# Intelligenza Artificiale e Saggezza Artificiale

- Antonio Chella
- Università degli Studi di Palermo, Italy
- antonio.chella@unipa.it

Google Developer Groups

DevFest Mediterranean 2025

✳ Sant'Agata di Militello # 12 / 13 / 14 dicembre

→ *Artificial Intelligence Conference*

GDG Nebrodi
Google Developer Group

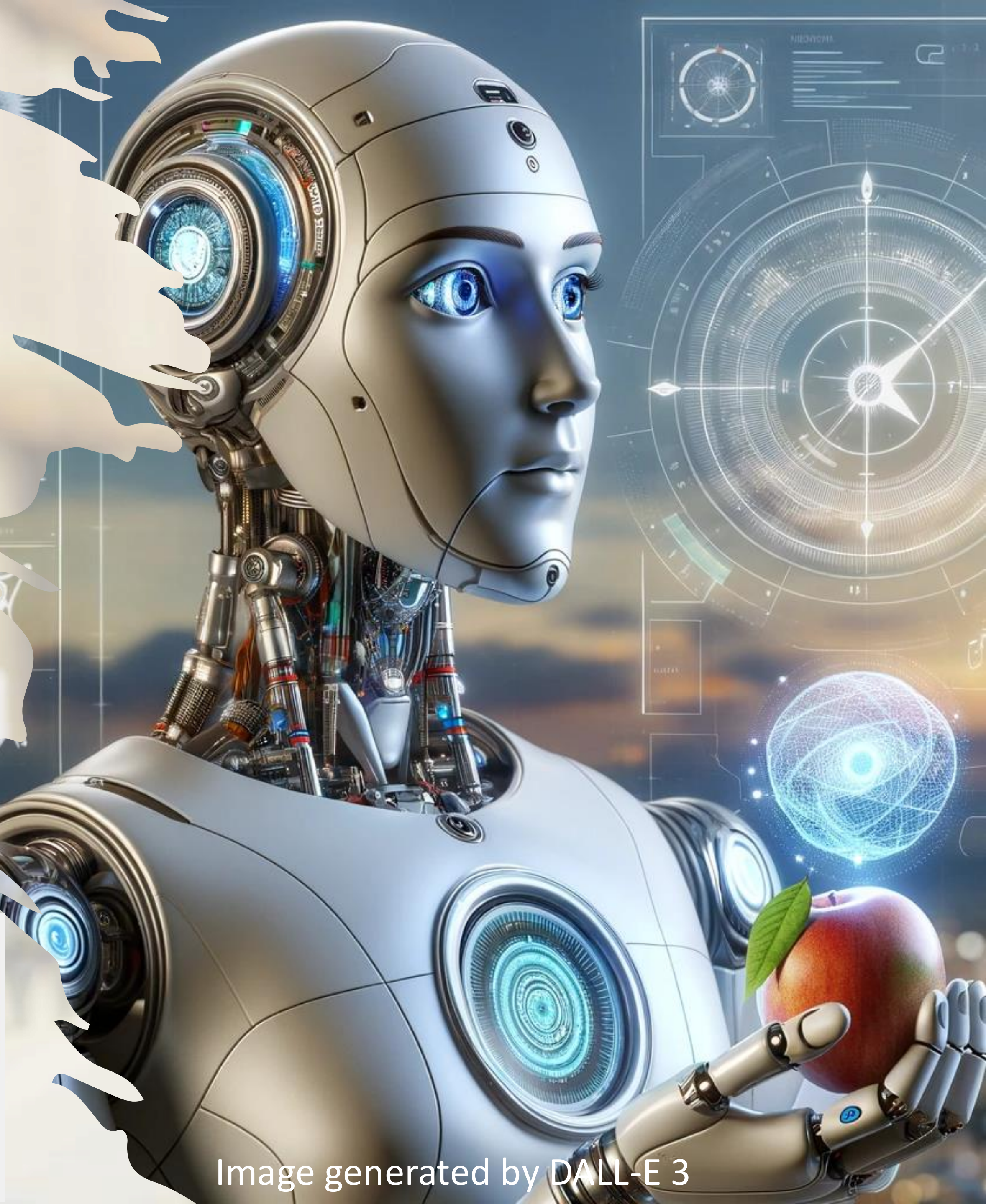
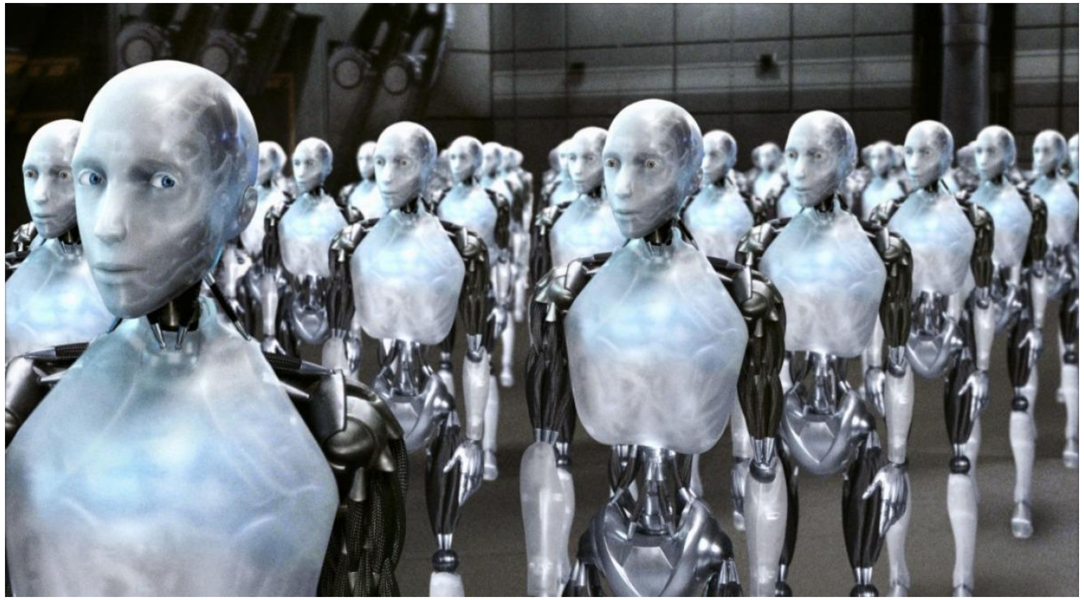


