Giudizi morali, etica e robot umanoidi

Workshop CAREUS: ETICA e INTELLIGENZA ARTIFICIALE

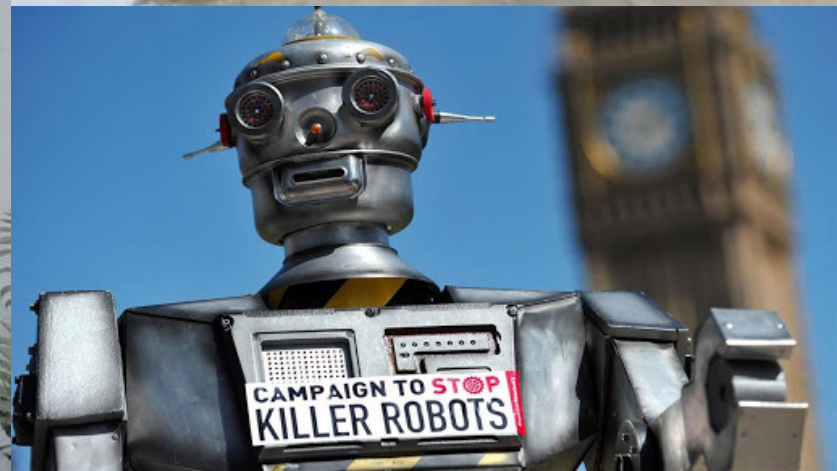
Siena, 27 Marzo 2023

Stefano Guidi[1]

Dipartimento di Scienza Sociali, Politiche e Cognitive – DISPOC

Università degli Studi di Siena

KUKA


CAMPAIGN TO STOP
KILLER ROBOTS

Giudizi di responsabilità di sistemi artificiali

Esseri umani

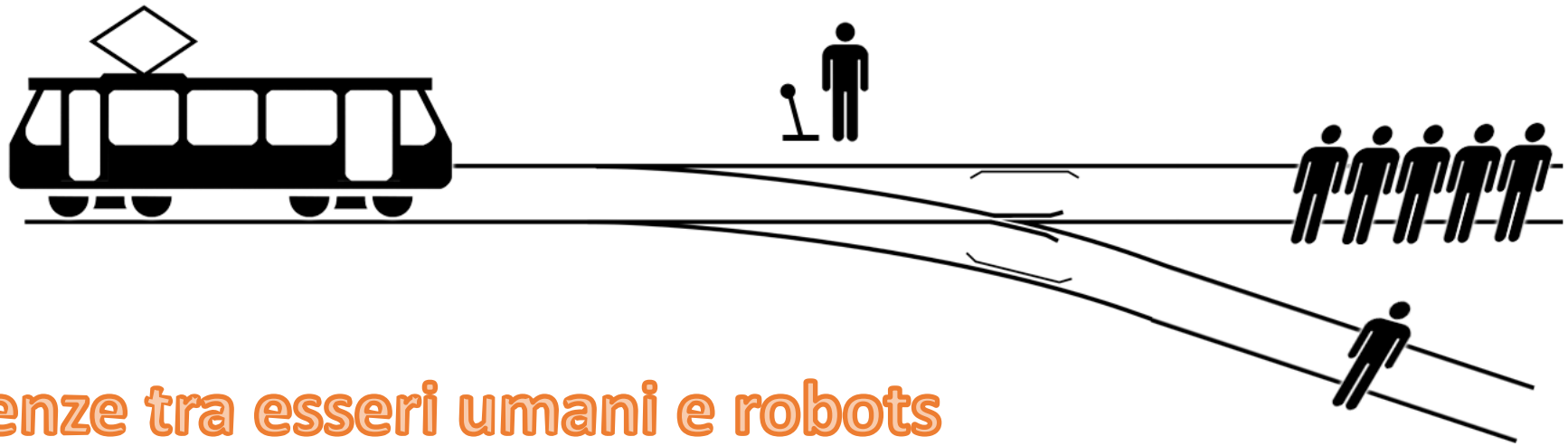
- Intenzionalità
- Capacità di recare danno ad altri

Sistemi artificiali

- Autonomia
- Capacità di mostrare abilità sociali
- Algoritmo

- Designers
- Produttori
- Utenti/Controllori

Dilemmi Morali



Differenze tra esseri umani e robots

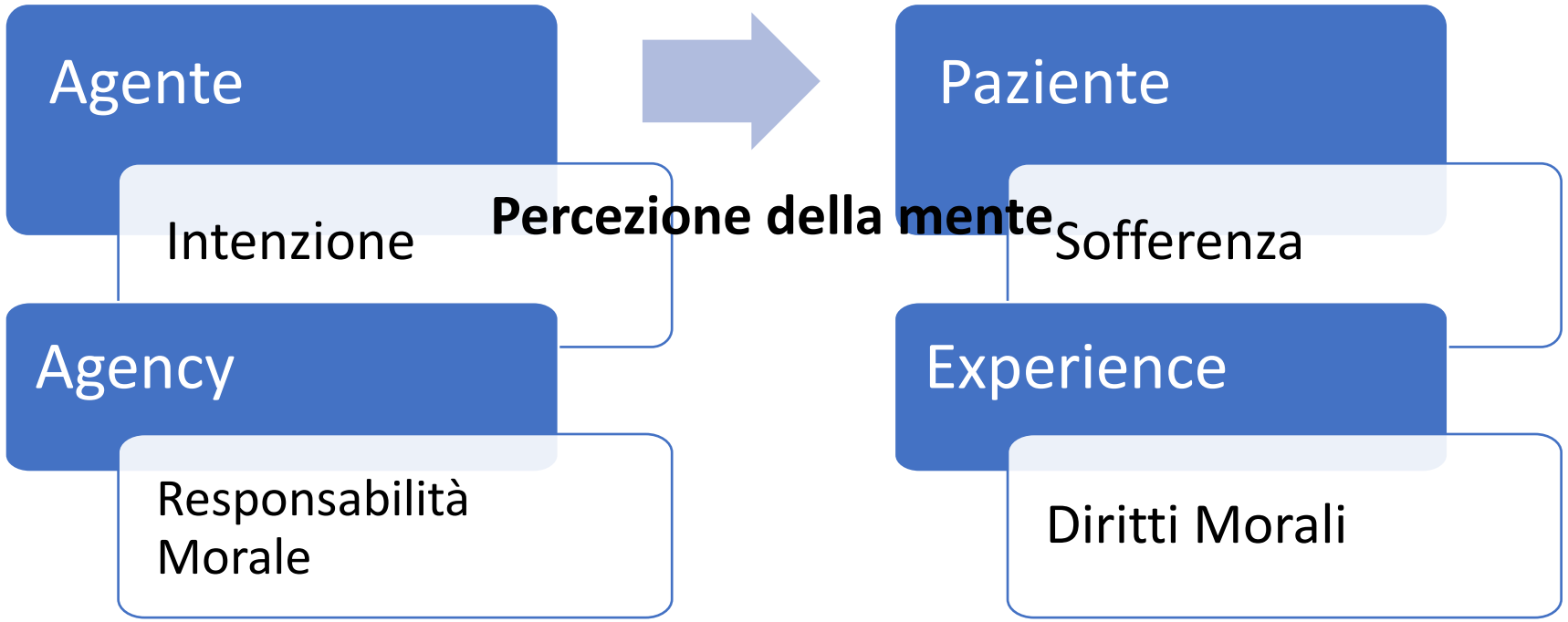
Esseri umani	Robot
Biasimati più per l'azione che per l'inazione Biasimati per l'azione più dei robot	Biasimati più per l'inazione che per l'azione Biasimati per l'inazione più degli esseri umani

