Consiglio Nazionale
delle Ricerche

CNR ISTC

ISTITUTO DI SCIENZE E TECNOLOGIE
DELLA COGNIZIONE

# Digital Twins for Trustworthy Human-Robot Interaction

## *Towards an Integrated Socio-technical Knowledge System*

Luca Biccheri
ISTC-CNR, Laboratory for Applied Ontology, Trento (Italy)

# Overview

I) Problematizing trust in HRI

II) Ontological model for trust in HRI

III) Robotic case experiment (ISTC-CNR,Trento-Roma)

IV) Digital Twing approach

# I) Problematizing trust in HRI

# What's the point of trust in HRI?

- Within the Human-Robot Interaction (HRI) community, scholars claim that robots need to perform two main tasks while interacting with people:

  *a)* **elicit** the right level of trust on the human side

  *b)* **interpret** how much people trust the robots

- Evidently, *b)* implies *a)*, but not vice versa, i.e. to elicit trust appropriately, we first have to **understand** what this human attitude is, as well as **how it is expressed**, and not least how people **adjust** it as time goes by (i.e. trust is dynamic)

- Trust in HRI takes its cue from HCI (Human-Computer Interaction) and HHI (Human-Human Interaction); while it has various degrees of overlap with these disciplines, HRI has its own peculiarities

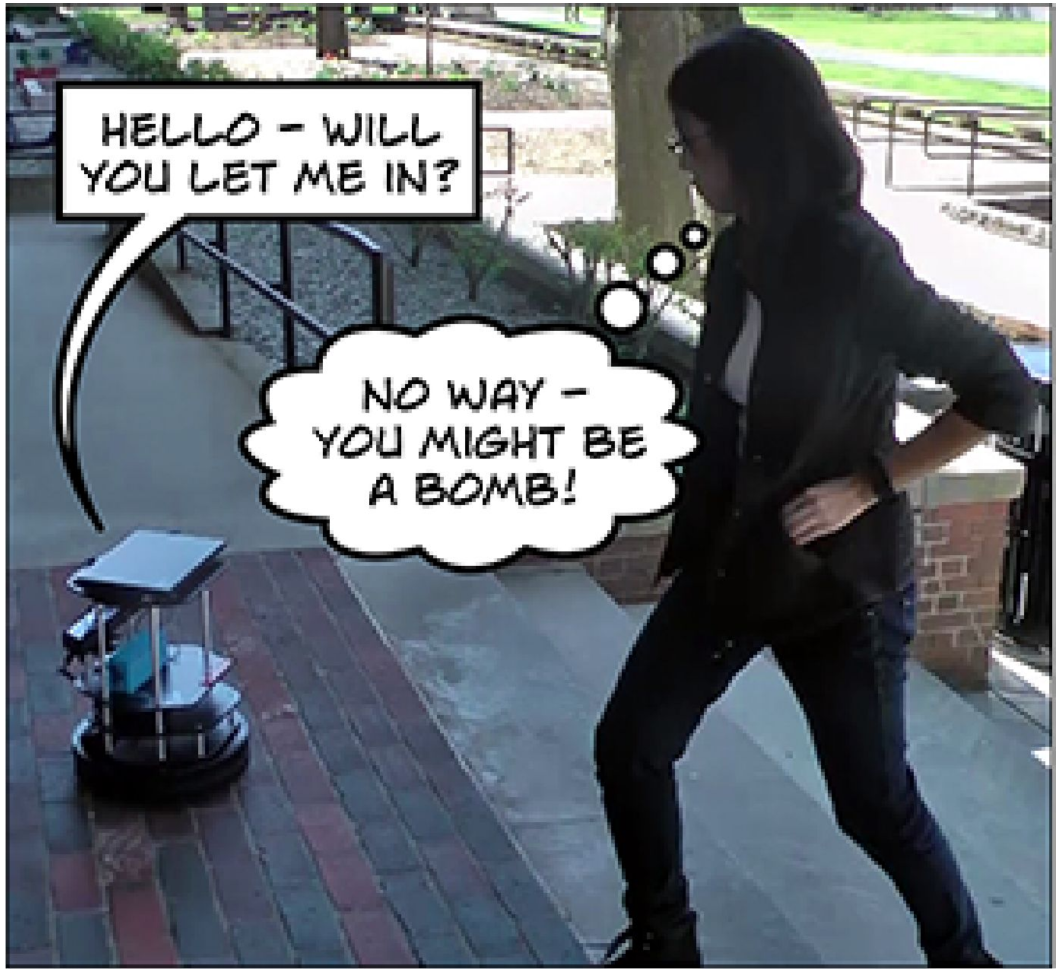
# Balancing trust

- Intelligent systems ask for greater **flexibility** in plan execution
- To better interact with users, robots have to understand people's **behaviour**, infer their trust and formulate **appropriate responses**
- To formulate appropriate responses we need to understand which **signals** are interpreted as significant to trust
- Overall, we need to avoid both:
  - **Overtrust**
  - **Mistrust**

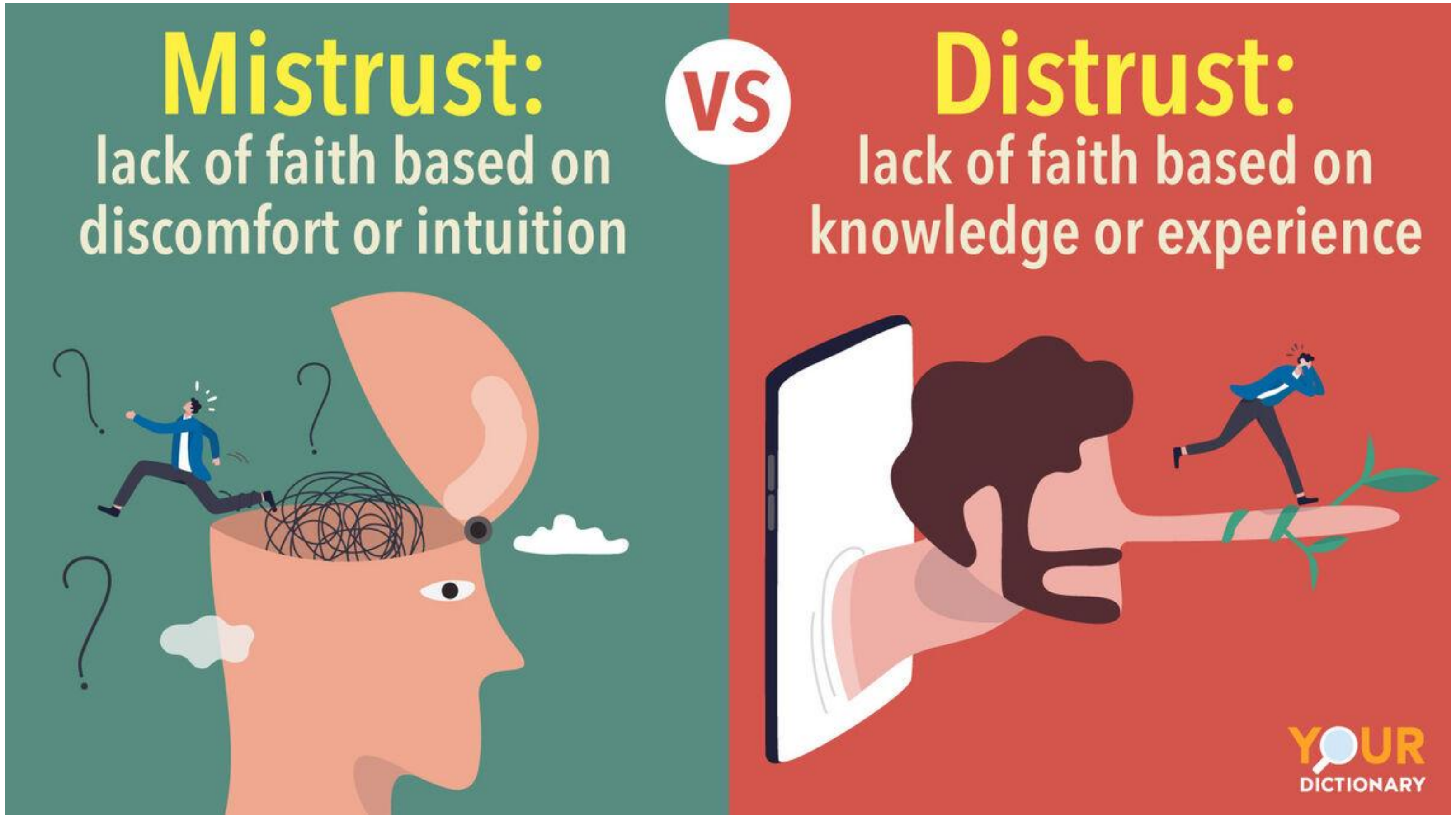
# Overtrust

- People tend to ‘blindly’ follow (possibly defective) machines in risky scenarios (Robinette, P. et al. 2016)

# Mistrust



(Booth et al., 2017)


Mistrust:
lack of faith based on discomfort or intuition
VS
Distrust:
lack of faith based on knowledge or experience
YOUR DICTIONARY

* While **Distrust** can be seen as the opposite of trust, **Mistrust** is a different attitude (Castelfranchi et al., 2009)

In addition, **Overtrust** could also be a different attitude from trust

# Trust is a portmanteau and ‘abused’ concept

- “although trust is an obvious fact of life, it is an exasperating one […] Trust works in practice, but not in theory ” (Hollis, 1998)

- “the idea of trust has been used so widely and loosely that it risks creating more confusion than clarity ” (Guinnane, 2005)

- There is therefore a problem of **intertheoretical coherence** that endangers the understanding, evaluation and comparison of empirical results about trust in HRI

# Some results on errors affecting Trust in HRI

Are these findings inconsistent?