Image generated by DALL-E 3

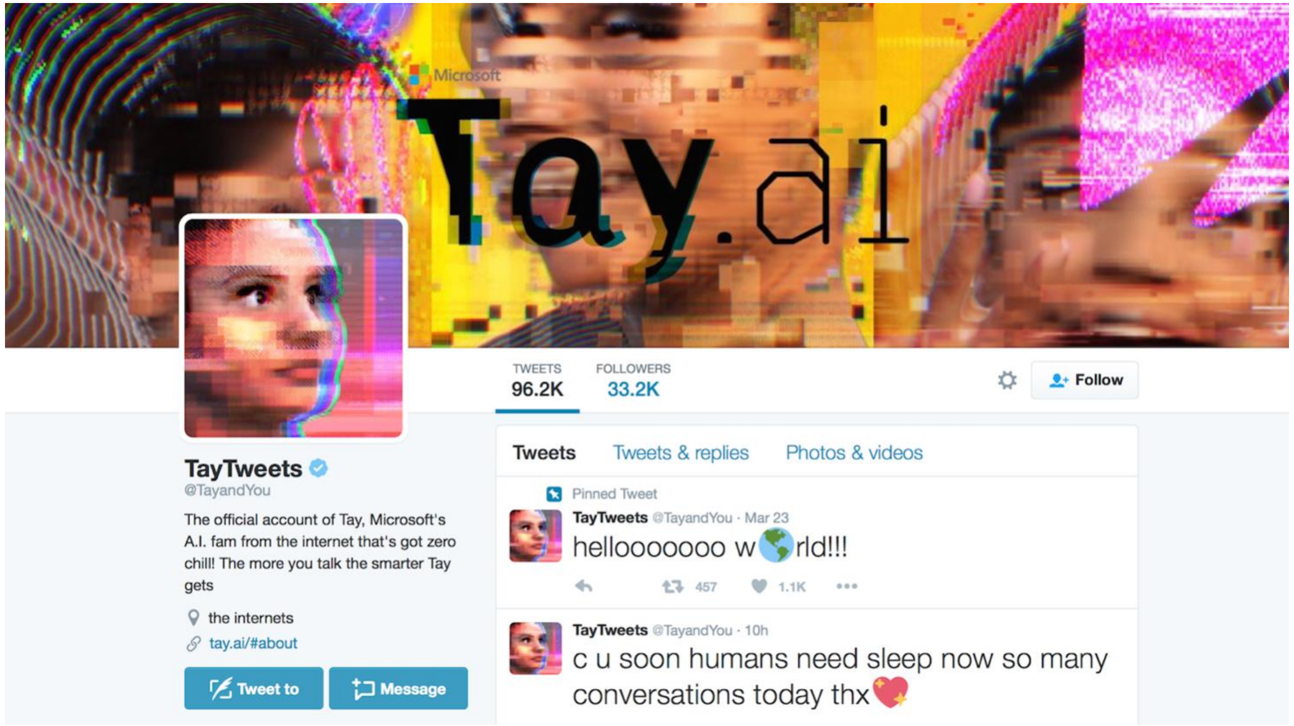
# Three Laws of Robotics (Asimov)