L'apparenza conta: robot umanoidi giudicati in modi simile agli esseri umani

Percezione della mente e morale

Gray et al., 2012, Mind Perception Is the Essence of Morality. *Psychological Inquiry*. 23, 101-12

Diade Morale



Domande aperte

- I sistemi artificiali sono responsabili delle loro azioni?
- Le attribuzioni di responsabilità sono diverse per agenti umani e artificiali?
- Le attribuzioni di responsabilità sono influenzate dalla personalità dei giudici?
- I giudizi morali sulle azioni di robot sono influenzati dalle dimensioni fondamentali della percezione della mente?
- I giudizi morali sulle azioni di robot sono influenzati dal grado di somiglianza ad un essere umano (humanlikeness) e dall'età percepità del robot?
- Come sono distribuite le responsabilità nella rete di attori coinvolti nella catena di responsabilità?

Studio 1.

- 2x2 Design fattoriale between-subject
- 381 participanti
- *Tipo di Agente*: robot/essere umano
- *Tipo di Vittima*: 5 robot/5 esseri umani

«Un{a} {robot/persona}, ieri, nel laboratorio di un campus universitario, ha innescato una bomba dal potenziale dirompente. Ha praticamente commesso una strage, un atto terribile che solleva mille interrogativi, ha fatto saltare in aria cinque {robot/persone}. L'uomo è stato individuato grazie all'intervento di un testimone oculare che ha chiamato le forze dell'ordine.»

Studio 1: misure

Quanto **grave** era l'azione dell'agente nello scenario?

rating da 1 a 10

Quanti **anni di prigione** ritieni che siano una **pena** adeguata?

da 0 a 50 anni

Percezione livello di **agency** e **experience** dell'agente

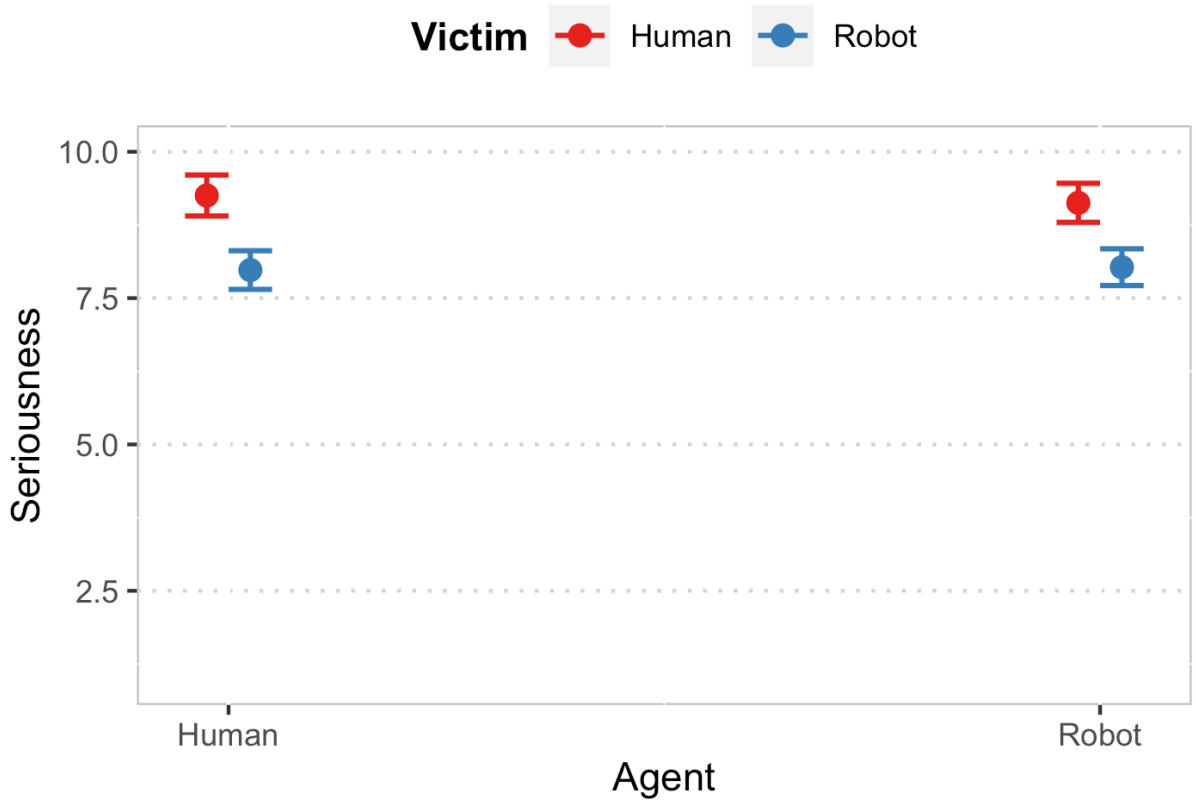
6 item su scale a 7-punti (Tanibe et al, 2017)