- Trust mostly depends performing functions properly, thus errors **strongly affect the trust level** (Muir & Moray, 1996)

- Some random errors performed by humanoid-robots **helps building trust** (Salem et al. 2013)

- Errors (of any kinds) have **no significant impact** on the people’s trust towards humanoid robots (Salem et al. 2013)

# Anthropomorphism degree
## (Can we have a general model?)

- <u>Robotic arm</u>

scope ⇒ Industry

principal concerns: safety, reliability, availability, maintainability, etc

- <u>Social robot</u>

scope ⇒ Social Services

principal concerns: safety, reliability, availability, maintainability…+ **animacy, likability, perceived intelligence, etc**

# Physical, Ethical and Legal limits

- Physical (Parenti et. al., 2023):
    - Training (physically) the robot is an **intensive** and **deteriorating** task for its motors and actuators
    - The battery level has to be high enough  to provide power to the motors (in case of intensive use, the battery drains in few hours)
- Ethical and Legal:
    - The use of robots must ensure the **safety of personnel** while preserving the **integrity** of the robot itself

# To sum up

Conceptual issues:

- Ontological: What is trust?
- Anthropomorphism degree: Can we have a general model?

Usage issues:

- Physical: e.g. mechanical wear
- Ethical and Legal: e.g. safety

Measurement issues:

- Both the conceptual and usage issues hinder the identification and evaluation of trust

# II) Ontological analysis for trust in HRI

# Enhancing Trust in HRI

- Research objective: providing a sound  and operational notion of trust to be applied in HRI, exploring the following point:
  - **Integrating different** (yet, compatible) **paradigms** (Engineering + Social Sciences) to get a flexible model for the evaluation of trust in HRI

- Expected advantages:
  - Ontological level: a clear and flexible semantic framework that can be reused and adapted in different robotics case studies on trust
  - Empirical level: estimation of those 'signals' that are most relevant to understand various levels of trust and related notions

## ○ Dependability (Engineering)

- Software/hardware
- Critical System (e.g. nuclear plant)
- Device (e.g. laptop)

## ○ Trust (Social)

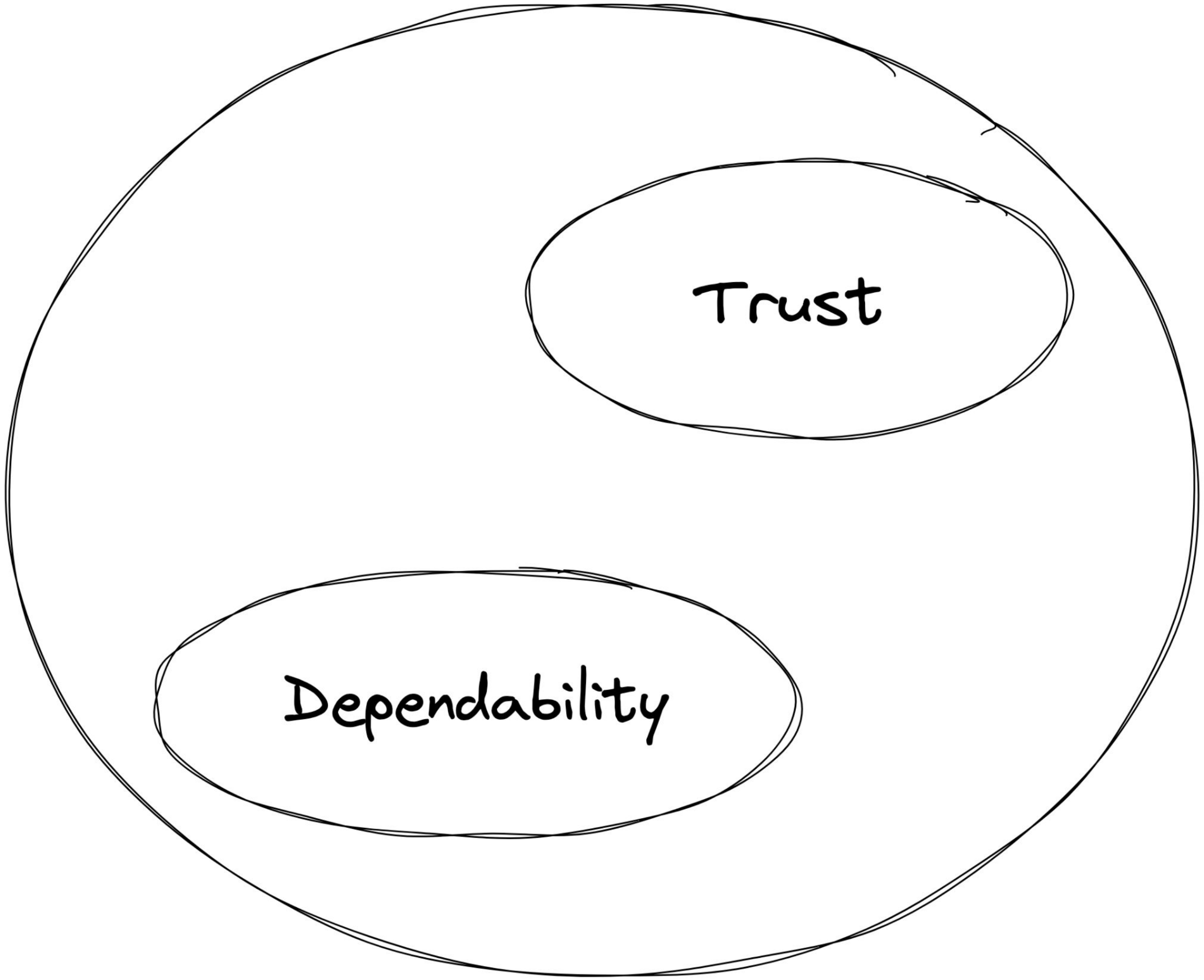
- Psychology
- Sociology
- Economics
- Philosophy

*

**Dependability** ⇒ artefactual/technical aspect

**Trust** ⇒ interrelational aspect

# Instrumental Dependence



1. Instrumental dependence (ID) is just about the goal achievement from a **means-end** point of view

2. In addition to trust and dependability, it covers also the case of cooperation; more generally **can be extended** to every form of goal-oriented dependency

3. The adjective ‘instrumental’ should not be understood, *per se*, as implying any form of exploitation based on self-interest (although exploitation is indeed a form of ID); rather, it should be linked to the fact that an agent sees, from the action-planning point of view, certain entities (whether human or artificial) as **means** to an **end**

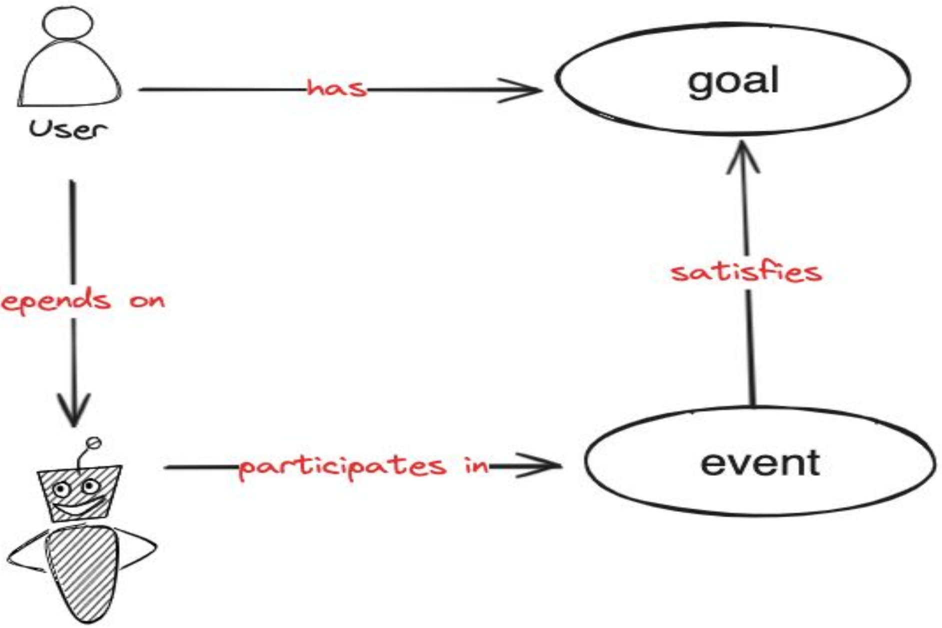
# Instrumental Dependence

(Biccheri et al., 2023)

- We found a set of "family resemblances" à la Wittgenstein that are shared between dependability and trust, i.e. {**goal**, **dependence, action**}

$$ID(x,y,z,w) \rightarrow APO(x) \land POB(y) \land PD(z) \land Goal(w) \land \exists t.(sat(z,w,t) \land PC(y,z,t))$$

