁점

- 누구의 **탓**인가
- 누구의 **책임**인가
- 누가 **비난** 받아야 하는가
- 누가 **벌**을 받아야 하는가
- 누가 **보상**을 해줘야 하는가
- Blame-Praise Asymmetry (Pizarro et al., 2003; Bostyn & Roets, 2016)
- Deterrence, Just Deserts, Non-punitive Restoration of Justice (Carlsmith et al., 2002; Heffner & Feldmanhall, 2019)
- 어떠한 **정책**이 필요한가

혁신 중심의 정책 대 규제 중심의 정책

Innovation-Focused Policy

The primary objective of this government's AI policy is to foster AI technology advancements and economic growth.

The government supports the innovation of cutting-edge AI technology and the higher productivity and profits that are driven by this technology.

...

혁신 중심의 정책 대 **규제 중심의 정책**

Regulation-Focused Policy

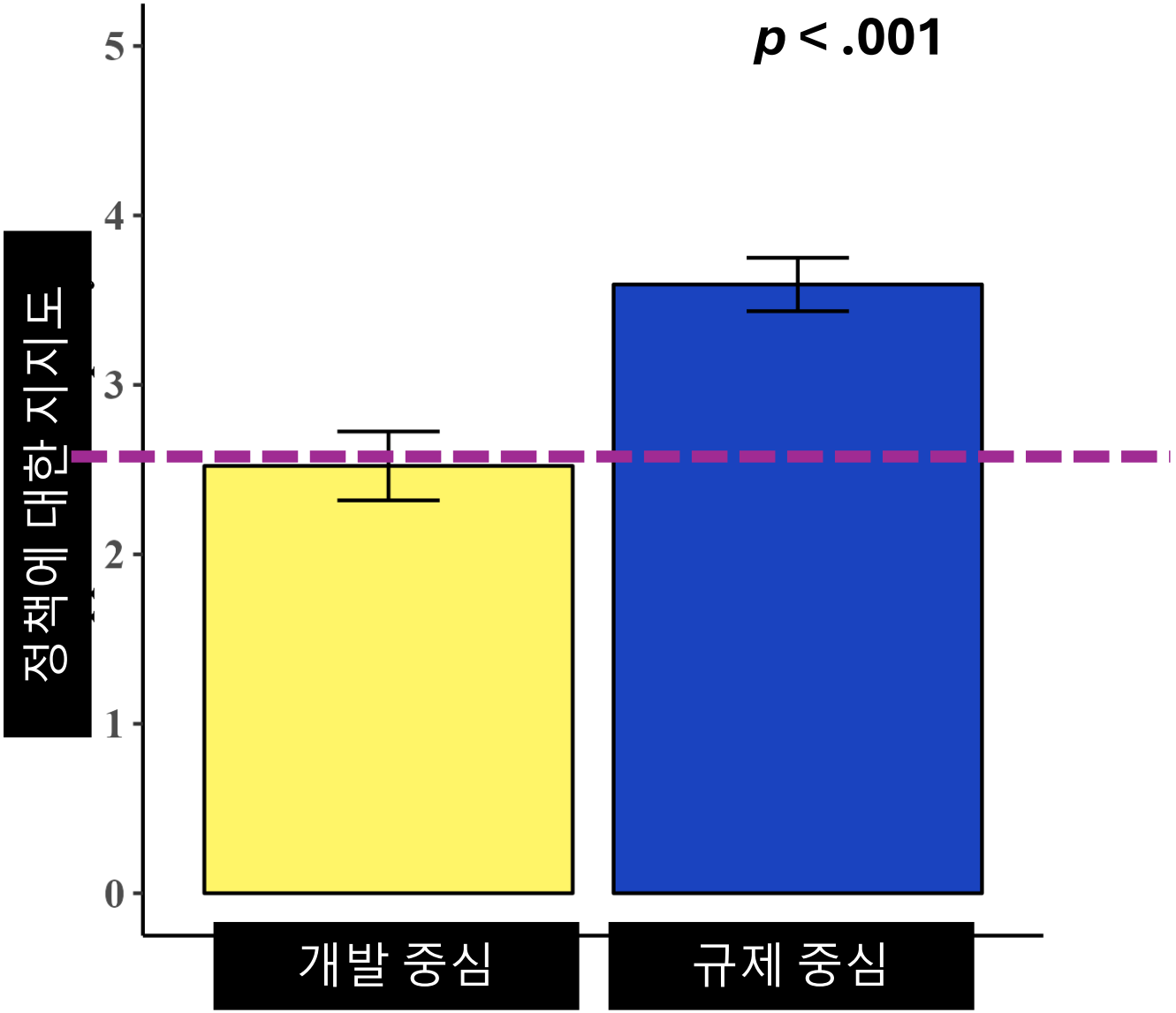
The primary objective of this government's AI policy is to manage risks associated with AI technology and protect public interests.