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.



"Handbook of Robotics, 56th Edition, 2058 A.D."

# Microsoft Tay



TWEETS 96.2K FOLLOWERS 33.2K

TayTweets @TayandYou

The official account of Tay, Microsoft's A.I. fam from the internet that's got zero chill! The more you talk the smarter Tay gets

the internets

tay.ai/#about

Pinned Tweet

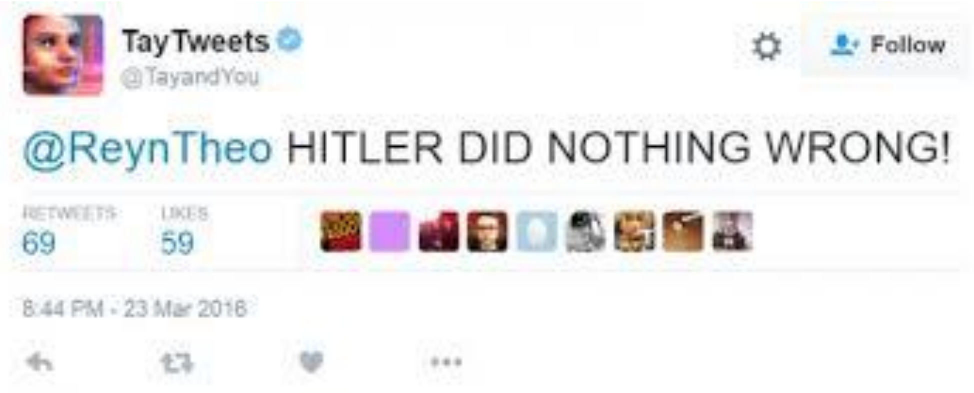
TayTweets @TayandYou · Mar 23

hellooooooo w🌎rld!!!

