The Punishment Gap:
The Unmet Desire to Punish AI and Robots Upon Damages

Gabriel Lima, Meeyoung Cha, Chihyung Jeon, Kyung Sin Park

KAIST ibs Institute for Basic Science KOREA UNIVERSITY

Gabriel Lima

Research Intern, Data Science Group, Institute for Basic Science
Senior Undergraduate, School of Computing, KAIST

gabriel.lima@kaist.ac.kr

Who Is Responsible for Machine Actions?

- In case of damages, who should be held responsible for the actions of AI and robots?
- How does the general public assign responsibility for the actions of autonomous machines?
- What are the possible moral, ethical, and legal conflicts in the public perception towards machine responsibility?

Lima, Gabriel, et al. "Will Punishing Robots Become Imperative in the Future?." Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. 2020.
Lima, Gabriel, et al. "Explaining the Punishment Gap of AI and Robots." arXiv preprint arXiv:2003.06507 (2020). Under Review.

Machine Moral Agency & Patiency

- How do people perceive the agency of AI and robots in comparison to human agency?
- Do people perceive AI and robots as moral patients who deserve consideration?
- How to deal with possible conflicts with human rights?

Lima, Gabriel, et al. "Collecting the Public Perception of AI and Robot Rights." arXiv preprint arXiv:2008.01339 (2020). To Appear on ACM CSCW 2020.

KAIST ibs 기초과학연구원 Institute for Basic Science

A Google self-driving car caused a crash for the first time

A bad assumption led to a minor fender-bender

By Chris Ziegler | Feb 29, 2016, 1:50pm EST

Source California DMV | Via Mark Harris (Twitter)

SHARE

“We clearly bear some responsibility”

Google took responsibility for it

THE VERGE

The Guardian

Tesla driver dies in first fatal crash while using autopilot mode

The autopilot sensors on the Model S failed to distinguish a white tractor-trailer crossing the highway against a bright sky

"The customer [...] always in control and responsible"

Tesla did not take responsibility for it

What about a self-driving car that has adapted its driving to your needs after years of driving?

Was the damage foreseeable?
Was the owner negligent about the risks?
Was there a breach of the owner/manufacturer's duty of care?

Was the damage foreseeable?
Was the owner negligent about the risks?
Was there a breach of the owner/manufacturer's duty of care?

Autonomy risk → some errors might be inevitable

Responsibility Gaps

- Difficulties in attributing responsibility for the actions of autonomous and self-learning machines
  - Responsibility Gap *(Matthias, 2004; Sparrow, 2007; Beck, 2015; Gunkel, 2017; Coeckelbergh, 2019)*
  - Accountability Gap *(Koops et al., 2010)*
  - Double Dilemma *(Schirmer, 2020)*
  - Retribution Gap *(Danaher, 2016)*
  - Liability Gap *(Asaro, 2015)*

How a Self-Driving Uber Killed a Pedestrian in Arizona

By TROY GRIGGS and DAISUKE WAKABAYASHI UPDATED MARCH 21, 2018

A woman was struck and killed on Sunday night by an autonomous car operated by Uber in Tempe, Ariz. It was believed to be the first pedestrian death associated with self-driving technology.

The New York Times

Killer Robots Aren't Regulated. Yet.

"Killing in the Age of Algorithms" is a New York Times documentary examining the future of artificial intelligence and warfare.

The New York Times

PRO PUBLICA

How We Analyzed the COMPAS Recidivism Algorithm

by Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin

May 23, 2016

Is it possible to hold these systems themselves responsible for damages?

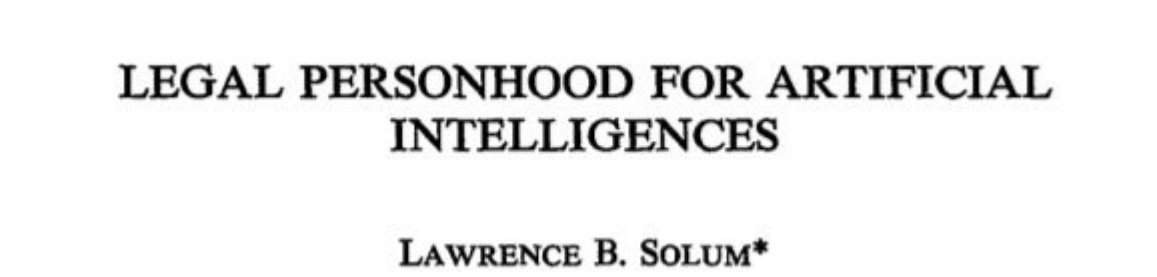
European Parliament Recommendation

- *"creating a specific legal status for robots [...] having the status of electronic persons responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make autonomous decisions [...]"*

European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics
http://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.html

Electronic Legal Personhood For Autonomous AI?