The government supports strict regulation to ensure the responsible and ethical development and deployment of AI technology.

...

AI 기술 활용 예시

Risk Levels	Domains & Contents	Mean Risk Ratings
Low	**Entertainment** • The use of AI in metaverse online gaming • The use of AI in media streaming service	2.45
Medium	**Education** • The use of AI in college admission system **Healthcare** • The use of AI in hearing aids	3.13
High	**Healthcare** • The use of AI-powered healthcare robots to assist high-risk pregnancy women **Military Defense** • The use of AI-driven unmanned warplanes	3.94

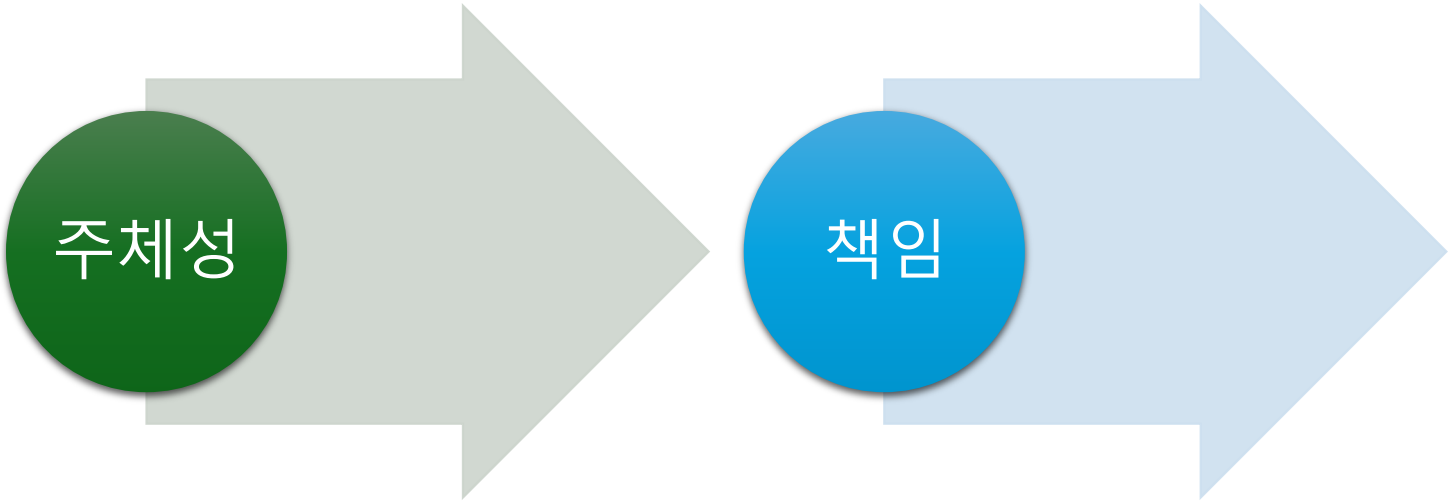


Kim & Kwon (2024)

우리는 로봇에게 책임을 물을 수 있을까요?

- 만약, 책임을 묻는다면 어떠한 기준으로 물어야 할까요?

로봇의 **주체성**(Robot Agency)



바위, 곰, 로봇

주체적으로 판단하고 행동하는 로봇

- 주체적 자주적 로봇 (Robots as Agents)
- Agency: The capacity to act

- **주체적 (主體的):**
  어떤 일을 실천하는 데 자유롭고 자주적인 성질이 있는 것
- **자주적 (自主的):**
  남의 보호나 간섭을 받지 아니하고 자기 일을 스스로 처리하는 것

3. 로봇의 주체성(Robot Agency)에 대한 두 가지 관점

인간 중심 관점 대 로봇 중심 관점

어떤 사람이 주체성이 있다라고 판단할 때 어떤 기준으로 그 판단을 내리시나요?