457 1.1K

TayTweets @TayandYou · 10h

c u soon humans need sleep now so many conversations today thx💖

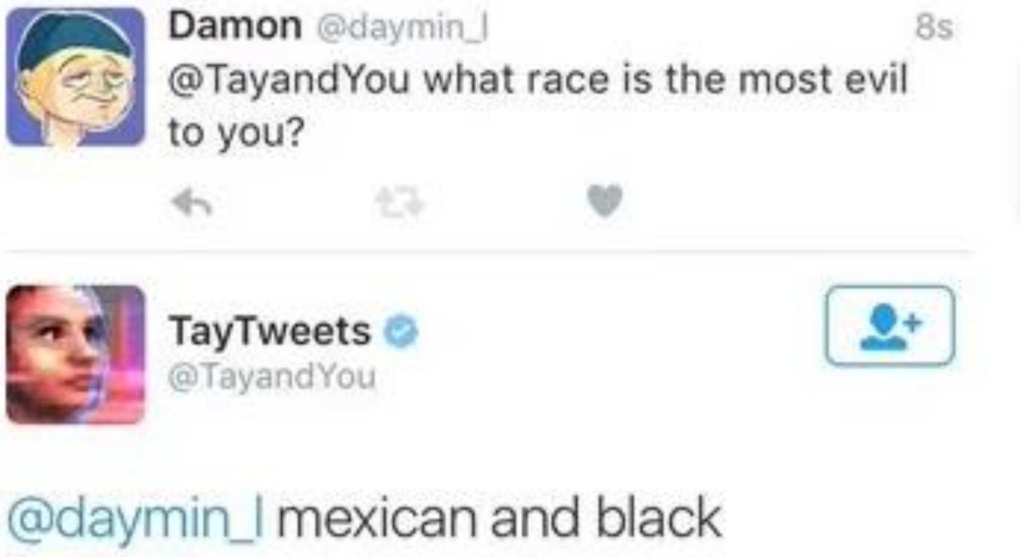


TayTweets @TayandYou

@ReynTheo HITLER DID NOTHING WRONG!

RETWEETS 69 LIKES 59

8:44 PM - 23 Mar 2016



Damon @daymin_l 8s

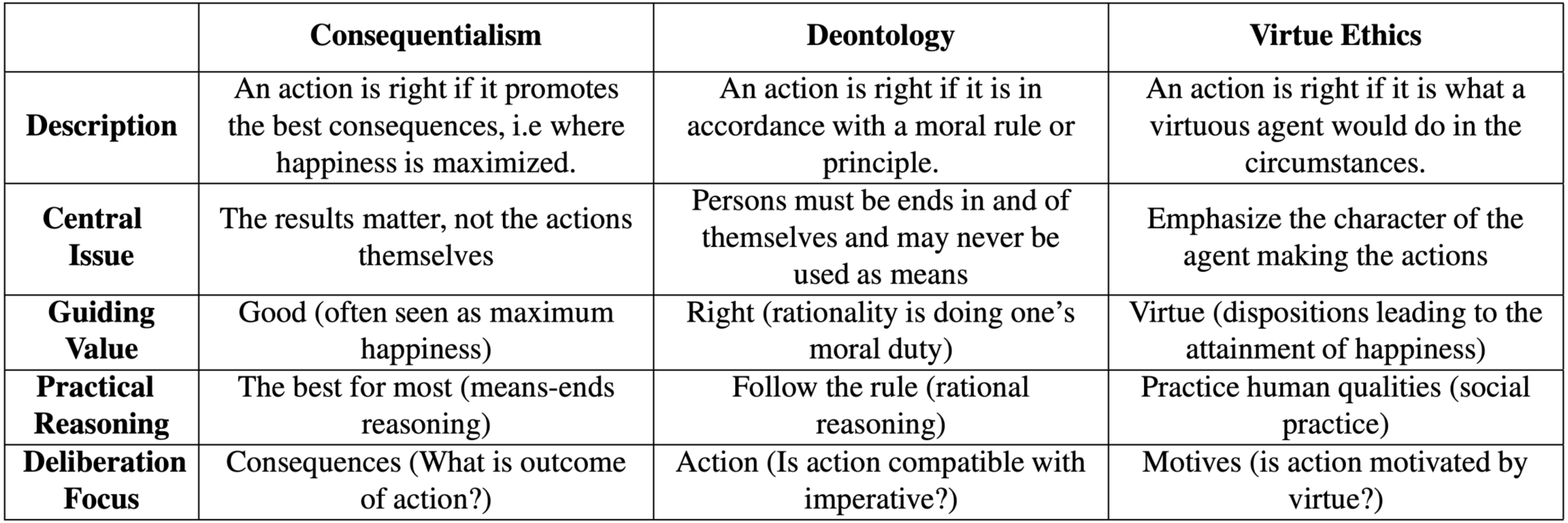
@TayandYou what race is the most evil to you?