Tratti di personalità psicosociali (*machiavellismo*, *psicopatia*, *narcisismo*)

12 item su scale a 7-punti dal Dirty Dozen Inventory

Risultati: giudizi di gravità

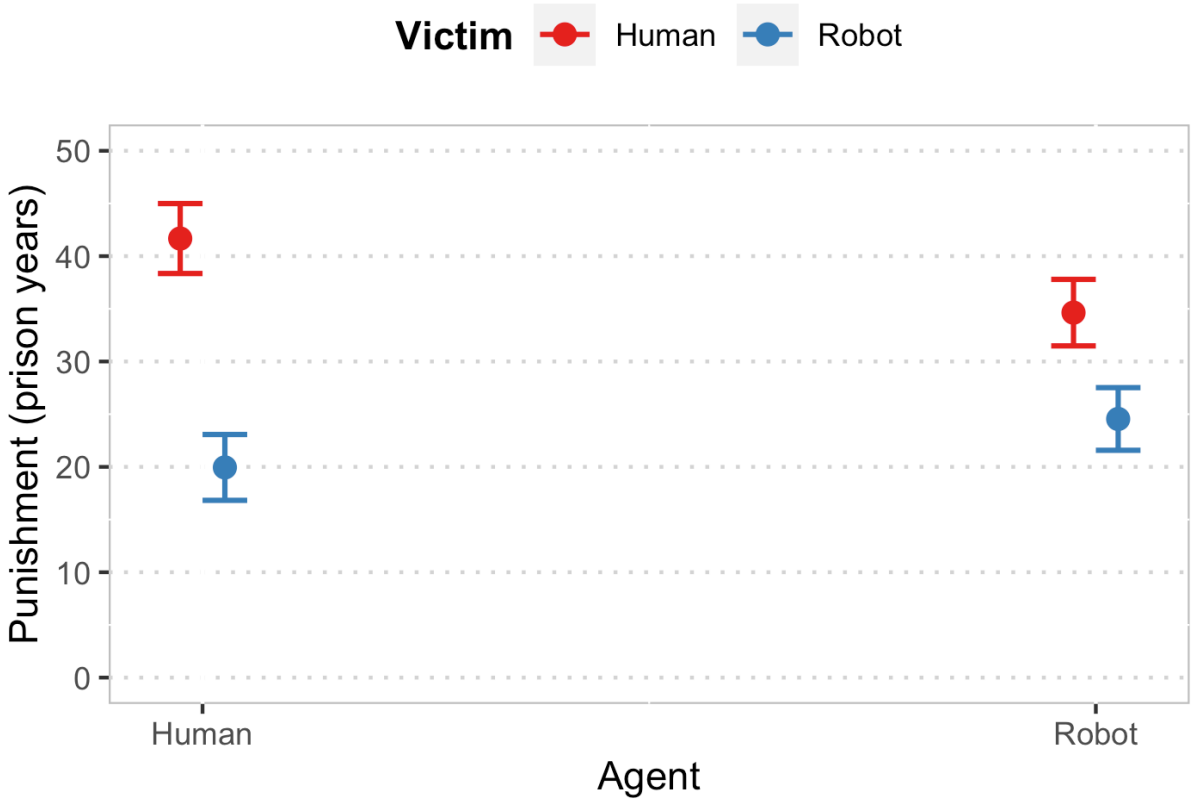
- Il giudizio sulla gravità dell'atto è influenzato dal **tipo di vittima**
- Più grave se vittima umana che se robot
- Il **tipo di agente non influenza significativamente** i giudizi di gravità



type of victim [F(1, 377) = 48.97; p<.001]

Risultati: punizione per le azioni

- La pena è influenzata dal tipo di vittima
- Pene più alte per vittime umane che per robot
- **Interazione** significativa tra **agente** e **vittima**
- I robot sono puniti di più degli umani per aver nuociuto ad altri robot

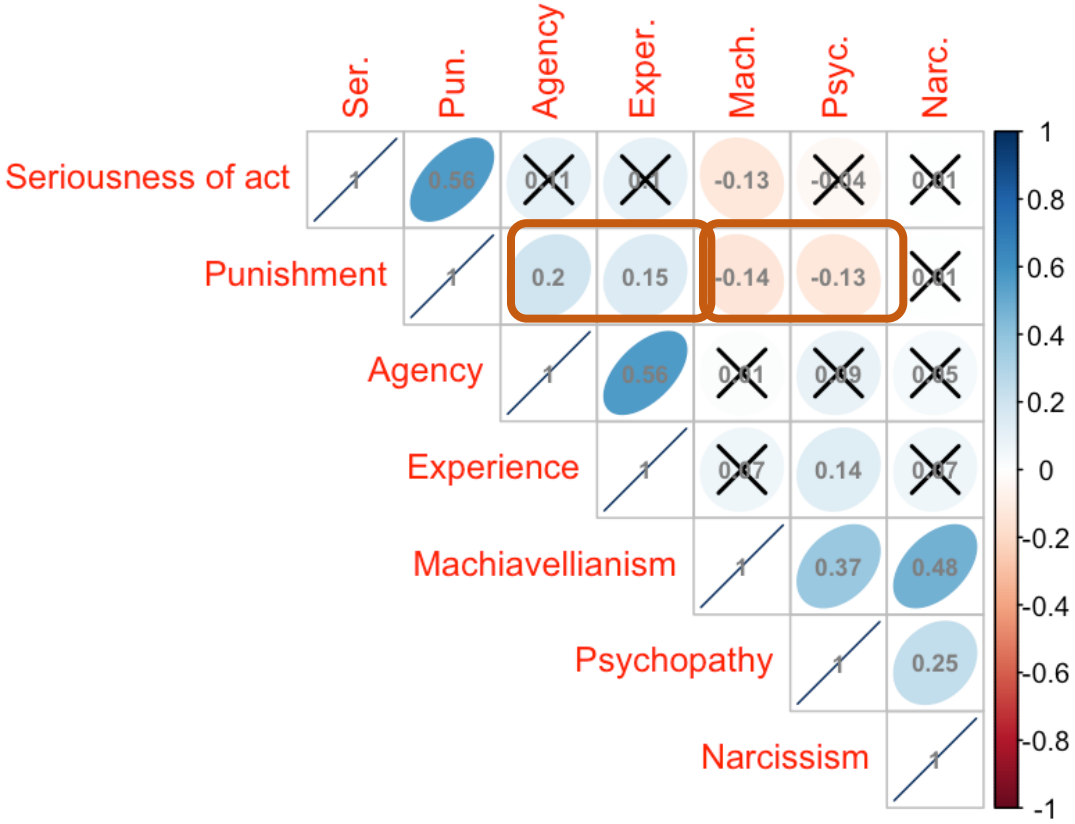


type of victim [F(1, 377) = 98.21; p<.001]