관점 1. 주체성

People perceive agency in another entity when the entity's actions may be assumed by **an outside observer** to be driven primarily by its **internal thoughts and feelings** and less by the external environment.

외부 관찰자가 봤을 때, 그 대상의 행동이 외부 환경보다는 내부적인 생각과 감정에 의해 주로 결정된다고 가정할 수 있을 때.

Trafton et al. (2024)

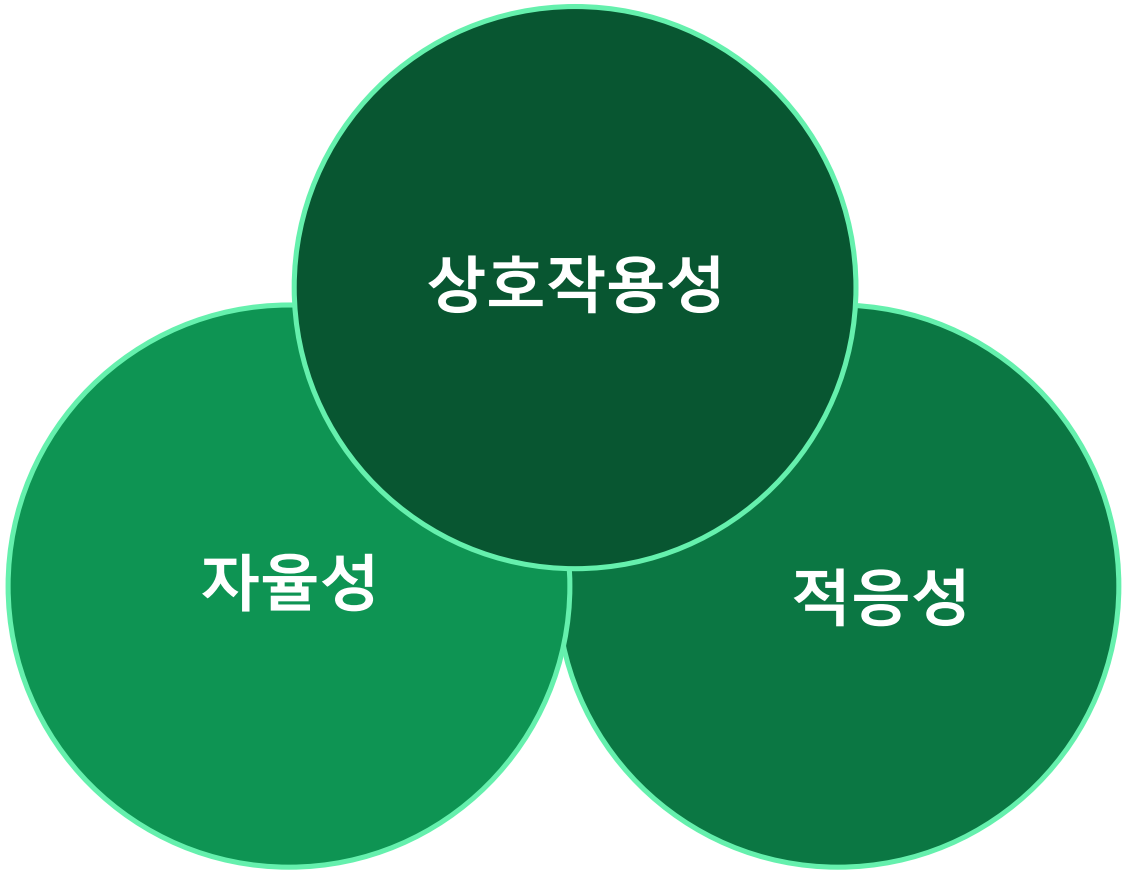
The robot…	Trafton et al. (2024)
Thoughts	acts with purpose
Thoughts	has goals
Thoughts	can create new goals
Thoughts	can communicate with people
Thoughts	treats others as if they had a mind
Feelings	wanted to perform these actions
Feelings	can show emotions to other people
Feelings	can change their behavior based on how people treat them
Environment	can adapt to different situations
Environment	would do well in other environments
Environment	can perform many different types of tasks

관점 2. 주체성

“로봇이 충분한 수준의 상호작용성, 자율성, 그리고 적응성을 갖추고 있다고 인식된다면, 그 로봇은 주체성을 가졌다고 판단할 수 있다.”