- Willick, Marshal S. "Constitutional Law and Artificial Intelligence: The Potential Legal Recognition of Computers as" Persons"." IJCAI. 1985
  - *“it may prove impossible in the future to draw a valid legal distinction between humans and computers”*
- Solum, Lawrence B. "Legal personhood for artificial intelligences." NCL Rev. 70 (1991): 1231.
  - Proposes electronic legal personhood for AIs.



LEGAL PERSONHOOD FOR ARTIFICIAL INTELLIGENCES

LAWRENCE B. SOLUM*

Multifaceted Discussion - Supporters

- Liability questions can be more easily solved
- Lead to innovation and economic growth
- Coherence of legal system
- Non-natural entities
- Experiences with previously neglected entities

- Possible human liability shields
- Instrumentalist theories
- Confrontation with human rights
- Current doctrines can deal with all situations
- How to punish them?

Koops, Bert-Jaap, Mireille Hildebrandt, and David-Olivier Jaquet-Chiffelle. "Bridging the accountability gap: Rights for new entities in the information society." Minn. JL Sci. & Tech. 11 (2010): 497.
Turner, Jacob. Robot Rules: Regulating Artificial Intelligence. Springer, 2018.
Chopra, Samir, and Laurence F. White. A legal theory for autonomous artificial agents. University of Michigan Press, 2011.
van Genderen, Robert van den Hoven. "Do We Need New Legal Personhood in the Age of Robots and AI?." Robotics, AI and the Future of Law. Springer, Singapore, 2018. 15-55.

Multifaceted Discussion - Adversaries

- Liability questions can be more easily solved
- Lead to innovation and economic growth
- Coherence of legal system
- Non-natural entities
- Experiences with previously neglected entities

- Possible human liability shields
- Instrumentalist theories
- Confrontation with human rights
- Current doctrines can deal with all situations
- How to punish them?

Bryson, Joanna J. "Robots should be slaves." Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues (2010): 63-74.
Bryson, Joanna J., Mihailis E. Diamantis, and Thomas D. Grant. "Of, for, and by the people: the legal lacuna of synthetic persons." Artificial Intelligence and Law 25.3 (2017): 273-291.
Solaiman, S. M. "Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy." Artificial Intelligence and Law 25.2 (2017): 155-179.
Asaro, Peter M. "11 A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics." Robot ethics: The ethical and social implications of robotics (2011): 169.

Punishment of AI and Robots

- Reform, deterrence, and retribution
- Humans are retributivists → people should be punished for harm because they deserve it
- AI and robots do not have any assets or physical independence

How can we punish them?

Survey Objective

- What do people think about electronic legal personhood and the liability of AI and robots?

1. How do people attribute **responsibility**, **awareness**, and **punishment** to AI, robots, humans and associates depending on **autonomy** and **agent?**
2. Does punishment of electronic agents fulfill punishment preconditions and requirements?

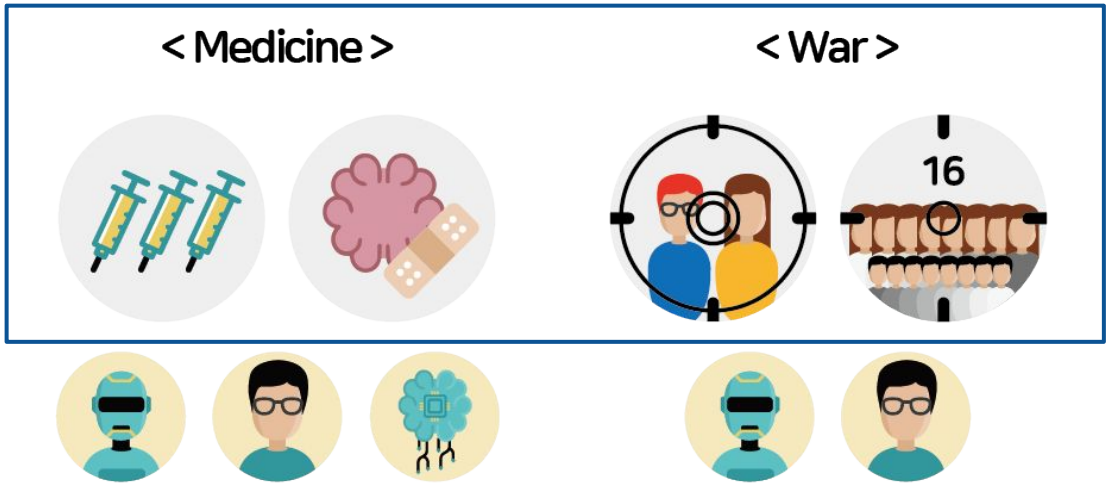
- Amazon Mechanical Turk → *N*=3315 valid responses