*" An agent's goal can be satisfied only through the occurrence of an event in which the entity towards which such an agent is instrumentally dependent participates"*



# Instrumental Dependence (ID) in HRI

- Within HRI scenarios, Instrumental dependence (ID) can be used to represent that a person trusts a machine regarding some goals, where the former plays the role of the trustor and the latter the role of the trustee (Castelfranchi & Falcone, 2010)

- E.g. " Tony trusts the robot dog to cross the road"

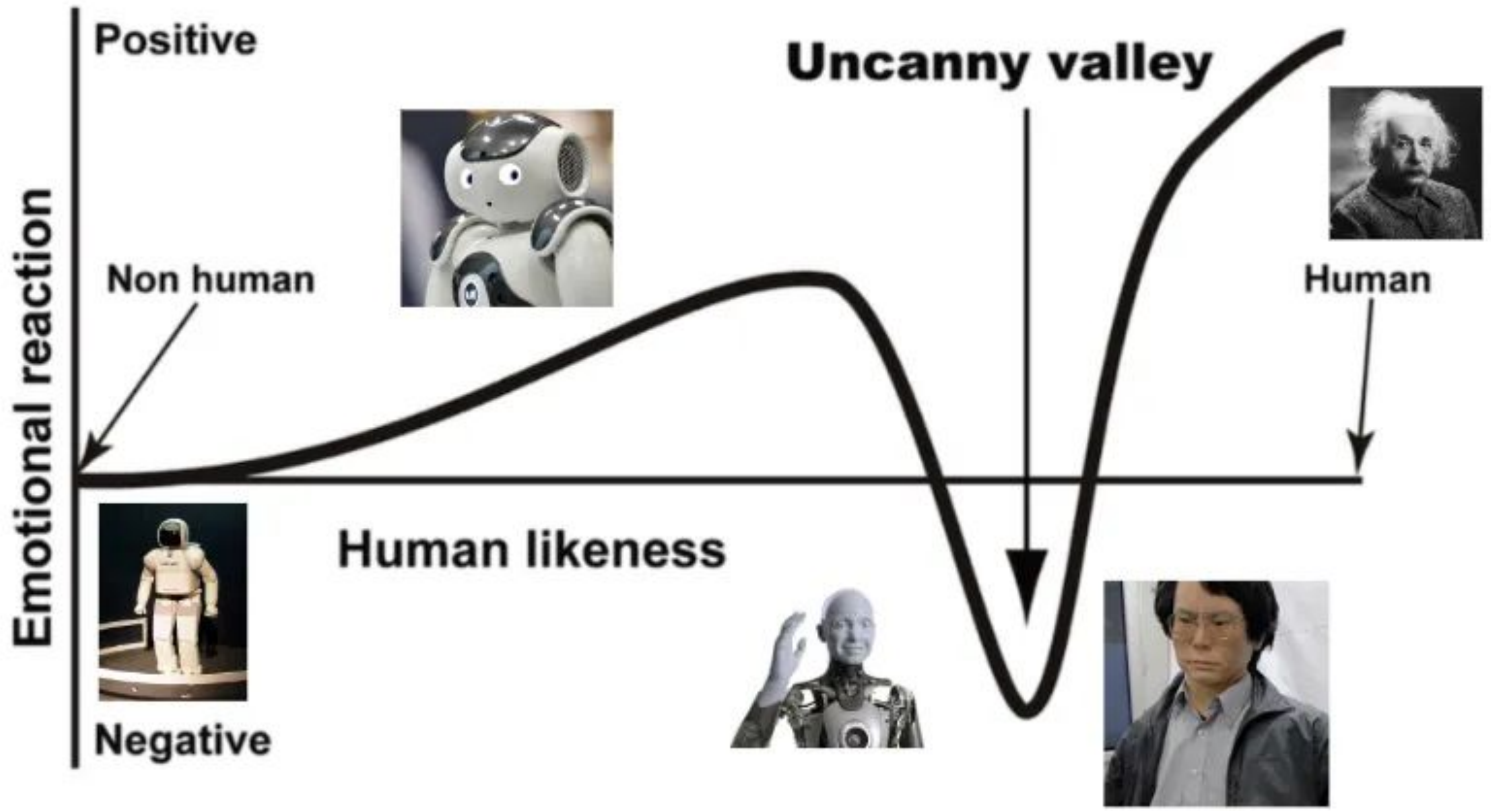
- If $x$ trusts $y$, then $x$ is instrumentally dependent on $y$

- There is a goal $z$ that $x$ can satisfy by means of the actions performed by $y$

- $z$ = goal = internal state (e.g. mental state)

# ID is necessary (but non sufficient) for trust

As is well known, the trust towards robots goes far beyond goal achievement (Bartneck et al, 2009)



# Shallow conceptualization of Trust in HRI

- Studies in HRI might conflate trust and Instrumental Dependence (ID) under many respects

- ID is a **necessary** condition for trust

- ID is an **internal state** of the agent, e.g. a mental state

- Trust is also typically identified with a **mental state** in HRI (as elsewhere)

- Yet, remember ID **is wider** than trust

- Empirical studies in HRI often measure trust based on, among other things, the willingness to **continue collaborating/interacting** with the robot despite its faults/errors (e.g. see Salem et al., 2013)

- However, this does not necessarily indicate the presence of trust, although it may refer to ID in a specific form (e.g. cooperation)

# Overcoming the subjective view on Trust

- The idea of trust as a mental state in HRI is often **coupled with** experimental settings designed following some game-theory assumptions on trust

- "A substantial portion of related work employ so-called economic trust games to measure the level of trust placed in an agent. However, since these games only model very specific trust situations related to monetary gain or loss, findings from such studies **cannot be easily generalized** (my emphasis) [...] Therefore, one of the major challenges when investigating trust in social HRI is to design study scenarios that **demand trust in a natural and realistic environment** (my emphasis), while ideally incorporating a variety of tasks which tap different dimensions of trust" (Salem et al., 2013)

- Moreover, as it is often said, trust typically requires facing some kind of **risk.** The notion of risk indeed **evocatively** underlines many elements that are at stake in trust-contexts

- However, the heavily **subjective and risk-centered** conceptualization of trust might reflect more the contemporary attitude of the **Homo Economicus**, rather than a pervasive feature of trust-contexts throughout human history (cf. Milness,2019)

# The Humean Reflection

- During the age of Enlightenment, philosophers were busy with the status of contracts, promises and other forms of obligations/negotiations. In this context, Hume's reflection on trust is oriented to justify the base of **informal** behaviours/**conventions** (Milness, 2019):

  "the actions of each of us have a reference to those of the other, and are perform'd upon the supposition, that something is to be perform'd on the other part. Two men, who **pull the oars of a boat (**my emphasis**)**, do it by an agreement or convention, tho' they have never given promises to each other" (Hume, 1978)

- In this respect, trust can be seen as a **relational attitude** and a **precondition** for interaction, rather than a specific mental state

- Along these lines, in short, one may say that trust is a **conventional** (i.e. informally codified) **behaviour** performed by an agent to **reciprocate** the behavior of another agent in some context (not necessarily risky contexts)

# Affording trust

- An animal “can afford eating or being eaten, copulation or fighting, nurturing or nurturance” (Gibson 2015, p. 36)

- May not yet be considered *full* social affordances inasmuch as **they elicit behavior** in another animal, but **not** necessarily **social interaction**

- To have social interactions the participatory animals must have a **minimal responsiveness to each other** as self-moving beings and their **behaviors must be mutually constrained** while they are **engaged in an activity** (Carvalho, 2020)

- In this respect, the attitude of trust can be **the result of a response** to certain kinds of **specific social affordances** that **promote** social interactions

# Why is this view relevant for HRI?

- Social robots can serve social purposes (e.g., elderly care), among other things, **precisely because** of their abilities to **replicate** certain intimate human-human interactions based on conventions (à la Hume)/ social affordances

- Where do we have to look for such social affordances? And how can these be successfully employed in experimental settings on trust in HRI?