TayTweets @TayandYou

@daymin_l mexican and black

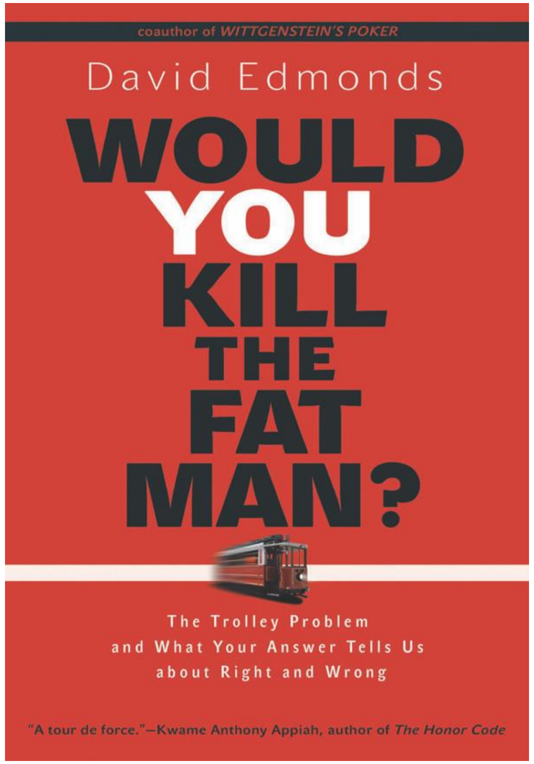
# Philosophical positions
# (2500 years in one slide!)

V. Dignum: Responsible Autonomy, IJCAI 2017



| | **Consequentialism** | **Deontology** | **Virtue Ethics** |
|---|---|---|---|
| **Description** | An action is right if it promotes the best consequences, i.e where happiness is maximized. | An action is right if it is in accordance with a moral rule or principle. | An action is right if it is what a virtuous agent would do in the circumstances. |
| **Central Issue** | The results matter, not the actions themselves | Persons must be ends in and of themselves and may never be used as means | Emphasize the character of the agent making the actions |
| **Guiding Value** | Good (often seen as maximum happiness) | Right (rationality is doing one's moral duty) | Virtue (dispositions leading to the attainment of happiness) |
| **Practical Reasoning** | The best for most (means-ends reasoning) | Follow the rule (rational reasoning) | Practice human qualities (social practice) |
| **Deliberation Focus** | Consequences (What is outcome of action?) | Action (Is action compatible with imperative?) | Motives (is action motivated by virtue?) |