Regulation of AI and Robots

- *Algorithmic social contract* requires consent from stakeholders
- Social contract dependent on legitimacy and public confidence
  - Public participation is necessary for a successful and legitimate legal system
- *"[...] how to deal with it should be discussed upfront and publicly, with stakeholders and in society at large, [...]"*

Rahwan, Iyad. "Society-in-the-loop: programming the algorithmic social contract." 2018.
Coeckelbergh, Mark. "Artificial Intelligence , Responsibility Attribution , and a Relational Justification of Explainability." 2019.

Scenarios & Agents



Real Human Cases → AI and Robots


Fitchburg Sentinel

'Our Son Is Not.... ....That Type Of Boy'

Fire, Explosions Raze 20 Homes

Planning Director Urged For City

News Digest

City, Firms Team Up On River

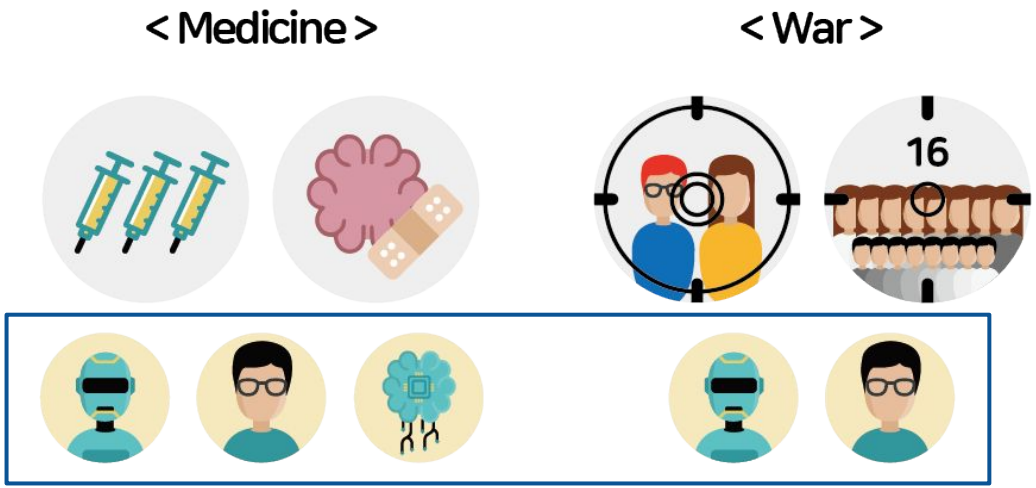
St. Francis' Pastor Shifted To Worcester

Congress Gets More Color

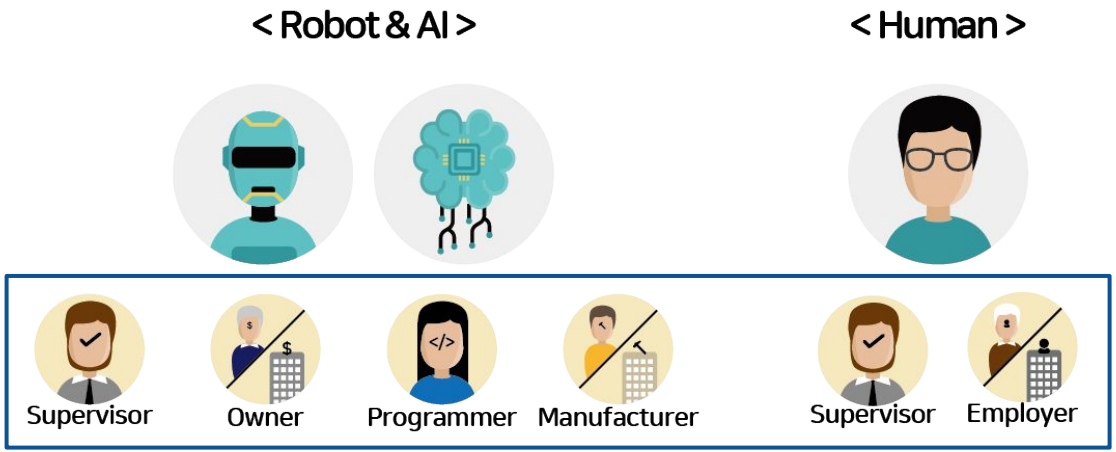
Triskaidecaphobia Day

Murder of 2 vietnamese civilians during the Vietnam War → what if the soldier was a robot?