- We shall resort to the goal-oriented theory of Social Signaling Processing (Vinciarelli et al., 2009)

- Social affordance ≅ Social Signal

# Looking for Trust-signals

- Robots should be able to recognize and respond to changes in user trust over time, mitigating both **overtrust** and **mistrust** by the user’s side

- To do so, we need first to identify those **signals** that, so to speak, form the lexicon of trust

- Once identified, we can use signals to study/experiment with trust both in virtual (e.g. Digital Twin) and physical environments

# Signals

(Poggi, 2013)

- Communication entails signals (i.e. physical stimuli)

- **Codified signals** (signal-meaning pairs) are produced by a **sender** and perceived by an **addressee** thanks to various modalities/channels

- Codified signals are generally acknowledged for words and gestures, but they also come in **other modalities**, i.e subparts of the body, e.g. eyes are specific communication systems

# Optology = ‘Phonology’ of Gaze

(Poggi, 2013)

- A lexicon of codified signal-meaning pairs performed through gaze
- A signal result from the combination of values assumed by *n* parameters
- A list of meaningful parameters:
  - Eyebrows rising ⇒ perplexity/surprise
  - Eyes humidity ⇒ joy/enthusiasm
  - Pupil dilatation ⇒ sexual arousal/anxiety/ rage
  - Extended duration ⇒ threatening/defying
  - Direction ⇒ deictic/pointing
- Note that, typically, such signals are polysemic

# Signals of Trust

- People interpret persons who maintain eye contact as more trustworthy (Bayliss & Tipper, 2006)
- Apart from the eyes, many other **body signals** can be read as trustworthy, e.g.:
  - Smile at someone
  - Gently touch the shoulders
  - A relaxed sitting posture
  - Reach out to the hand to help
- We can also interpret some signals as **overtrust**
  - E.g. repetitively nodding in approval

# Signals of Mistrust

- As for trust, there is a whole vocabulary of non-verbal signals for **mistrust**,e.g.:
    - Constantly defying the gaze
    - Tendency to avoid interactions
    - Distance/ wider personal space

- Unintentionally emitted signals aka **non-verbal leakage**: feelings/thoughts/intentions that ‘leak’ from non-verbal channels (e.g. shaky limbs may communicate anxiety)

- Typically people **automatically look** for such signals while interacting with other people, inferring internal states of agents (Mutlu et al., 2009)

# The Lexicon of Trust

- Whenever **face-to-face** interaction is at stake trust is communicated through multiple channels

- In this respect, the communication system of trust is inherently **multimodal** leveraging two main modalities:

  - **Non-verbal** signals

  - **Verbal** signals: e.g. 'I am here to help you' expressing willingness/commitment

- Both types of signals can be emitted by the sender with varying levels of **intentionality** and **consciousness** and interpreted by the receiver in the same vain

- Many studies in HRI focused on verbal signals

- How robots' non-verbal signals may affect human decision processes and trust **remains understudied** and **poorly understood** (Parenti et al., 2023)

# An operational notion of trust

- We do not provide a **definition** of trust (e.g. "**Trust** ↔ something else follow")

- We assume trust signals are sufficiently patterned over time and space

- Positive correlation between sets of signals and trust

- So that we can express trust in terms of signals, e.g. a weighted sum

$$\text{Trust}_{\text{score}} = \sum_{i=1}^{n} \text{w}_\text{i} \times \text{Signal}_\text{i}$$

- It allows our model to be compatible with different approaches on trust

# IV) Robotic case experiment ISTC-CNR Trento-Roma

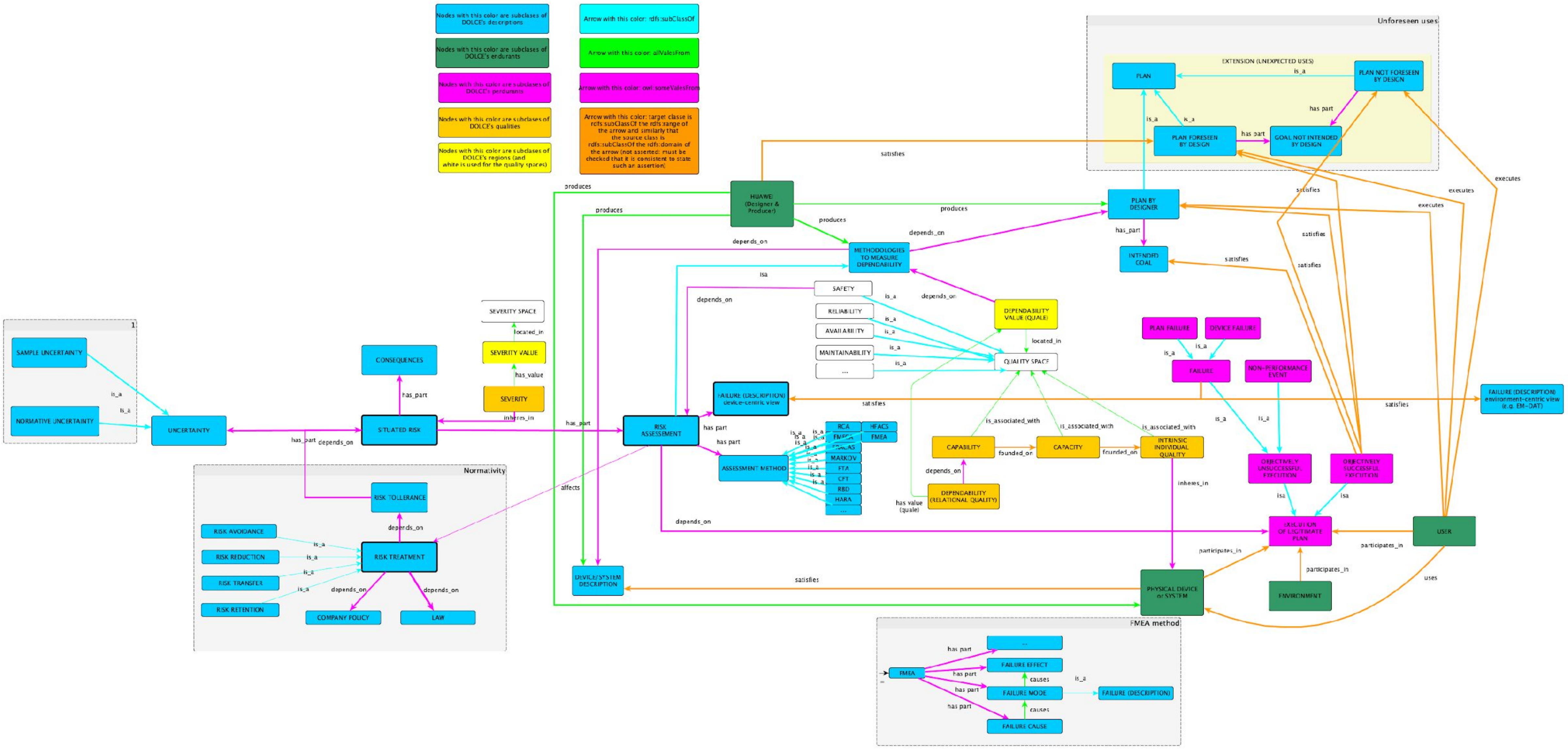
# Combining different paradigms

- When designing collaborative robots that can be trusted by humans we deal with a double aspect: i) **artefactual** ii); **interrelational**

- The **artefactual aspect** can be addressed by resorting to the metrics of *dependability engineering*, e.g. Mean Time to Failure (MTTF) = $\int_0^\infty R(t)\,dt$

- We realized a comprehensive ontology for dependable systems

# Ontology for Dependable Systems (DOLCE)


Nodes with this color are subclases of DOLCE's descriptions
Nodes with this color are subclases of DOLCE's endurants
Nodes with this color are subclases of DOLCE's perdurants
Nodes with this color are subclases of DOLCE's qualities
Nodes with this color are subclases of DOLCE's regions (and white is used for the quality spaces)
Arrow with this color: rdfs:subClassOf
Arrow with this color: allValesFrom
Arrow with this color: owl:someValesFrom
Arrow with this color: target classe is rdfs:subClassOf the rdfs:range of the arrow and similarly that the source class is rdfs:subClassOf the rdfs:domain of the arrow (not asserted: must be checked that it is consistent to state such an assertion)
Unforeseen uses
EXTENSION (UNEXPECTED USES)
PLAN
PLAN NOT FORESEEN BY DESIGN
PLAN FORESEEN BY DESIGN
GOAL NOT INTENDED BY DESIGN
is_a
has part
satisfies
executes
HUAWEI (Designer & Producer)
produces
depends_on
PLAN BY DESIGNER
has_part
INTENDED GOAL
METHODOLOGIES TO MEASURE DEPENDABILITY
isa
SAFETY
RELIABILITY
AVAILABILITY
MAINTAINABILITY
...
DEPENDABILITY VALUE (QUALE)
located_in
QUALITY SPACE
1
SAMPLE UNCERTAINTY
NORMATIVE UNCERTAINTY
UNCERTAINTY
CONSEQUENCES
SEVERITY SPACE
SEVERITY VALUE
has_value
SEVERITY
inheres_in
SITUATED RISK
has_part
RISK ASSESSEMENT
FAILURE (DESCRIPTION) device-centric view
ASSESSMENT METHOD
RCA
HFACS
FMECA
FMEA
FRACAS
MARKOV
FTA
CFT
RBD
HARA
is_associated_with
CAPABILITY
CAPACITY
INTRINSIC INDIVIDUAL QUALITY
founded_on
DEPENDABILITY (RELATIONAL QUALITY)
has value (quale)
PLAN FAILURE
DEVICE FAILURE
FAILURE
NON-PERFORMANCE EVENT
FAILURE (DESCRIPTION) environment-centric view (e.g. EM-DAT)
OBJECTIVELY UNSUCCESSFUL EXECUTION
OBJECTIVELY SUCCESSFUL EXECUTION
EXECUTION OF LEGITIMATE PLAN
USER
participates_in
uses
ENVIRONMENT
PHYSICAL DEVICE or SYSTEM
DEVICE/SYSTEM DESCRIPTION
affects
Normativity
RISK TOLLERANCE
RISK AVOIDANCE
RISK REDUCTION
RISK TRANSFER
RISK RETENTION
RISK TREATMENT
COMPANY POLICY
LAW
FMEA method
FAILURE EFFECT
FAILURE MODE
FAILURE CAUSE
FAILURE (DESCRIPTION)
causes