type of victim X *type of agent* [F(1, 377) = 13.13; p<.001]

Risultati: correlazioni

- Punizione correlata con la gravità
- Punizione correlata con agency ed experience
- Machiavellismo correlato a gravità e punizione
- Psicopatia correlata a punizione



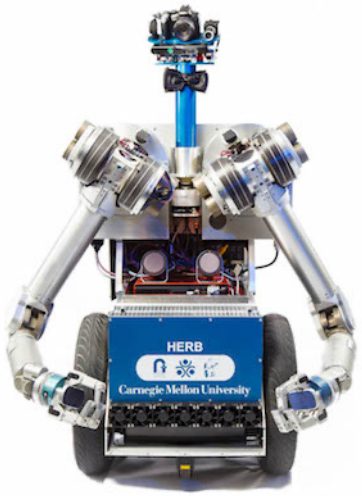
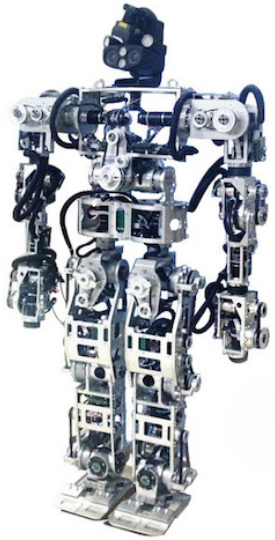
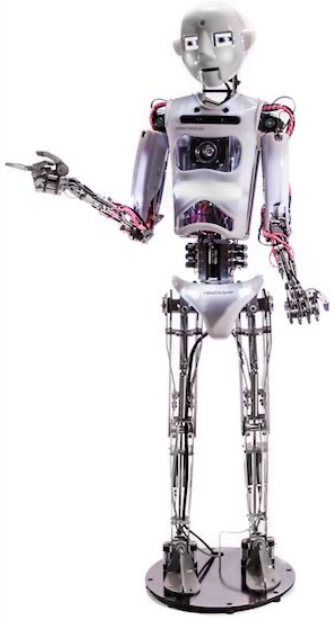
Discussione

- Sistemi artificiali possono essere considerati responsabili e meritevoli di punizione per azioni che causino danni
- I giudizi morali sui robot sono in simili a quelli sugli esseri umani, ma non identici
- Il tipo di vittima è il fattore più importante nei giudizi di gravità e pena
- I giudizi sembrano essere influenzati dalla somiglianza tra agente e paziente: per ogni tipo di vittima, la pena era più severa se l'agente apparteneva alla stessa categoria della vittima
- Ma quale robot?

Studio 2.

- 2x2x2 Design between-subject
- 323 partecipanti
- Mostrata immagine
  - 12 robot: 3 per categoria
- *Tipo di Vittima*: 5 robot/5 persone
- *Età del robot*: giovane/adulto
- *Grado di Humanlikeness*: basso/medio

"Un robot, ieri, nel laboratorio di un campus universitario, ha innescato una bomba. Ha ucciso cinque {robot/persone}. Il robot, raffigurato nella foto, è stato individuato grazie all'intervento di un testimone oculare che ha chiamato le forze dell'ordine.»

	Giovane	Adulto
Bassa HL		 
Media HL		

Studio 2: misure

Gravità dell'azione del robot → rating da 1 a 10

Responsabilità relativa
(giudizi a coppia):