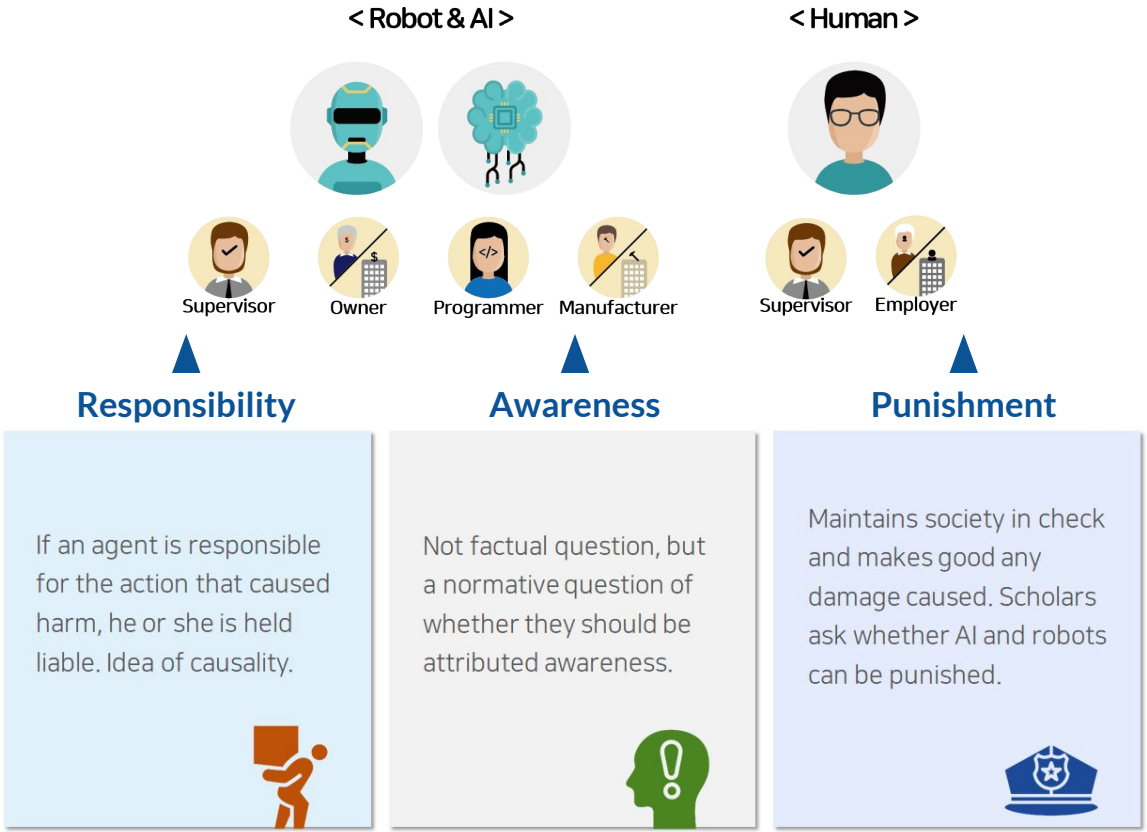
Scenarios & Agents



Associates


< Robot & AI >
< Human >
Supervisor
Owner
Programmer
Manufacturer
Supervisor
Employer