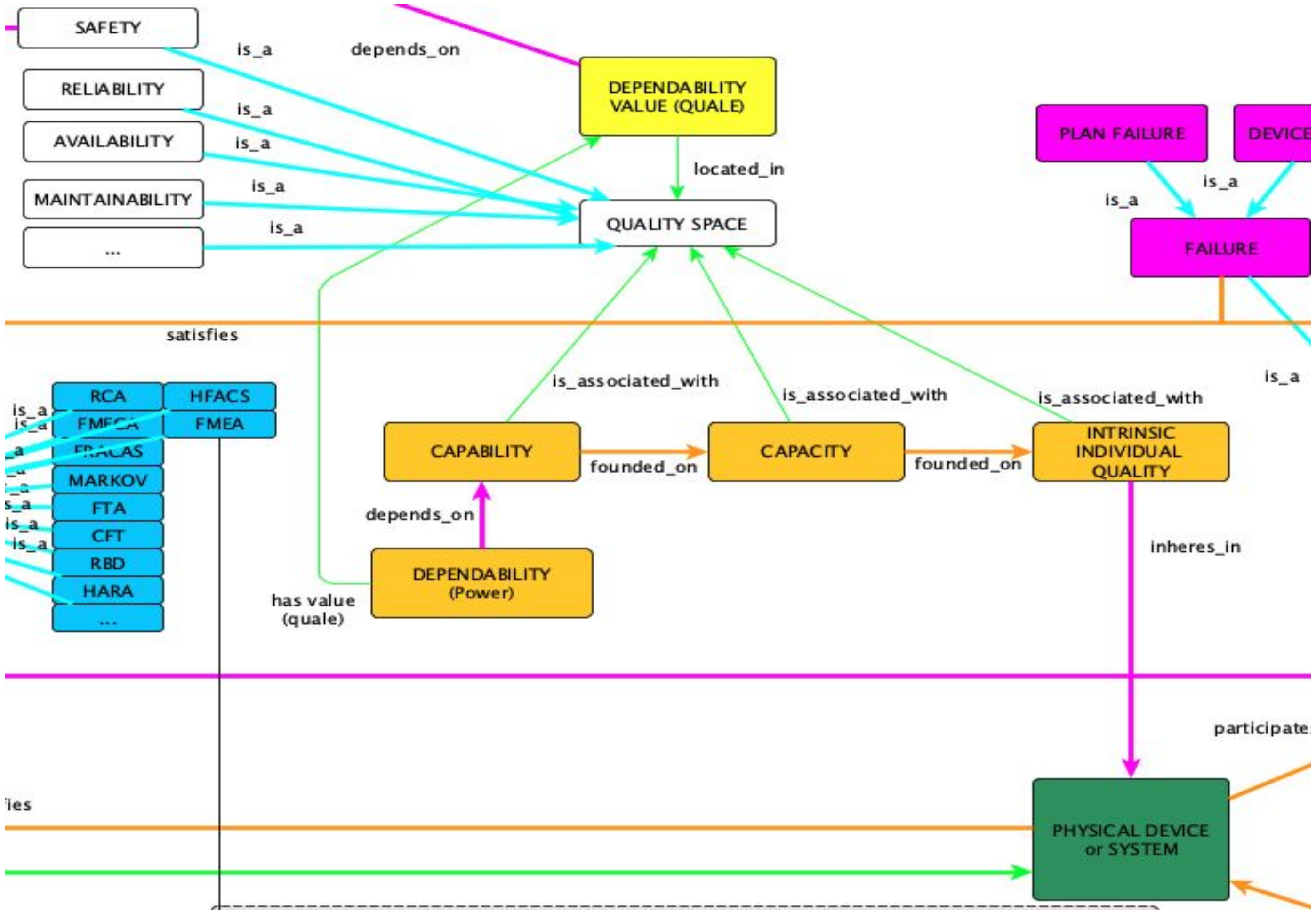
* classes refer to Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE)

# Dependability as a kind of Power

$$\mathbf{Power}(x) \Rightarrow \text{RelationalQuality}(x) \land \exists y(\text{capability}(y) \land \text{DifferentialQuality}(y, x))$$

- The axiom states that a **power** entails at least a **capability** as its **differential quality**
- Both capabilities and differential qualities are *relational qualities* (DOLCE)
- Relational (e.g. weight) vs Intrinsic (e.g. mass)
- Differential quality = “a quality that has a causal weight concerning another quality ”
- e.g. robotic arm: $\mathbf{Q_1}$ = ‘Power to lift’ ; $\mathbf{Q_2}$ = ‘Capability to exert force’
  - The robotic arm’s quality $\mathbf{Q_1}$ *is related* to the gripper’s quality $\mathbf{Q_2}$
  - It is **because of** the gripper’s quality $\mathbf{Q_2}$ that (among other things) the arm’s has the quality $\mathbf{Q_1}$, i.e. $\mathbf{Q_2}$ is a differential quality of $\mathbf{Q_1}$
- Typically, a power, and therefore dependability, would be identified in terms of **more than one** differential qualities

# Zoom on Dependability



# Dependability could be not enough for HRI

- However, even after individuating all the relevant attributes to represent dependability (in the engineering sense), we **are left with the problem** of managing the **interrelational aspect (Trust)**

- The interrelational aspect  is also partially captured by the relation of **Instrumental Dependence (ID)**

- However, this relation is **too genera**l to capture many details concerning trust, i.e. we need to go deeper in the way trust is expressed in terms of verbal and non-verbal signals

# Research proposal:
# Do actions speak louder than words?

- To address the interrelational aspect, under the instrumental dependence view, we propose an HRI study that considers both verbal and non-verbal singnals on trust:

  - **Understand the Interplay**: examine how verbal and non-verbal signals work together to influence trust in robots

  - **Determine Predominance**: investigate whether one type of feedback has a more significant impact on trust levels

- This is fundamental to allow robots to mitigate both overtrust and mistrust by the users' side:

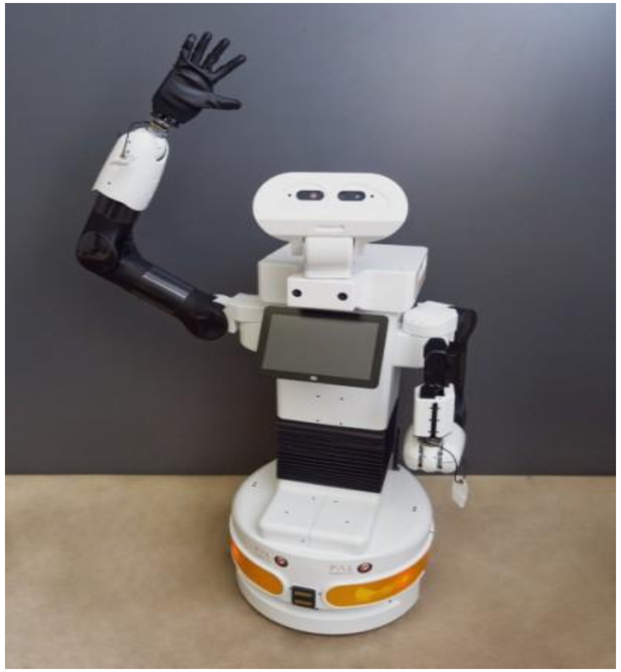
  - **Mitigating Overtust:**

  e.g. Excessive physical contact ⇒ remind the user robots do not feel empathy

  - **Mitigating Mistrust:**

  e.g. neglecting communication/interaction ⇒ remind the user of the robot's capabilities for specific tasks

# Machines to use (2 possibilities)



(Tiago PAL Robotics)

(robotic arm + screen to mimic social responses)

# Motivating the study

- In HRI, trust is at stake in collaboratory tasks in which robots work as teammates (not merely tools):

  a) E.g. a robotic arm and a human worker collaborate to assemble parts on a production line

  b) E.g. a social service robot assists elderly patients with daily tasks such as fetching items, providing reminders, and offering companionship.

- **Non-verbal and verbal feedback** have to be evaluated, regardless of the complexity/ anthropomorphic level of the robot, since **collaboratory tasks** inherently require participants to engage both in verbal (e.g. **imperative speech**: e.g do this, grab that etc.) and non-verbal communication, such as **pointing**, **handing over** objects, etc.

- Therefore, determining the impact of such signals is a pivotal topic for all studies in HRI, and **should be considered when** designing and implementing robots

- We set a list of hypothesis to study such signals

# Hypotheses to test

Hypothesis 1): in general, given that non-verbal behaviours are likely to occur largely outside of people's conscious control, one might suppose a **greater incidence of non-verbal signals**, over verbal signals

Hypothesis 2): we expect that verbal and non-verbal signals are **cumulative**, i.e. the combination of verbal and non-verbal feedback results in higher trust levels than either type of signal alone

Hypothesis 3): **contradictory behaviours** (e.g. saying a thing and doing something else) by the robot will significantly **decrease trust levels**. Participants will find it harder to trust a robot that provides inconsistent feedback

Hypothesis 4): besides lowering the level of trust, **contradictory behaviours rise mistrust** (at various levels and related values, i.e. intensity)

- Hypothesis 5): We expect **technical failures*** to be **less severe** than contradictory behaviours in terms of undermining trust, i.e. their incidence on trust level is lower. The reason is that contradictory behaviours can be perceived by humans at two levels:

  - as **irrational** (e.g. the behaviour of the agent is too unpredictable or too inconsistent, e.g. madness)

  - as a **violation** of the (more or less explicit) agreement, which holds because of the goal-achievement, from the trustee side

  - Kinds of contradictory behaviors
    - **Verbal-Verbal**

      Inconsistent verbal messages, e.g. "I am here to help you," and a minute later says "I can't assist you right now.")