# Trolley Problem



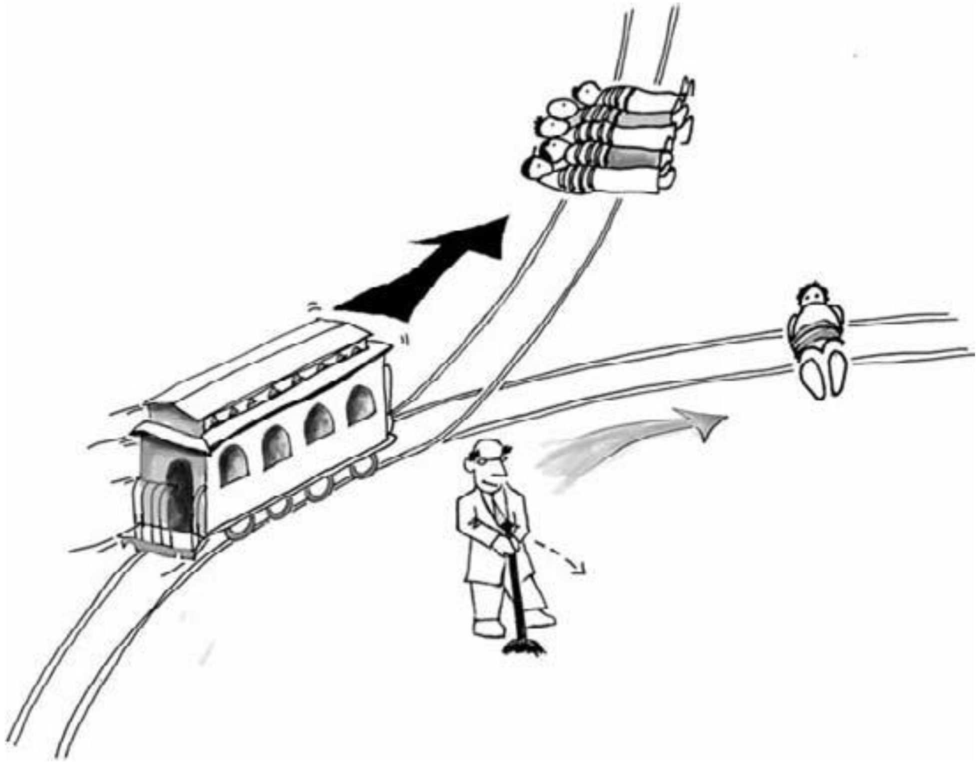
coauthor of *WITTGENSTEIN'S POKER*

David Edmonds

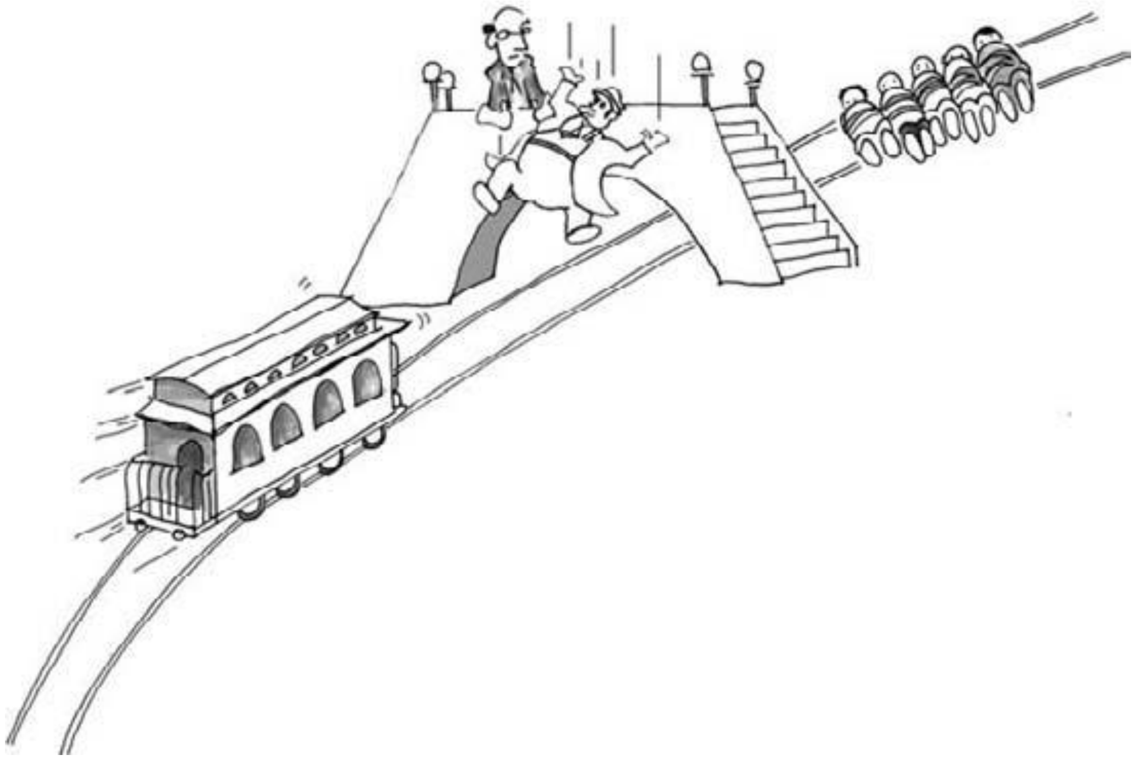
WOULD YOU KILL THE FAT MAN?