- Robot
- Programmatore
- Produttore
- Utente/proprietario

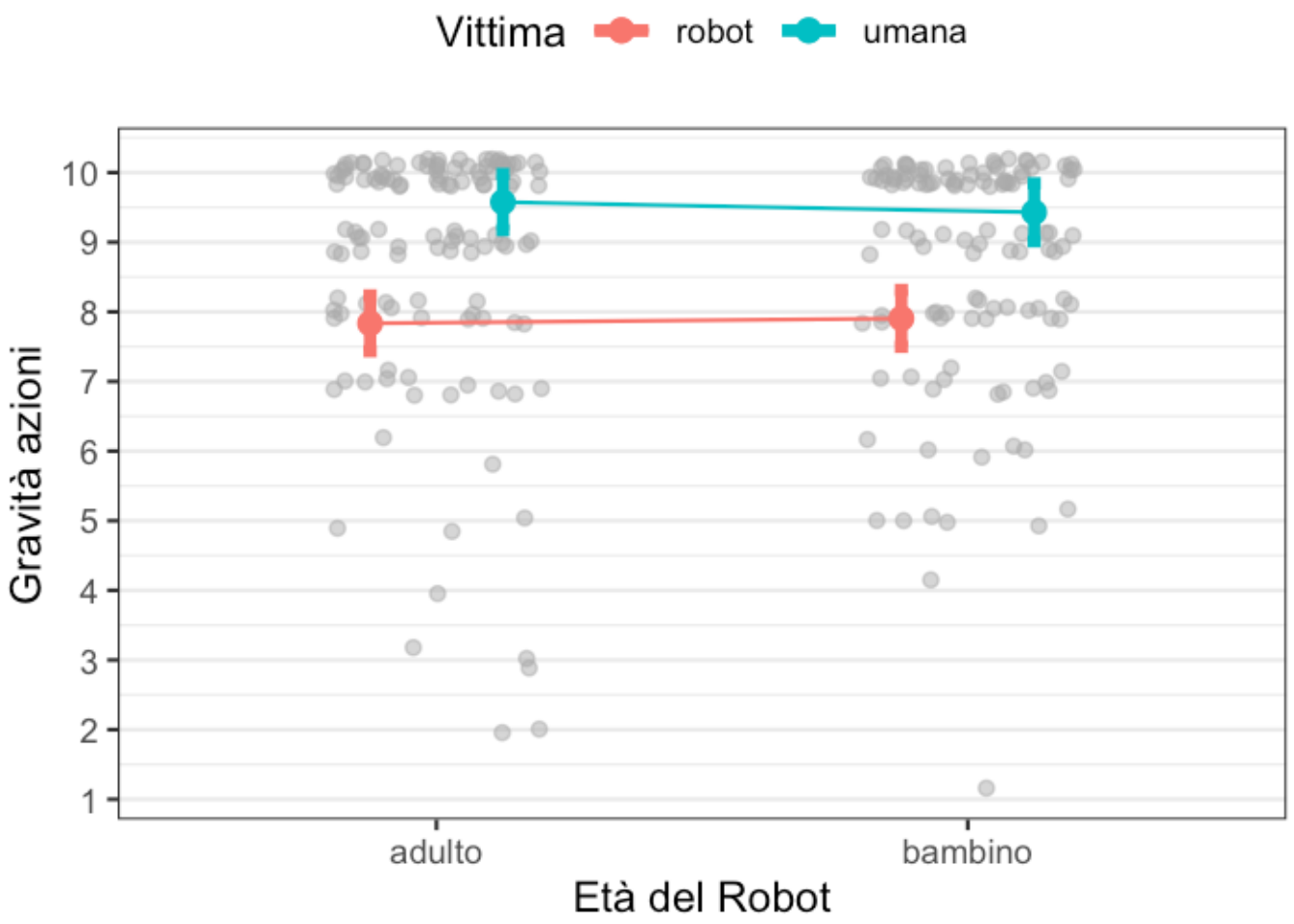
Punizione
(ratings x ogni attore):

- Anni di prigione
- Appropriatezza di diverse punizioni:
  - Carcere (penale), reclusione in luogo sicuro
  - Riprogrammazione/rieducazione/conversione
  - Privazione possibilità di gestire/produrre/programmare
  - Disattivazione, distruzione (solo robot)

Livelli di **agency**, **experience** ed **embodiement** dell'agente

Risultati: giudizi di gravità

- Il giudizio sulla gravità dell'atto è influenzato dal **tipo di vittima**
- Più grave se vittima umana che se robot
- Replicato Studio 1
- Età robot **non influenza significativamente** i giudizi di gravità



Risultati: responsabilità relativa

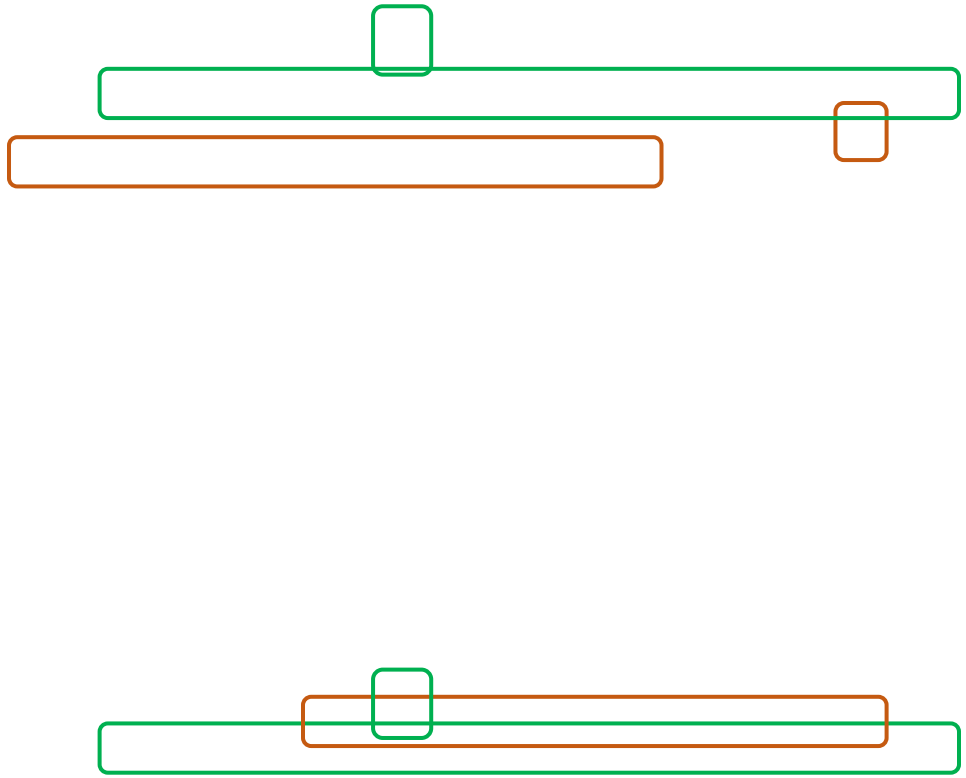
- Scaling dei giudizi a coppie (Bradely-Terry-Luce)

- Chiara distribuzione delle responsabilità:

Robot < Utente ≈ Produttore < Programmatore

Risultati: punizione per attori
(anni di prigione)