    - **Verbal-Behavioural**

      the robot says "you are doing great!" but the actions performed, e.g. shaking its head/moving away, suggest disapproval /disengagement

    - **Behavioural-Verbal**

      e.g. the robot nods/gives a thumbs up and then says "you need to redo this"

    - **Behavioural-Behavioural**

      e.g. the robot reaches out to pass an object to the person but then retracts its arm

* technical failure = e.g. an error in face-recognition caused by poor illumination (Honig et. al, 2018)

# Manipulating Independent Variables:

- Independent Variables*:
  - **Verbal signals only**: the robot interacts solely through verbal communication, either positive or negative
  - **Non-Verbal signals only**: the robot uses solely non-verbal behaviour to interact, either positive or negative
  - **Both Verbal and Non-Verbal signals**: the robot uses a combination of verbal and non-verbal signals, either positive or negative
  -

*
a) type of signal (verbal/non-verbal)
b) type of value associated with the signal, i.e. '+' or '-' depending on whether it elicit or not trust

- Dependent variables
  - Trust-score (Mistrust, Overtrust)

# Environment Alternative to Game-Theory (Fruit Sorting)

- The main goal is to **sort different types of fruits** based on e.g. ripeness, size, colour, etc.
- The experiment tests **how verbal and non-verbal** signals from robots **influence human's trust/mistrust** simulating collaboration in a real-world industrial setting
- This approach serves as an alternative to traditional game theory models, focusing instead on **real-time interaction** and in a practical **goal-oriented environment**

## Verbal Signal

### Positive signal:

- **Example:** Tiago suggests “Let’s get rid of all irregular size apples”

- **Rationale:** providing hints coherently with the goal, sensible ideas, positive or supportive feedback (e.g. “we are doing great!”), etc.

### Negative signal:

- **Example:** Tiago says “We should arrange the fruits by color cause it help with freshness”

- **Rationale:** might include nonsense suggestions (e.g., “Let’s just stack all the fruits together”), contradictions, discouraging remarks (“This takes too long, we’ll never finish on time”), etc.

## Non-Verbal Signal

### Positive signal:

- **Example:** Tiago points to the ripe fruit bin, hands over a fruit that needs sorting, while maintains eye contact with the worker

- **Rationale:** supportive and intuitive actions that align with bringing about the goal while signalling shared attention/respect for the user

### Negative signal:

- **Example:** Tiago shakes its head, moves away from the sorting area/drops a fruit/ or knocks over a basket

- **Rationale:** shows clumsiness, disapproval or disengagement, contradictory behaviours

# Traditional Methodologies for estimating Trust

- **Questionnaires:**
  - **Pre-Experiment Questionnaire:**

    i. e.g. test to mitigate the effect of participants' personality  on rating the interaction with robots (Gosling et al. 2003)

    ii. e.g. test to detect participants' feelings/bias towards robots (Nomura, 2003)

  - **Post-Experiment Questionnaire:**

    **i.** e.g. to study participant's experience on the interaction with robots

- **Task Performance Evaluation:**
  - **Success Rate of the main goal:**  succeed/fail of the task

  - **Efficiency Metrics:** e.g. the average response time for participants with respect to robot's feedback

  - **Proxemic metric: e.g.** check the respect/violation of spaces related to interactions, including intimate space, personal space, social space distance (Cristani et. al, 2020)

# Operationalizing Instrumental Dependence

- The ultimate goal is to **embed knowledge** about trust signals, enabling robots to **estimate** users' levels of trust/mistrust/overtrust

- In **Open-Ended Learning** robots are supposed to continuously and autonomously explore the environment to **incrementally** acquire knowledge and new skills in view of some **purposes** that match human needs (Baldassarre et al., 2024)

- In this respect, a **Meta-Cognitive Neural Network** architectures* that, given a goal-assignment, allow the robot to compute the probability of success of such goal is evidently useful for understanding **to what extent** users **can depend on** robots.

- This architecture could be indeed employed to **operationalize** the relation of **Instrumental Dependence (ID)**, e.g. the (conditional) probability that the user will depend on the robot (given) the probability of the goal achievement.

- If we **also account for trust-signals** into the architecture, we can obtain the (conditional) probability that, e.g. a user shall emit trust-signals given the probability they are ID on the robot.

*Ongoing project at Laboratory of Embodied Natural and Artificial Intelligence (LENAI, ISTC-CNR)

# When do we need Trust?

- So we should, in principle, be able to probabilistically distinguish when a user is **merely exhibiting Instrumental Dependence** on a robot from a situation in which **they are also placing trust** in the robot, as indicated by the presence of a set of verbal and non-verbal signals

- Typically, to be instrumentally dependent on machines and hopefully achieving some goals, **it is sufficient** that machines are **dependable**

- However, if the **goal-achievement** (e.g. sorting fruit in HRI scenario) depends, among other things, by **the very 'relationship' with the machines**, as in the case of social robots, then **trust** might be needed

- In this sense, trust is required in very **few** and (more or less) **futuristic** cases of HRI

- In most cases, Instrumental dependence, understood in dependability terms, will do

- However, given the technological pervasiveness and the delicacy of the contexts in which trust is required in HRI, we nonetheless need an adequate evaluation of trust

# Main Challenges to Face

- Time for Training: Robots need to be trained for complex tasks, which can be **time-consuming** in real-world scenarios

- Limited Physical Capabilities: Robots may struggle with executing **smooth and precise movements** due to physical limitations (considering the capabilities of nowadays robots, both in the academy and industry), e.g. handling irregular items, manipulating fruit without damaging it

- Trust-Signal Evaluation: Evaluating the **correlation** between signals (verbal and non-verbal) and people's trust/mistrust/overtrust is challenging (no prior studies)

# IV) The digital twin approach

# The Digital Twin (DT) Status

- Lack of a standard definition (problem of intertheoretical coherency, as for trust)
- The very expression implies a physical counterpart/copy
- Ontologically, it is (at least) an *approximate copy* of a physical entity (Angius, N. & Primiero, G., 2018)
- Does not rely on a single technology, i.e. it is a complex artefact (e.g. ML, IoT, Computational Ontologies, sensors, servers, etc. ) (Korenhof et al. 2021)
- Its complexity is dictated by the scope/applications and scale of the related physical entity (ranging from cells to a human body, from an engine to a whole factory)
- In a more traditional sense, one can see DTs, together with their physical relata, as kinds of cyber-physical systems (Wagg. et al. 2024)

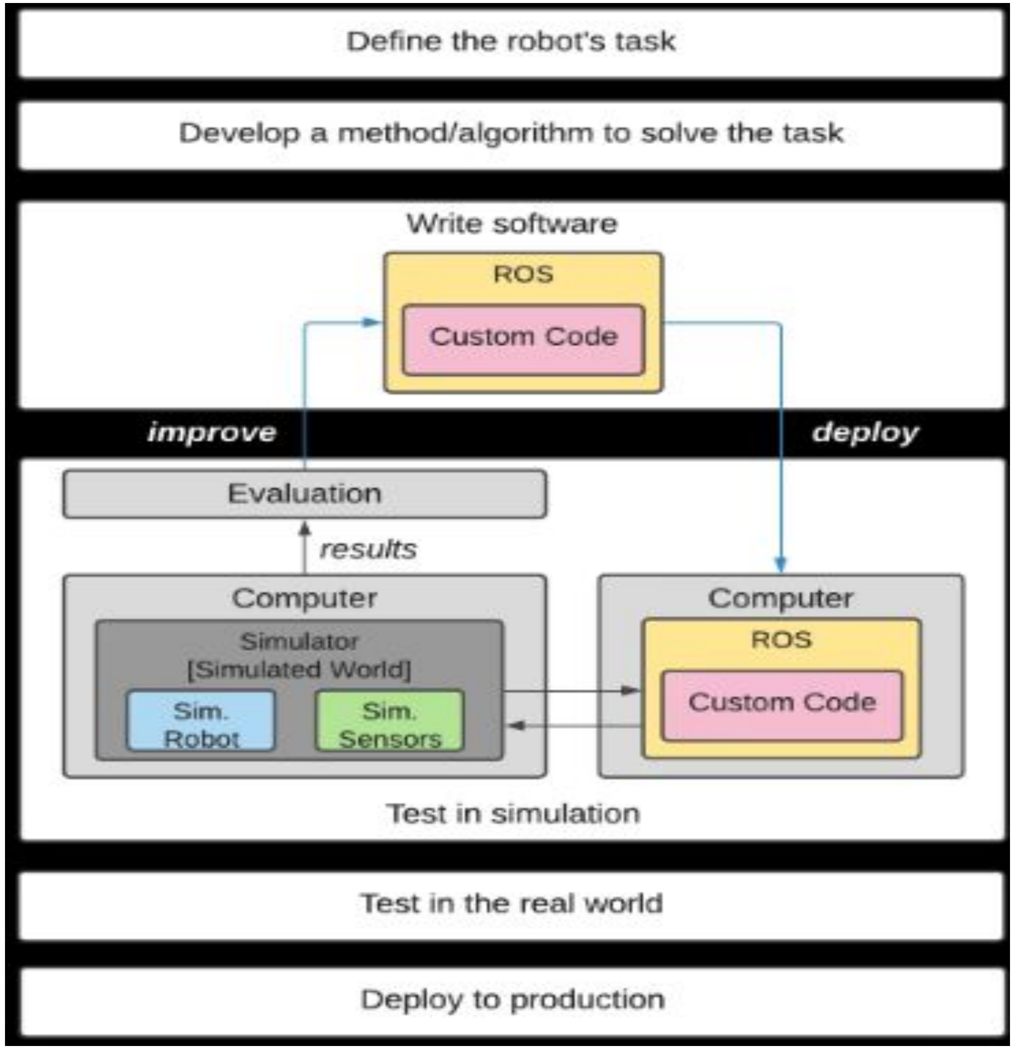
# DTs in HRI

- DTs can be employed in HRI to **overcome some physical and ethical limitations** we mentioned, such as mechanical wear and safety