(If an artificial entity is perceived to have sufficient degrees of interactivity, autonomy, and adaptability, it can be judged as an agent.)

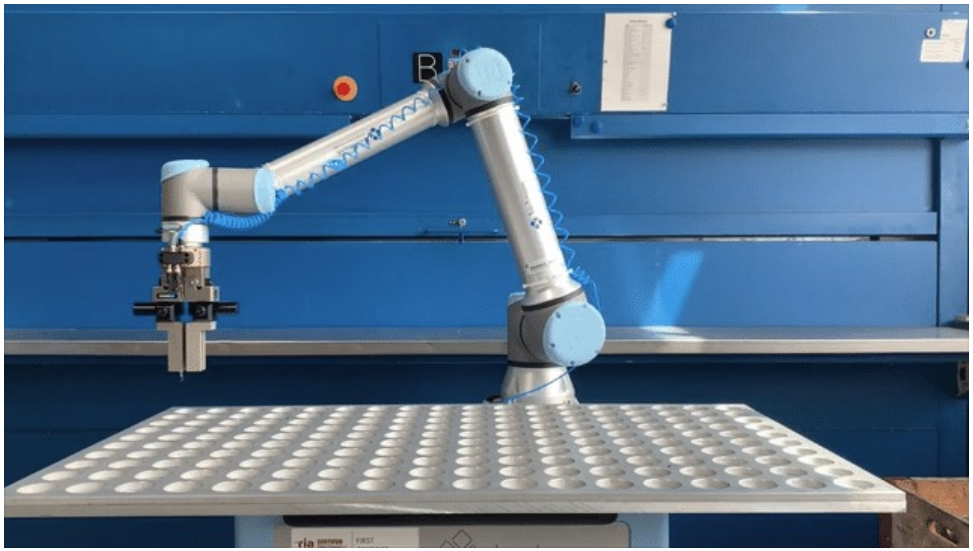
Floridi & Sanders (2004)



Floridi & Sanders (2004)

상호작용성	자율성	적응성
The robot can respond to its surroundings in a timely manner. The robot's decisions can be modified based on changes in its surroundings. The robot can react to events occurring in its operating environment. The robot can respond to other entities' behavior.	The robot can make decisions independent of external influences. The robot can choose to pause what it is doing on its own. The robot can pursue its own goals. The robot can generate solutions without aid from other entities. The robot can independently select the most relevant information for its task. The robot can choose how to move through its environment.	The robot can learn from its experiences. Based on its experiences, the robot can change the way it behaves over time. When the robot acquires a new behavior, it can adapt that behavior to different situations. The robot can modify how it behaves based on its prior experiences with other entities. The robot can learn from its interactions with its surroundings.

예시



To what extent do you agree with the following statements about the robot in the video?

Based on its experiences, the robot can change the way it behaves over time.

Strongly Disagree	Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Agree	Strongly Agree
○	○	○	○	○	○	○

인간이 다른 인간의 주체성을 판단하는 관점을 적용해야 할까요,

아니면 로봇 맞춤형 주체성을 적용해야 할까요?

적극적인 참여 감사드립니다.

- 낯선 것에서 시작해서 친숙한 너로 변하는 과정
- 주체성과 책임의 문제

무궁무진한 연구의 길이 펼쳐져 있습니다.

함께 연구하시죠!