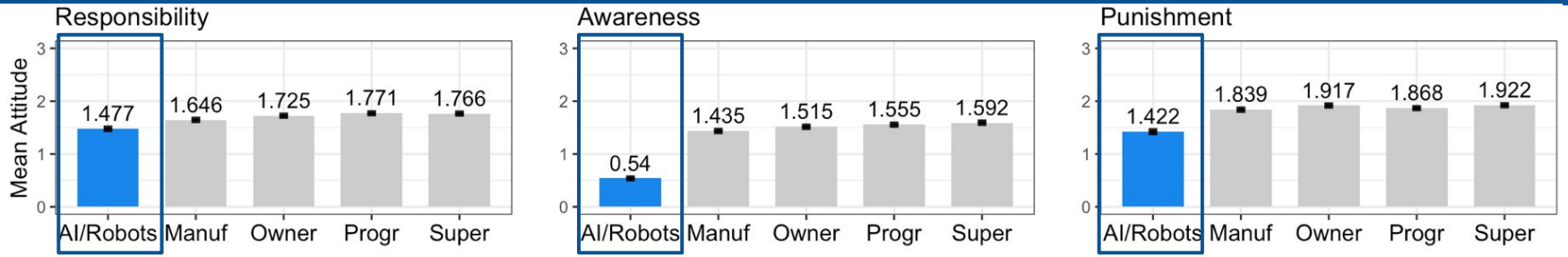
Assignment of Variables to Entities



Question:

How do people attribute **responsibility, awareness, and punishment** to **AI, robots, humans and associates** depending on **autonomy and agent**?

Liability Assignment - AI/Robots & Associates

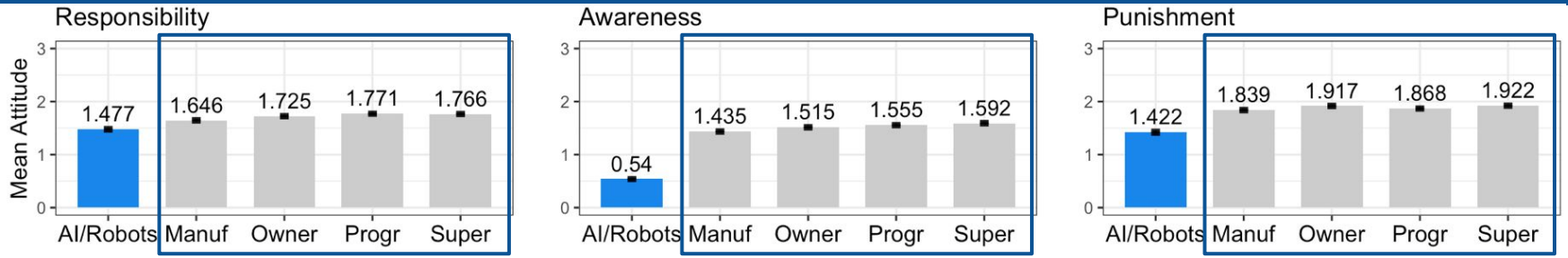


0 - Not at all, 1 - A little, 2 - Some, 3 - Very

- AI and Robots are responsible and should be punished
- AI and Robots are not aware → lack of mental state, important for criminal and civil liability
- Punishment as a two step process: causality → mind perception

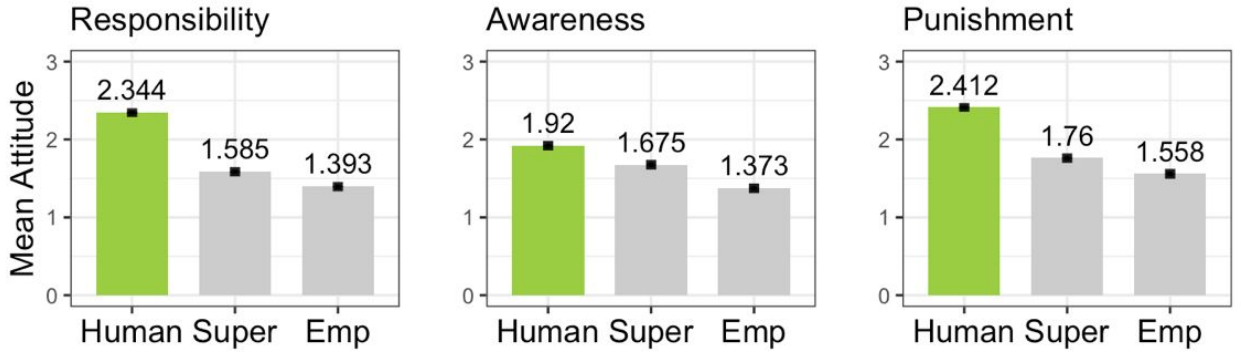
Cushman, Fiery. "Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment." Cognition 108.2 (2008): 353-380.