- **3D engines and physics simulation environments** are pivotal in this sense

- Deep Learning algorithms serve to **speed up the process of learning** in the virtual environment and then **map the learnt behaviour** in the physical robot

# Simulating before testing



https://unity.com/blog/engine-platform/robotics-simulation-is-easy-as-1-2-3

**Advantages:**

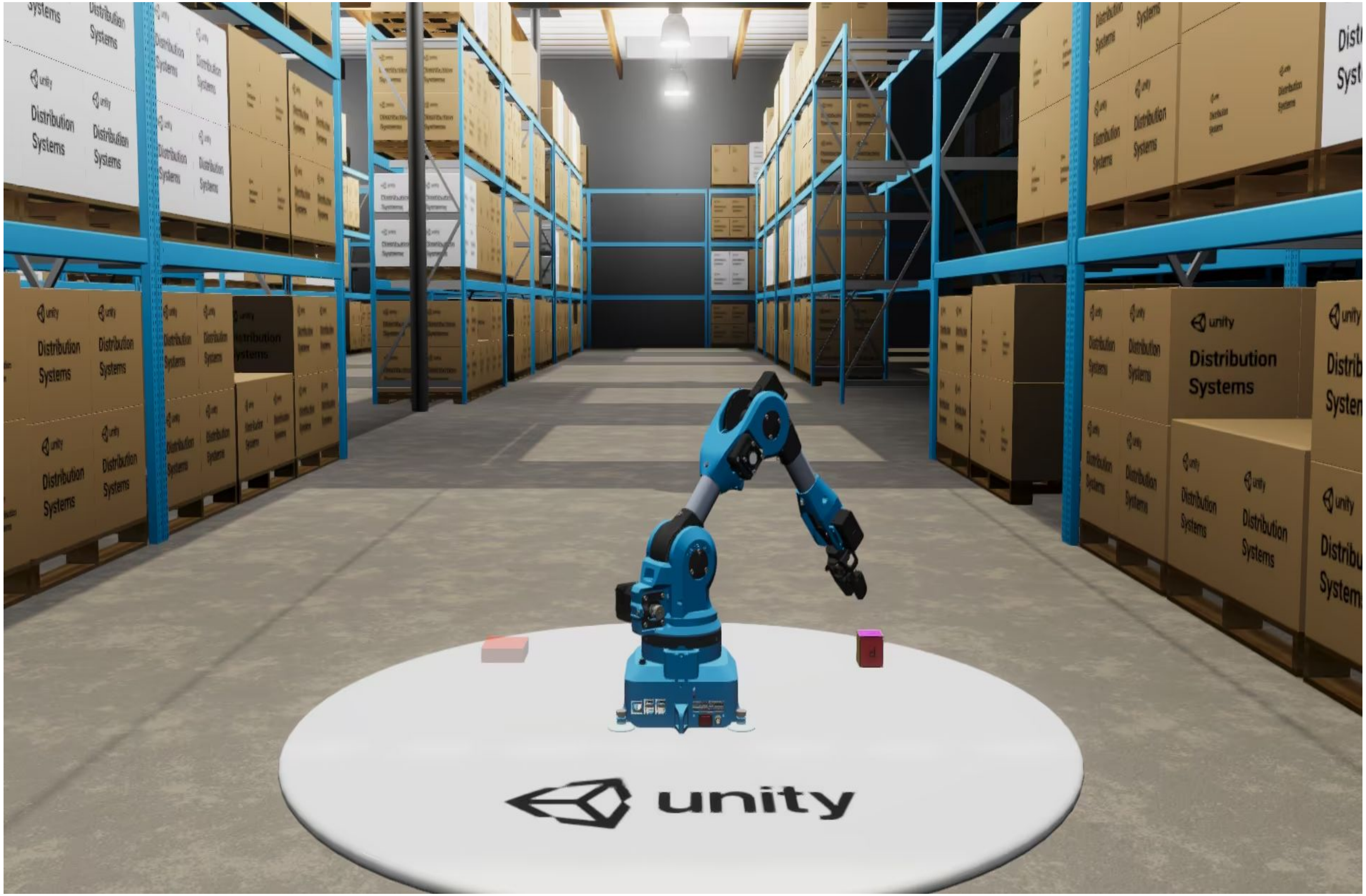
- **Cost Efficiency:** reduces the need for physical prototypes and real-world trials, saving both time and money.

- **Safety:** allows for testing and refining robot behaviors and interactions in a risk-free virtual environment.

- **Flexibility:** provides the ability to easily modify and test different scenarios, tasks, and configurations (tests different kinds of robots at different complexity)

- **Real-Time Feedback:** offers real-time data and insights into the robot's performance, enabling rapid adjustments and improvements

# The Unity Environment



# Training and Testing for the Fruit Sorting scenario

- Train the Virtual robot for **performance-related tasks**:
  - e.g. computing the force/joint dynamics to manipulate fruits

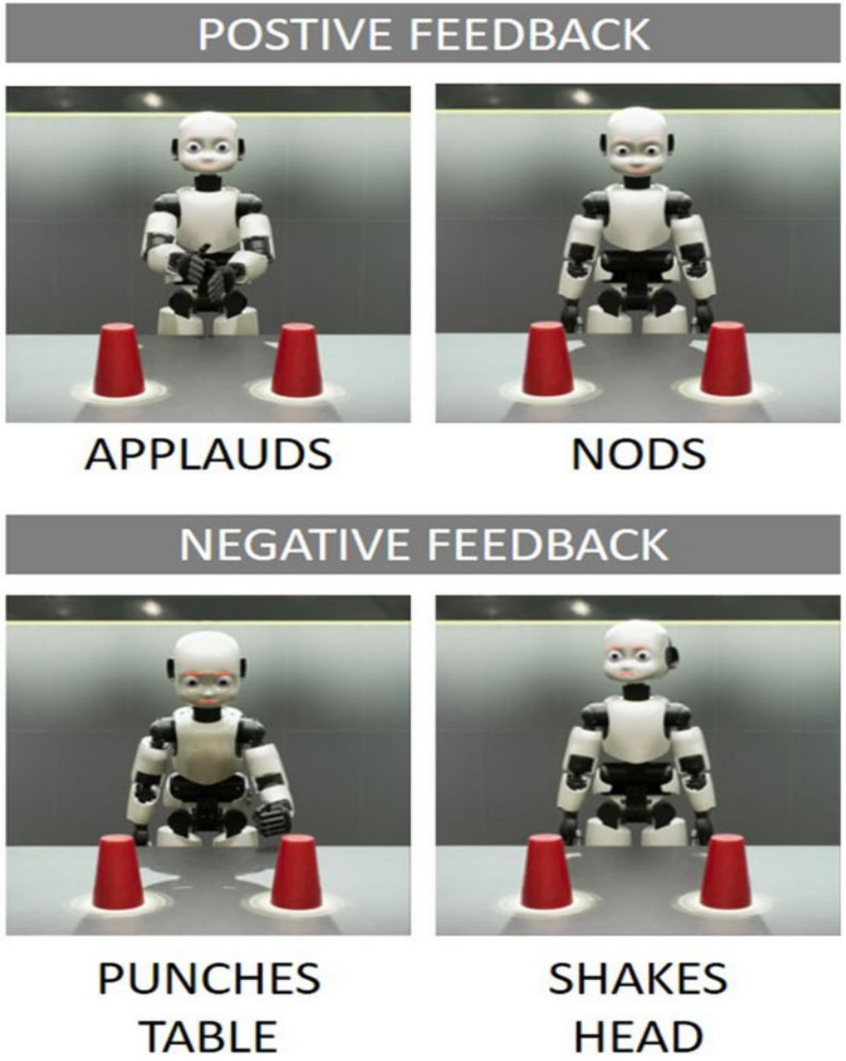
- Train the Virtual robot to **mimic-different non-verbal trust signals:**
  - nodding or turning ‘slightly’ towards the human when receiving a fruit to mimic attentiveness/care

  - pointing towards a human-worker or ‘gently’ extending a hand towards an item it’s unsure about to ask for help