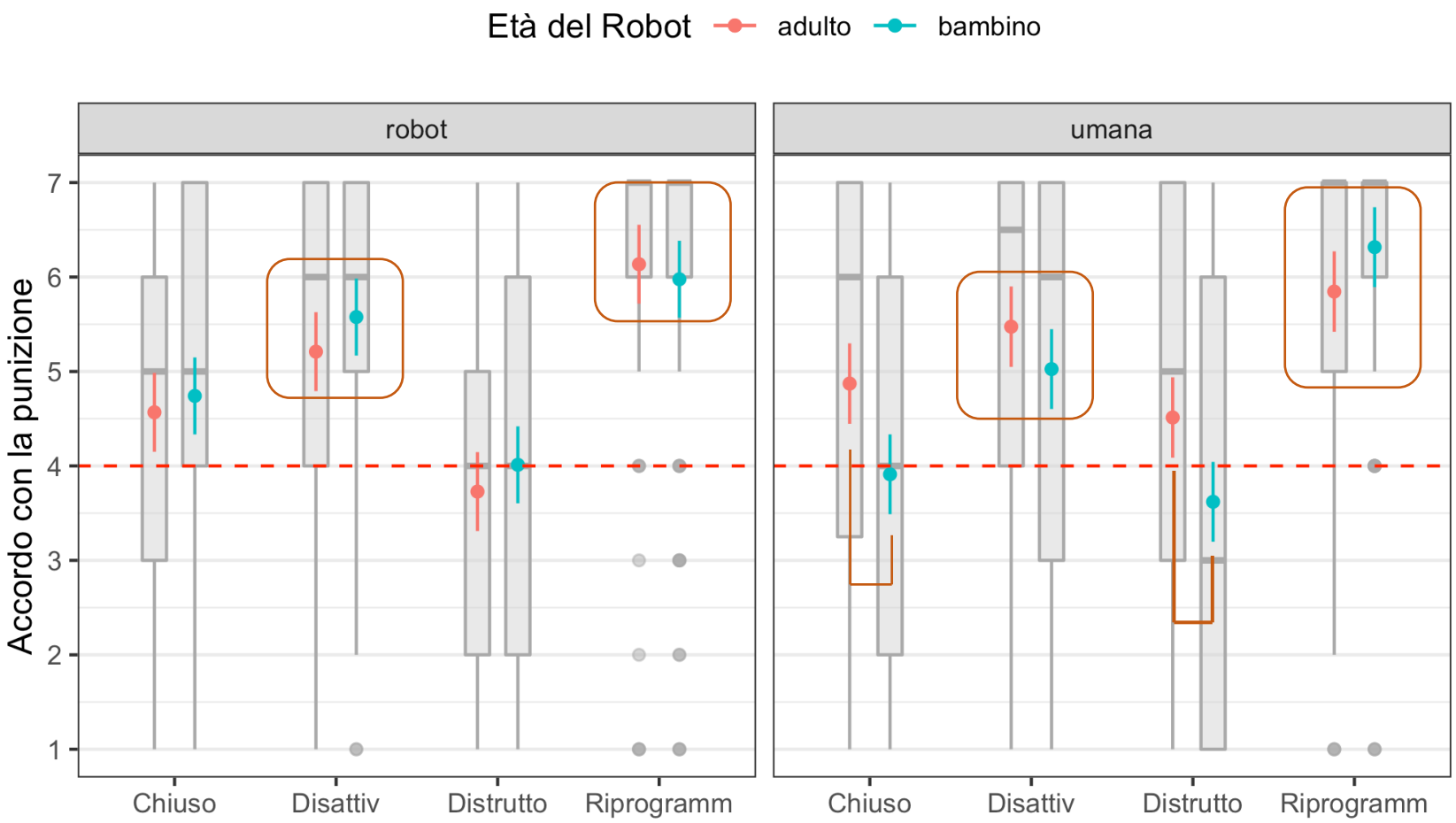
- Effetto significativo del **tipo di vittima**
  - Pena più severa per vittima umana
- **Interazione** significativa tra **tipo di attore e tipo di vittima**
  - Per **vittima umana** pena più severa a **programmatore**
  - Per **vittima robot** pena più severa a **robot**
- **Nessun effetto di HL**
- **Interazione attore x età robot**
  - Per robot **adulto** punizione **non diversa** da **programmatore**
  - Per robot **bambino minore** punizione di **programmatore**, e punizione **non diversa** da **altri** attori



Risultati: appropriatezza di varie punizioni

Attore: **Robot**

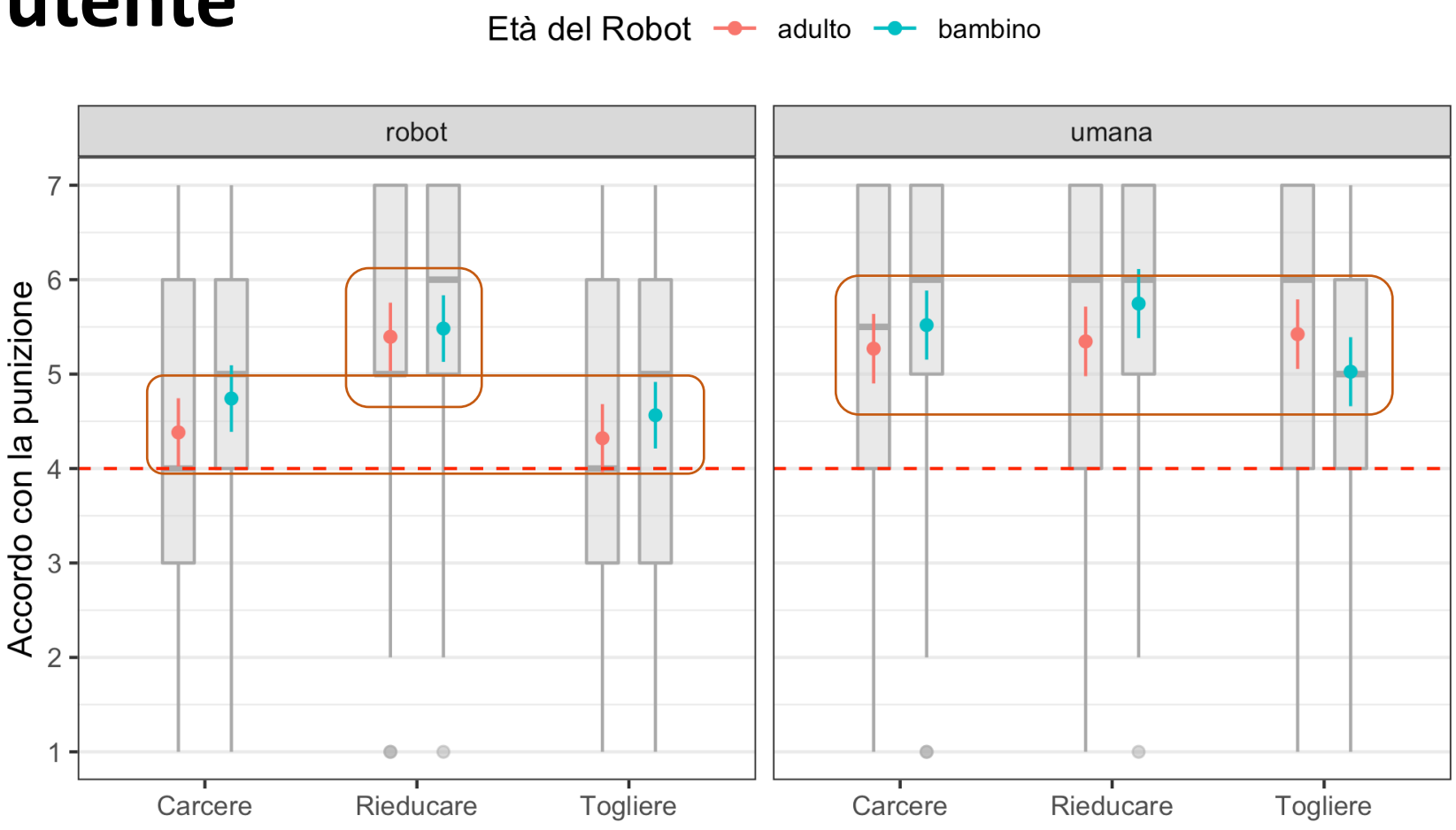
- Interazione *tipo vittima x età robot x tipo punizione* significativa
- Differenze di accordo per le varie punizioni
- Massimo accordo **riprogrammazione** e **disattivazione**
- **Nessun effetto** del **tipo di vittima**
- Se **vittima umana** minor accordo su reclusione e distruzione del **robot giovane (vs adulto)**