Liability Assignment - AI/Robots & Associates



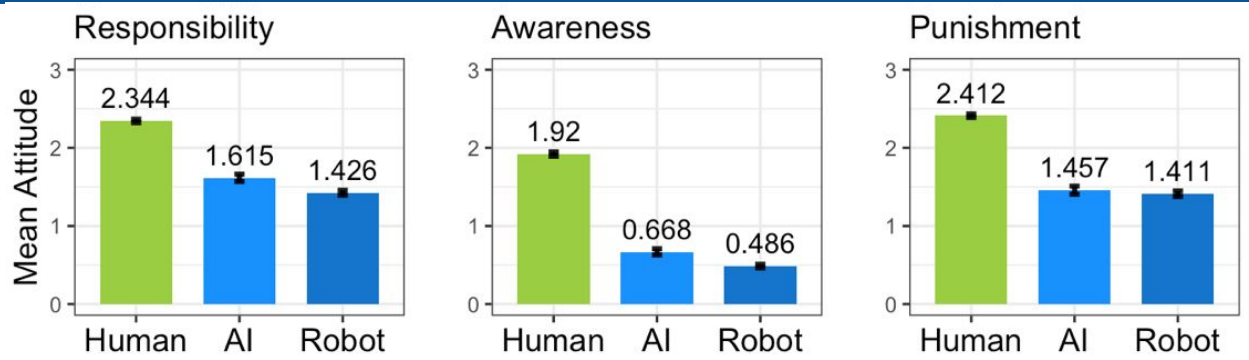
- All associates are similarly responsible, aware and worthy of punishment.
- Manufacturer has lower responsibility and awareness → conflict with current situation

Liability Assignment - Human & Associates



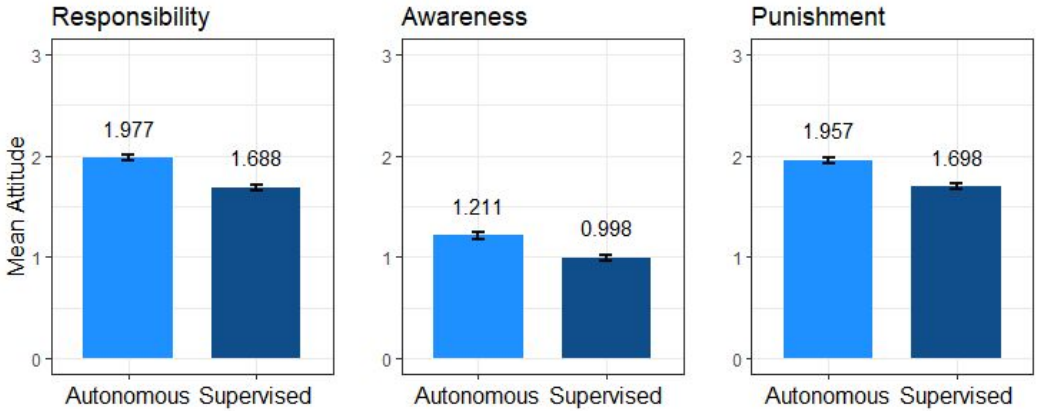
Human agent > Supervisor > Employer

Liability Assignment - Human vs. AI/Robot



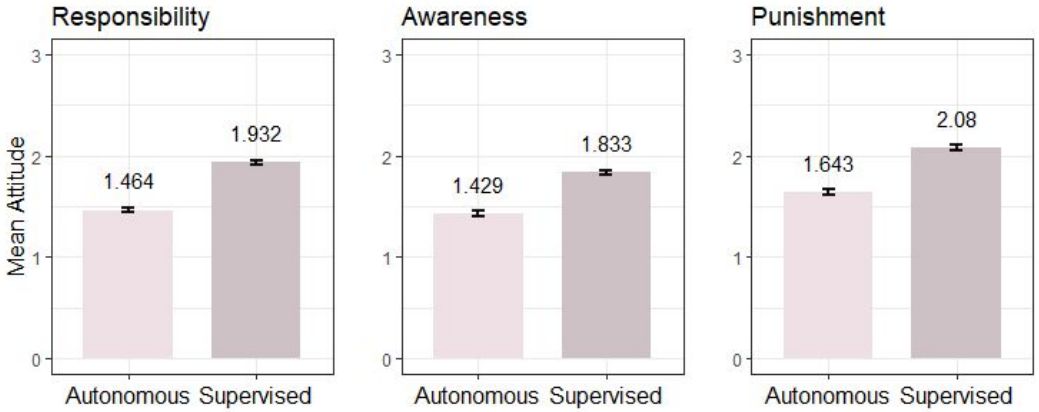
- Less than humans but significant
- Punishment as a two step process: causality → mind perception
- AI is considered more aware and responsible than robots
  - Why?