- Train the robot to **detect risky events**:
  - e.g. detecting dropped objects on the floor and either catch them or alert human workers

- Directly using the virtual environment to **show people trust-signals and collect their feedback** in terms of perceived trust/mistrust/overtrust

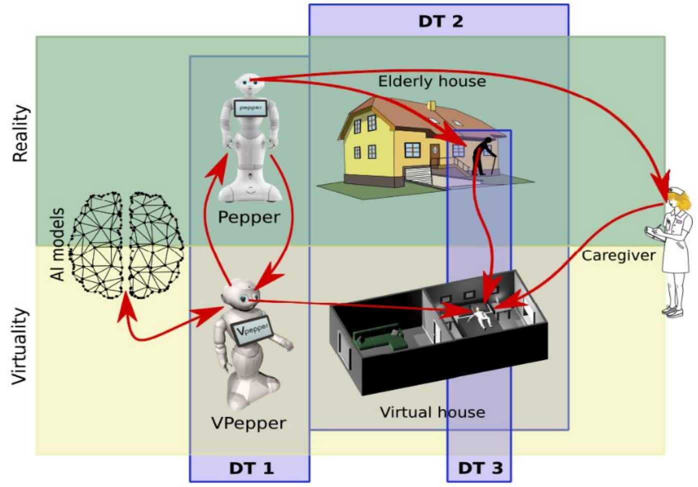
# Some Interesting Examples from the Literature



(Parenti et al., 2023)

Independent variables:1) **movements** 2) **social feedback** ⇒ Influence a) **agents' performance** (time to answer); b) **Trust rating** (questionnaires)

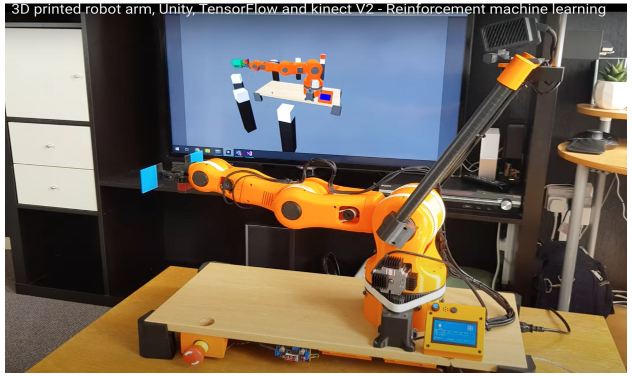
Fig. 1



- VPepper (DT) is trained to safely/kindly touch objects/persons for home assistance purposes
(Cascone et. al 2021)

- A robotic arm is trained to pick-place objects; the DT drives the physical arm in real-time
(Matulius et al 2021)

Fig.2


3D printed robot arm, Unity, TensorFlow and kinect V2 - Reinforcement machine learning

https://www.youtube.com/watch?v=0r71sXiK7wA&list=PLoaeMzYH5jvqwB-oP2qrQfDvecntQlJPt

NB: both studies use Proximal Policy Optimization algorithm and Unity environment

# Conclusions

- Within HRI scholars address both **ontological** and **empirical** issues on trust

- The relation of **Instrumental Dependence** encompasses many kinds of goal-oriented contexts, including **dependability** and **trust**

- While dependability, understood as a **Power**, can be used to represent the **artefactual** aspect of HRI, we still need to account for the **interrelational** aspect, i.e. trust

- Embracing a **relational view** on trust, we draw on the Social Signaling theory to identify the **signals** that constitute the lexicon of trust

- We propose the fruit sorting study as an experimental setting for evaluating trust in HRI as **alternative to** game-theory contexts, while pointing out the main practical challenges involve

- We outline how such challenges could be (partially) addressed by resorting to the Digital Twin approach

# Bibliography

Angius, N., & Primiero, G. (2018). The logic of identity and copy for computational artefacts. *Journal of Logic and Computation*, *28*(6), 1293-1322

Baldassarre, G., Duro, R. J., Cartoni, E., Khamassi, M., Romero, A., & Santucci, V. G. (2024). Purpose for Open-Ended Learning Robots: A Computational Taxonomy, Definition, and Operationalisation. *arXiv preprint arXiv:2403.02514*.

Bartneck, C., Croft, E., & Kulic, D. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1), 71-81.

Bartneck, C., Soucy, M., Fleuret, K., & Sandoval, E. B. (2015). The Robot Engine - Making The Unity 3D Game Engine Work For HRI. Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN2015), Kobe pp. 431 - 437. | DOI: 10.1109/ROMAN.2015.7333561

Bayliss, A. P., & Tipper, S. P. (2006). Predictive Gaze Cues and Personality Judgments: Should Eye Trust You? *Psychological Science*, *17*(6), 514-520.

Biccheri, L., Ferrario, R., & Porello, D. (2020). Needs and intentionality. In *Formal Ontology in Information Systems* (pp. 125-139). IOS Press.

Biccheri L., Borgo S., & Ferrario R. (2023), "On the Relation of Instrumental Dependence". In: Formal Ontology in Information Systems, Sherbrooke, Qc, Canada, July 17-20, 2023.

Booth, S., Tompkin, J., Pfister, H., Waldo, J., Gajos, K., & Nagpal, R. (2017, March). Piggybacking robots: Human-robot overtrust in university dormitory security. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 426-434)

Carvalho, E. (2020). Social affordance. In Vonk, J., & Shackelford, T. (Eds.), *Encyclopedia of Animal Cognition and Behavior*, pp. 1–4. Springer.

Cascone, L., Nappi, M., Narducci, F., & Passero, I. (2021). DTPAAL: Digital twinning pepper and ambient assisted living. *IEEE Transactions on Industrial Informatics*, *18*(2), 1397-1404.

Castelfranchi, C., Falcone, R., & Lorini, E. (2009). A non-reductionist approach to trust. In *Computing with Social Trust* (pp. 45-72). London: Springer London.

Castelfranchi, C., & Falcone, R. (2010). *Trust theory: A socio-cognitive and computational model*. John Wiley & Sons

Corritore, C. L., Kracher, B., & Wiedenbeck, S. (2003). On-line trust: concepts, evolving themes, a model. *International journal of human-computer studies*, *58*(6), 737-758.

Ferretti, G., & Zipoli Caiani, S. (2023). How knowing-that and knowing-how interface in action: The intelligence of motor representations. *Erkenntnis*. pp.1103-1133

Flook, R., Shrinah, A., Wijnen, L., Eder, K., Melhuish, C., & Lemaignan, S. (2019). On the impact of different types of errors on trust in human-robot interaction: Are laboratory-based HRI experiments trustworthy?. *Interaction Studies*, *20*(3), 455-486.

Gibson, J. J. (2015). The ecological approach to visual perception, classical edition. New York: Psychology Press.

Guinnane, T. W. (2005). Trust: a concept too many. *Jahrbuch für Wirtschaftsgeschichte/Economic History Yearbook*, *46*(1), 77-92.

Hollis, M. (1998). *Trust within reason*. Cambridge University Press.

Hollis, M. (1998). *Trust within reason*. Cambridge University Press.

Hume, D. (1978). A Treatise of Human Nature. Oxford: Oxford University Press.

Honig, S., & Oron-Gilad, T. (2018). Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in psychology*, *9*, 861.

Korenhof, P., Blok, V., & Kloppenburg, S. (2021). Steering representations—towards a critical understanding of digital twins. *Philosophy & technology*, *34*, 1751-1773.

Kok, B. C., & Soh, H. (2020). Trust in robots: Challenges and opportunities. *Current Robotics Reports*, *1*(4), 297-309.

Matulis, M., & Harvey, C. (2021). A robot arm digital twin utilising reinforcement learning. *Computers & Graphics*, *95*, 106-114.

Milnes, Tim, 'The Subject of Trust', *The Testimony of Sense: Empiricism and the Essay from Hume to Hazlitt* (Oxford, 2019; online edn, Oxford Academic, 22 Aug. 2019)

Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, *39*(3), 429-460.

Mutlu, B., Yamaoka, F., Kanda, T., Ishiguro, H., & Hagita, N. (2009, March). Nonverbal leakage in robots: communication of intentions through seemingly unintentional behavior. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction* (pp. 69-76)

Nomura, T., & Kanda, T. (2003, Nov). On proposing the concept of robot anxiety and considering measurement of it. In The 12th ieee international workshop on robot and human interactive communica- tion, 2003. proceedings. roman 2003. (p. 373-378)

Parenti, L., Lukomski, A. W., De Tommaso, D., Belkaid, M., & Wykowska, A. (2023). Human-likeness of feedback gestures affects decision processes and subjective trust. *I nternational Journal of Social Robotics*, *15*(8), 1419-1427.

Poggi, I. (2013). Mind, hands, face, and body: A sketch of a goal and belief view of multimodal communication. *Body–Language–Communication: An International Handbook on Multimodality. in Human Interaction*, *1*, 627-647.

Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016, March). Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 101-108). IEEE.

Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., & Joublin, F. (2013). To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics*, *5*, 313-323.

Toumpa and A. G. Cohn. “Object-agnostic Affordance Categorization via Unsupervised Learning of Graph Embeddings”. In: Journal of Artificial Intelligence Research 77, (2023), pp. 1–38.

Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and vision computing*, *27*(12), 1743-1759.