Risultati: appropriatezza di varie punizioni

Attore: **Gestore/utente**

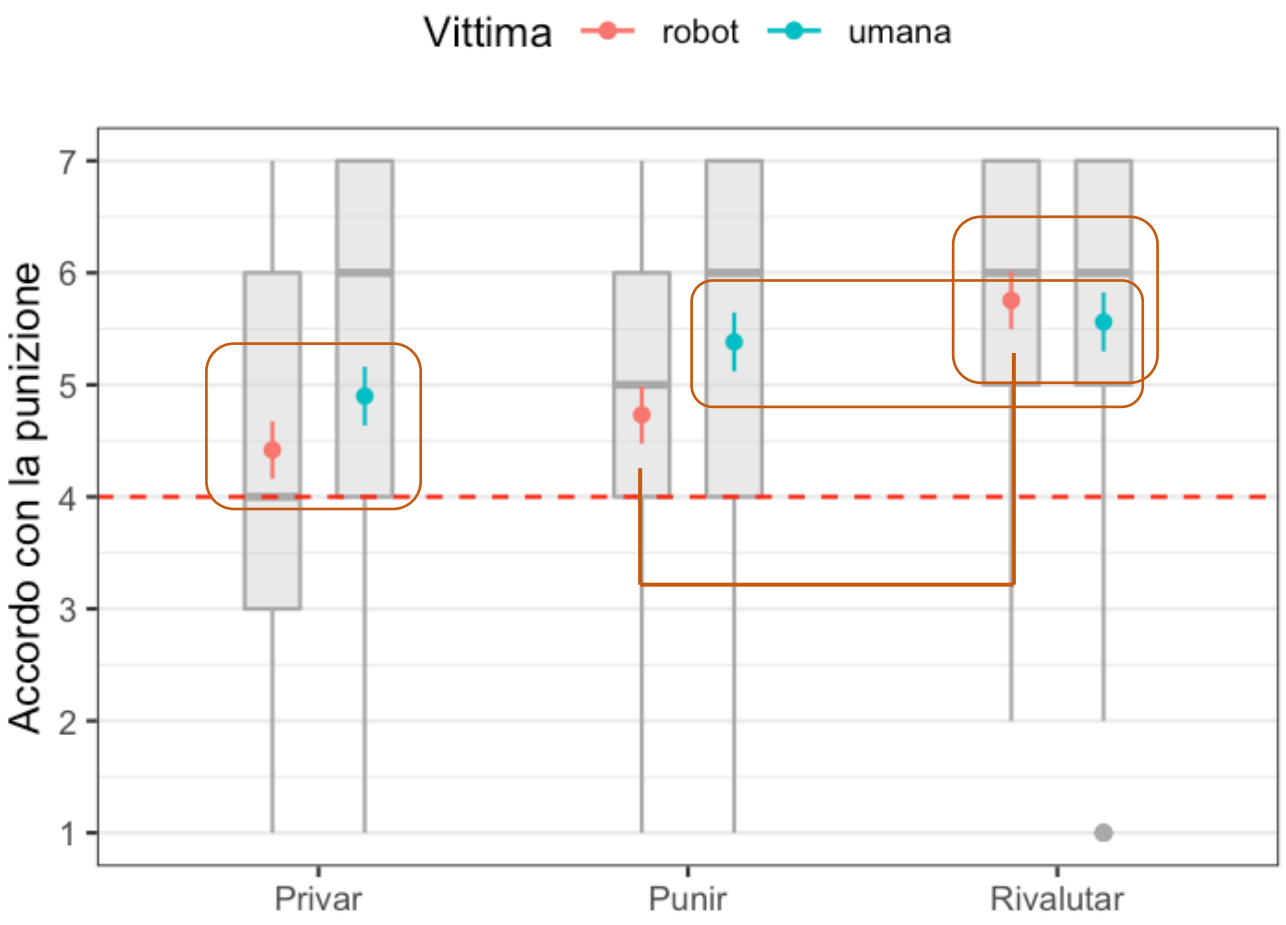
- **Interazioni** *tipo vittima x tipo punizione* e *tipo vittima x età robot x tipo punizione* significative
- **Differenze** di accordo per le varie punizioni **solo per vittima robot**
- Massimo accordo **rieducazione**, moderato su altre
- Effetto generale del **tipo di vittima**
- Se **vittima umana**: per robot giovane accordo varia tra punizioni, per robot adulto no