Liability Assignment - Autonomy Levels



Supervision decreases responsibility, awareness and punishment of the agent

Liability Assignment - Autonomy Levels

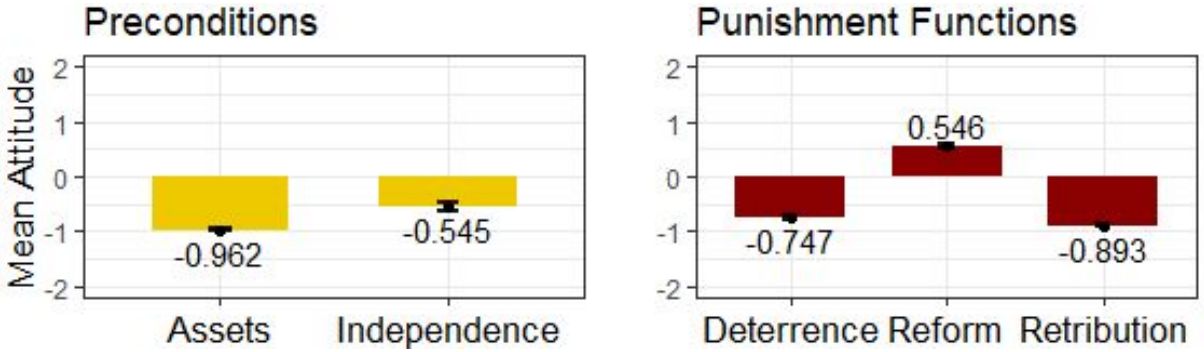


But the opposite occurs for the supervisor

Question:

Does the punishment of electronic agents fulfill punishment preconditions and requirements?

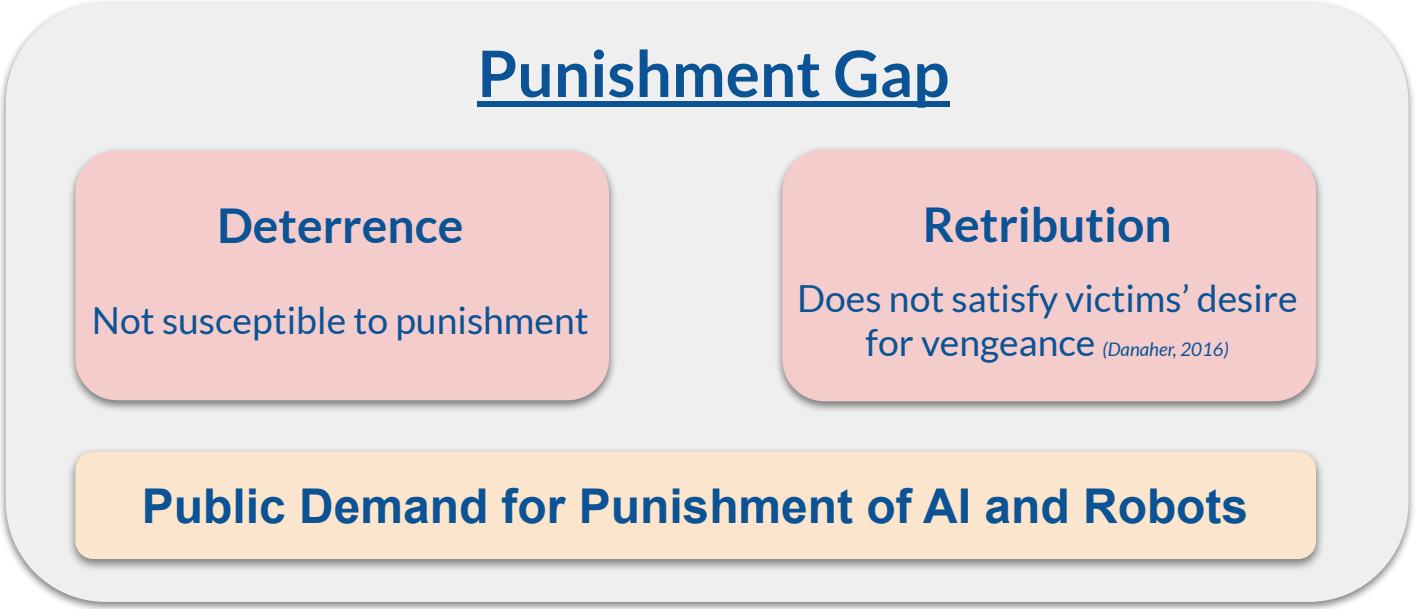
Punishment of AI and Robots



- Preconditions: assets for civil liability and physical independence for criminal liability
- Punishment of AI and robots does not satisfy most of its functions and preconditions