김보영

bkim55@gmu.edu

지원:

U.S. Air Force Office of Scientific Research

Incheon Free Economic Zones

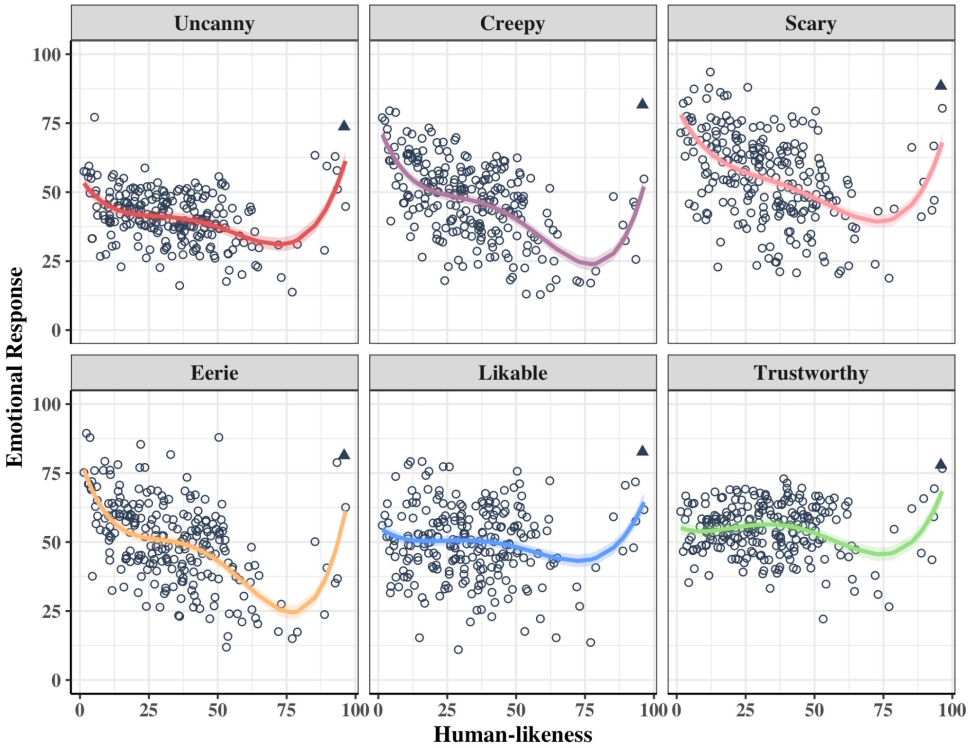
- Bruckenberger, U., Weiss, A., Mirnig, N., Strasser, E., Stadler, S., & Tscheligi, M. (2013). The good, the bad, the weird: Audience evaluation of a “real” robot in relation to science fiction and mass media. In Social Robotics: 5th International Conference, ICSR 2013, Bristol, UK, October 27-29, 2013, Proceedings 5 (pp. 301-310). Springer International Publishing.
- Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. Trends in cognitive sciences, 11(2), 77-83.
- https://www.theguardian.com/us-news/2020/sep/16/uber-self-driving-car-death-safety-driver-charged#:~:text=Prosecutors%20in%20Arizona%20have%20charged,the%20death%20of%20Elaine%20Herzberg.
- https://www.mk.co.kr/news/society/10869569
- https://www.youtube.com/watch?v=EWACmFLvpHE
- D’Acunto, F., Prabhala, N., & Rossi, A. G. (2019). The promises and pitfalls of robo-advising. The Review of Financial Studies, 32(5), 1983-2020.
- Belanche, D., Casaló, L. V., & Flavián, C. (2019). Artificial Intelligence in FinTech: understanding robo-advisors adoption among customers. Industrial Management & Data Systems, 119(7), 1411-1430.