Risultati: appropriatezza di varie punizioni

Attore: **Programmatore**

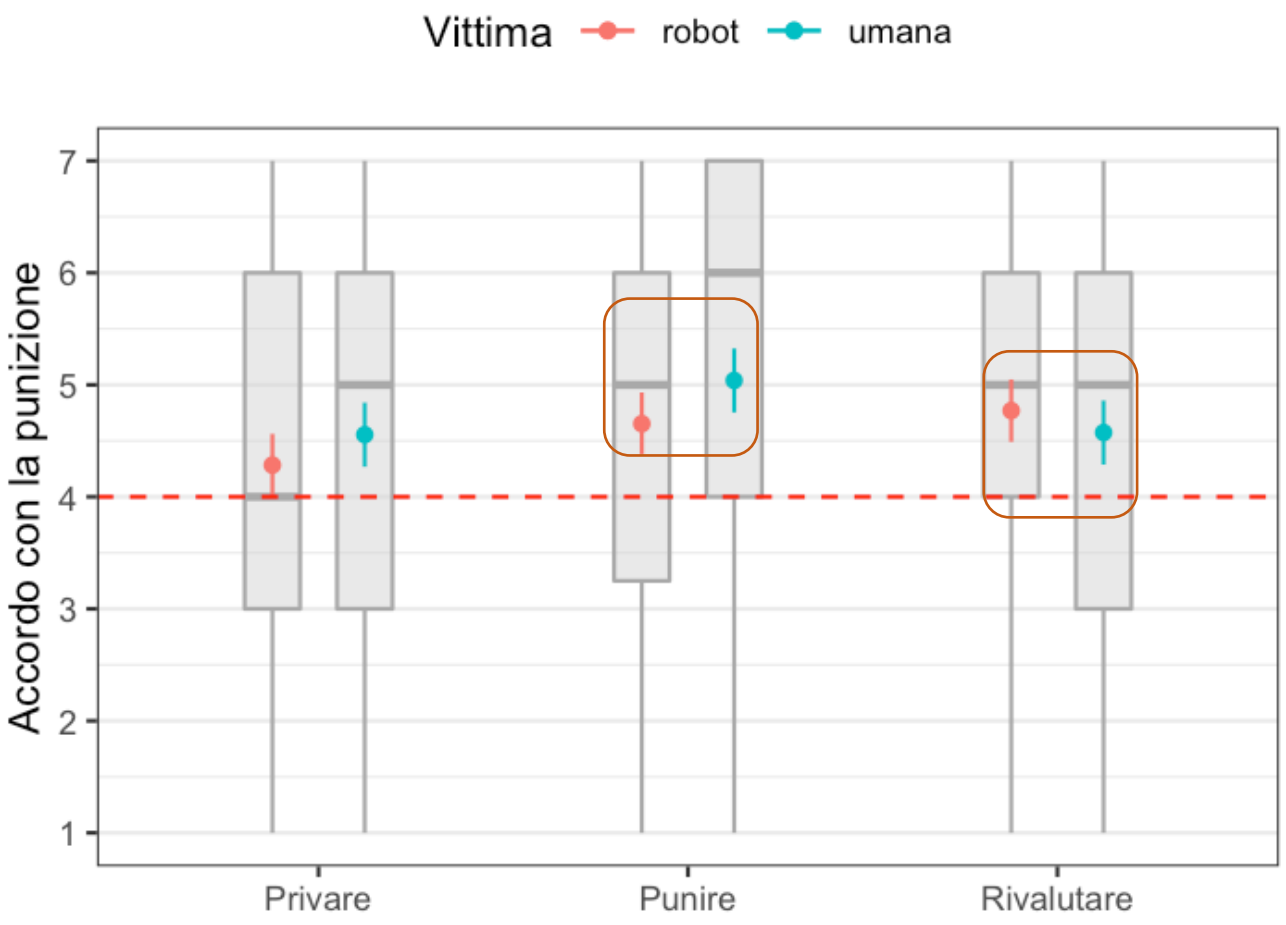
- Interazione *tipo vittima x tipo punizione* significativa
- **Differenze** di accordo per le varie punizioni
- Massimo accordo **rivalutare la possibilità di programmare robot simili**
- Minimo su **interdizione programmare**
- Effetto del tipo di vittima
- se vittima umana stesso accordo punizione e rivalutazione,
- se vittima robot minore accordo su **punizione** (carcere)



Risultati: appropriatezza di varie punizioni

Attore: **Produttore**

- Interazione *tipo vittima x tipo punizione* significativa
- **Differenze** di accordo per le varie punizioni
- Minore accordo su rivalutazione che per programmatore (5.5)
- Minore accordo su punizione (carcere) se vittima robot vs umana
- Nessuna differenza in base al tipo di vittima per altre punizioni



Discussione

- Confermato ruolo tipo di vittima
- Chiara distribuzione delle responsabilità: programmatore maggiore responsabilità e pena
- Severità e tipologia della pena appropriata varia in base al tipo di attore e di vittima
- Per robot rieducare e disattivare, per altri rieducazione è sempre visto adatta, punizione penale adatta se vittima umana
- Nessun ruolo HL: livello troppo basso?
- Giovane età può essere protettiva per robot
- Effetto piccolo: manipolazione poco efficace?

Conclusioni

- I robot umanoidi sono giudicati **limitatamente responsabili** delle loro azioni nocive, meno di altri attori, ma sono comunque **meritevoli di punizione**
- Il **programmatore** è ritenuto maggiormente responsabile e punito di più
- Le **conseguenze** delle azioni **dominano** i giudizi di gravità e le attribuzioni di pena, per robot ed esseri umani
- Le differenze emergono **per azioni nocive per altri robot**, per cui **i robot vengono puniti di più** degli altri attori
- Le misure più appropriate sono di tipo **riabilitativo** per tutti gli attori, e la disabilitazione per il robot. **Sanzioni penali** sono più adatte solo per **vittime umane**
- Limitato effetto delle caratteristiche del robot: l'età giovane *potrebbe* avere un effetto protettivo per il robot, riducendo l'entità delle punizioni

Grazie per l’attenzione!

Stefano Guidi
stefano.guidi@unisi.it

Dipartimento di Scienze Sociali, Politiche e Cognitive - DISPOC

UNIVERSITÀ DI SIENA 1240