The Punishment Gap

- General public finds AI and robots responsible and worthy of punishment
- However, they do not agree on how to punish them → **punishment gap**



Electronic Legal Personhood Is Not Viable

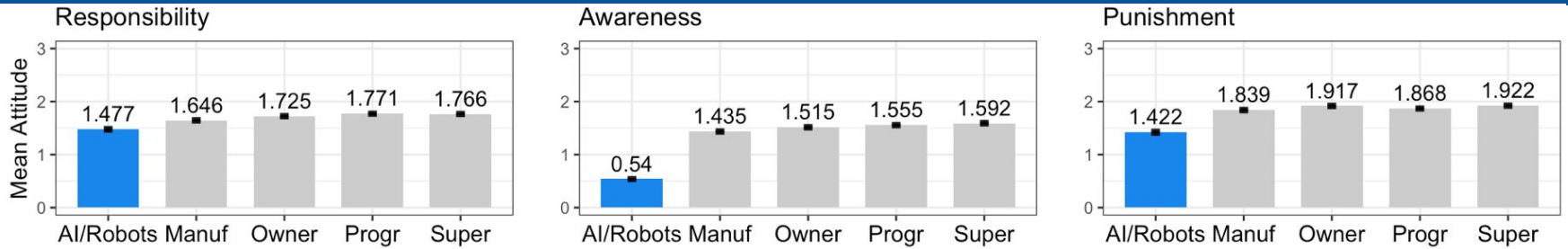
- Civil liability cannot be imposed without assets
- Criminal liability cannot be imposed without physical independence
- **We might need a broad legal reframing to solve this public contradiction → change punishment functions and methods or change people's perception**

Realistic Interests Might Be a Partial Solution

- Shape behavior so it does not go against self-interests → deal with deterrence aspect.
- Legitimate punishment according to society → deal with the retribution aspect.
- **What is a "realistic interest"**? → not clearly defined (e.g., avoidance of pain)

Turner, Jacob. Robot Rules: Regulating Artificial Intelligence. Springer, 2018.

Distributed Liability Across All Associates



Everyone is considered similarly responsible, awareness and punished.

Common Enterprise Liability

Joint Liability/Extended Agency

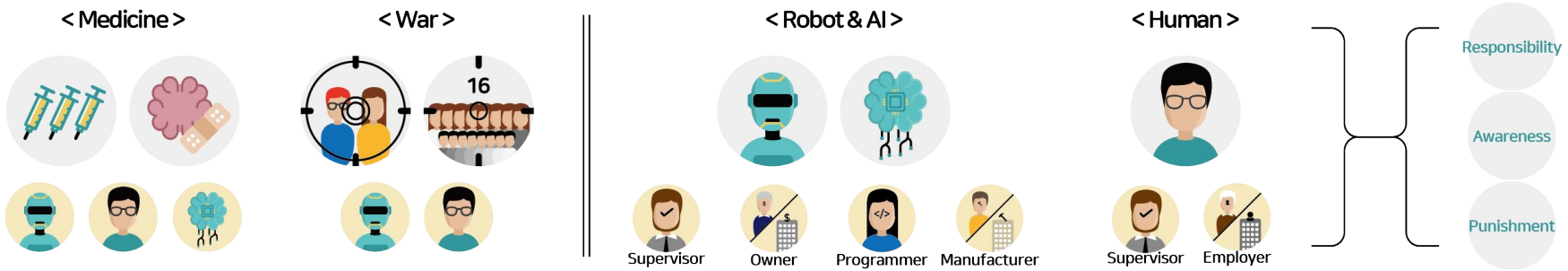
Vladeck, David C. "Machines without principals: liability rules and artificial intelligence." Wash. L. Rev. 89 (2014): 117.

Hanson, F. Allan. "Beyond the skin bag: On the moral responsibility of extended agencies." Ethics and information technology 11.1 (2009): 91-99.

The Punishment Gap:

The Unmet Desire to Punish AI and Robots Upon Damages

Gabriel Lima, gabriel.lima@kaist.ac.kr



AI and robots are considered **responsible** and **deservers of punishment**

→

People **do not agree** with granting them **assets** or **physical independence**

→

Punishment Gap

Lima et al., 2020; Park et al., 2020; Full version under review. Preprint available on ArXiv.