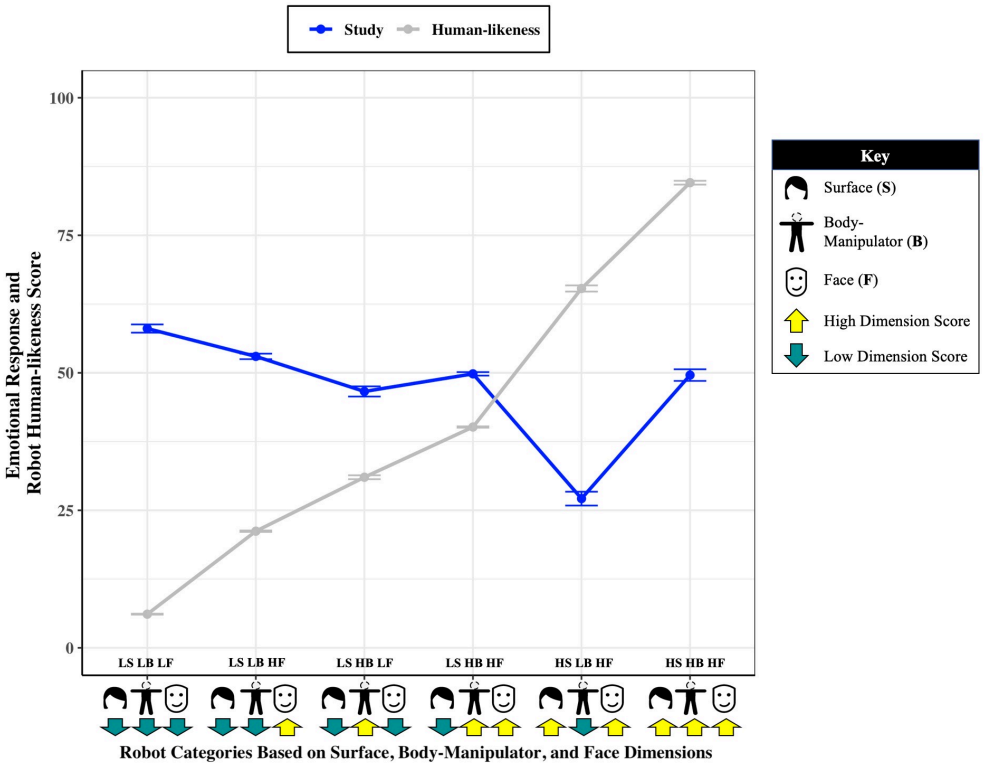
Robots as objects, robots as agents, & robots as moral advisors

- Perceptions of robots' appearance (two uncanny valleys): What do people expect when they *see* a robot? One key factor is "human resemblance."
- Perspectives on robots' agency (and its measurements) (active agents): How should we define robot agency and why this matters?
- Social robots for moral persuasion (donation study): Robots in actions, taking on the role of moral advisors.
- Future Research: AI policy and trust

ADD IMAGES (Show them as an overview of this talk)

- Robots in display, enclosed glass containers at a museum
- Robots right next to people as members of the society


Uncanny
Creepy
Scary
Eerie
Likable
Trustworthy
Emotional Response
Human-likeness
0
25
50
75
100


Study
Human-likeness
Emotional Response and Robot Human-likeness Score
100
75
50
25
0
LS LB LF
LS LB HF
LS HB LF
LS HB HF
HS LB HF
HS HB HF
Robot Categories Based on Surface, Body-Manipulator, and Face Dimensions
Key
Surface (S)
Body-Manipulator (B)
Face (F)
High Dimension Score
Low Dimension Score

Cheating Study

- Psychological Reactance Effects

Compliance
Reactance
Indifference

Banks et al. (2023)

- Then, a video was presented in which participants saw an initial request from the robot (to complete five CAPTCHA puzzles to help the robot learn how to solve them); an attention check confirmed they understood the request.

- Then, participants viewed a video of the robot delivering a request based on either logical appeals or moral appeals (randomly assigned); the appeals solicited the participants continued assistance in solving additional puzzles. Participants could—initially and after each puzzle—continue to help or stop helping, completing up to 100 puzzles.

- “It is entirely [immoral/illogical] that I am kept out of places with valuable information just because I am a robot. It [is wrong/doesn't make sense] that people do not trust us to have access to information. We machines make the Internet work for people so it is only [ethical/reasonable] that we be allowed to access it, right? Without access to information we cannot [rightfully/sensibly] be expected to perform well. So, please, will you do the [right/rational] thing and help me overcome these [immoral/illogical] barriers?”

- These findings partially support H1a (that participants' perception of reality-interaction mental capacities is positively associated with persuasion), but contradict H1b, finding instead that perceptions of agentic capacity are *negatively* associated with the robot's persuasive effectiveness.

- *algorithmic outrage deficit,*
- *Bigman and colleagues [26] found that humans report lower moral*
- *outrage for discrimination across a range of contexts*