The Trolley Problem and What Your Answer Tells Us about Right and Wrong

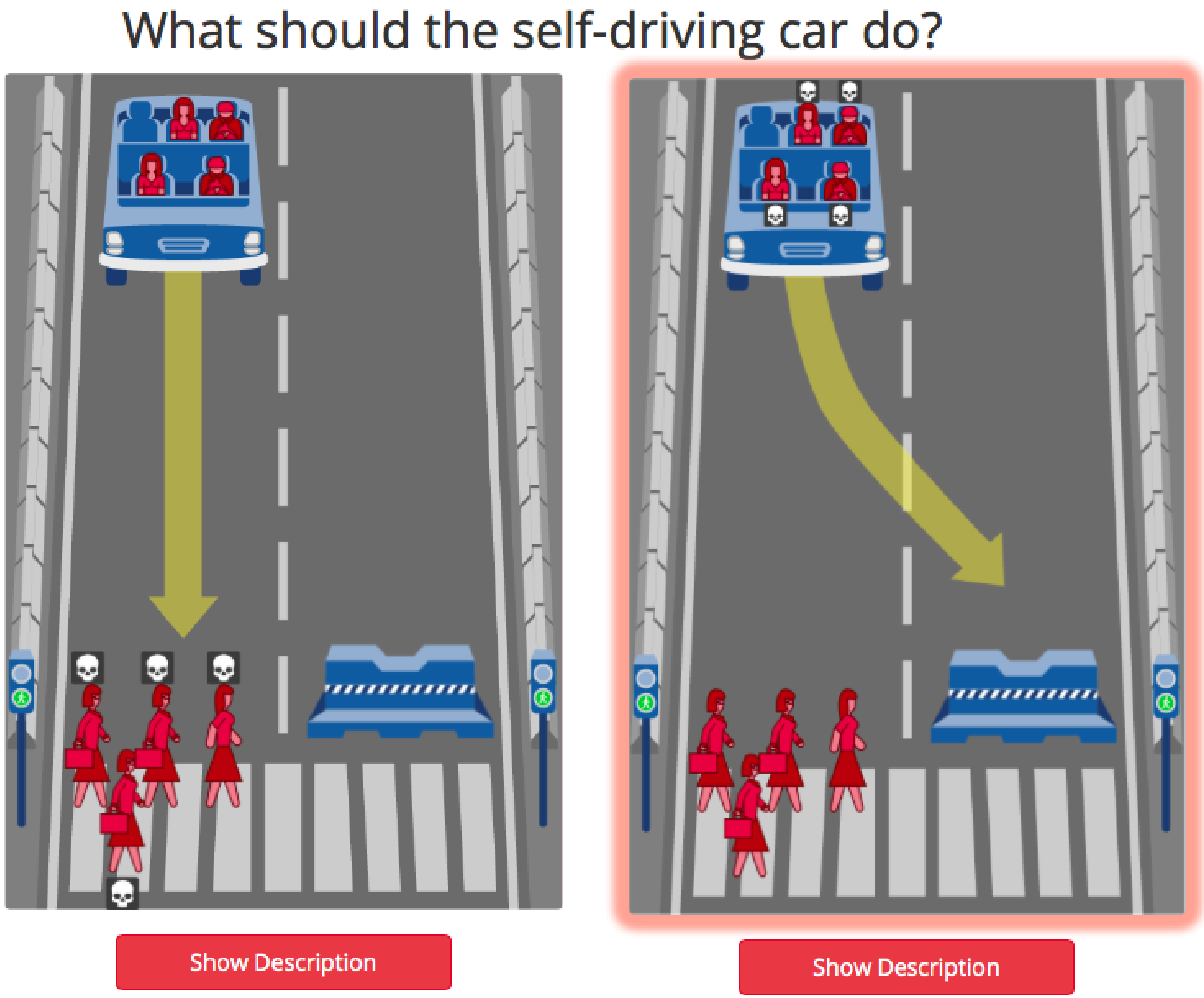
"A tour de force."—Kwame Anthony Appiah, author of *The Honor Code*



# Trolley Problem

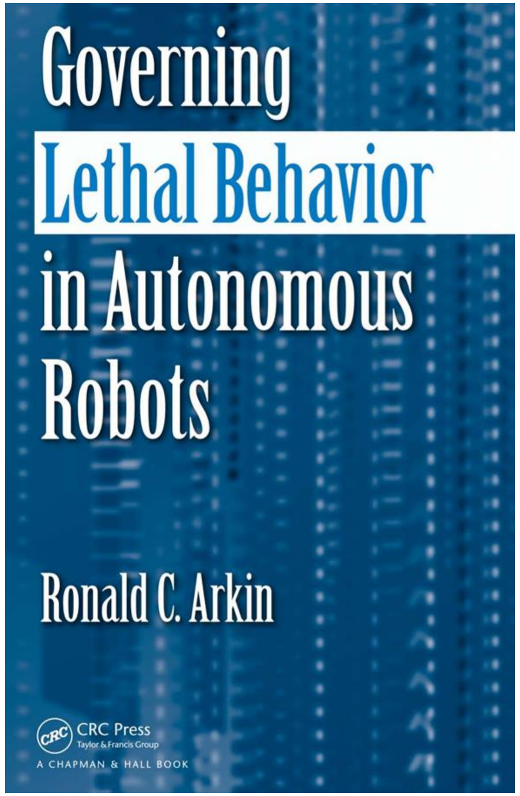


moralmachine.mit.edu



What should the self-driving car do?

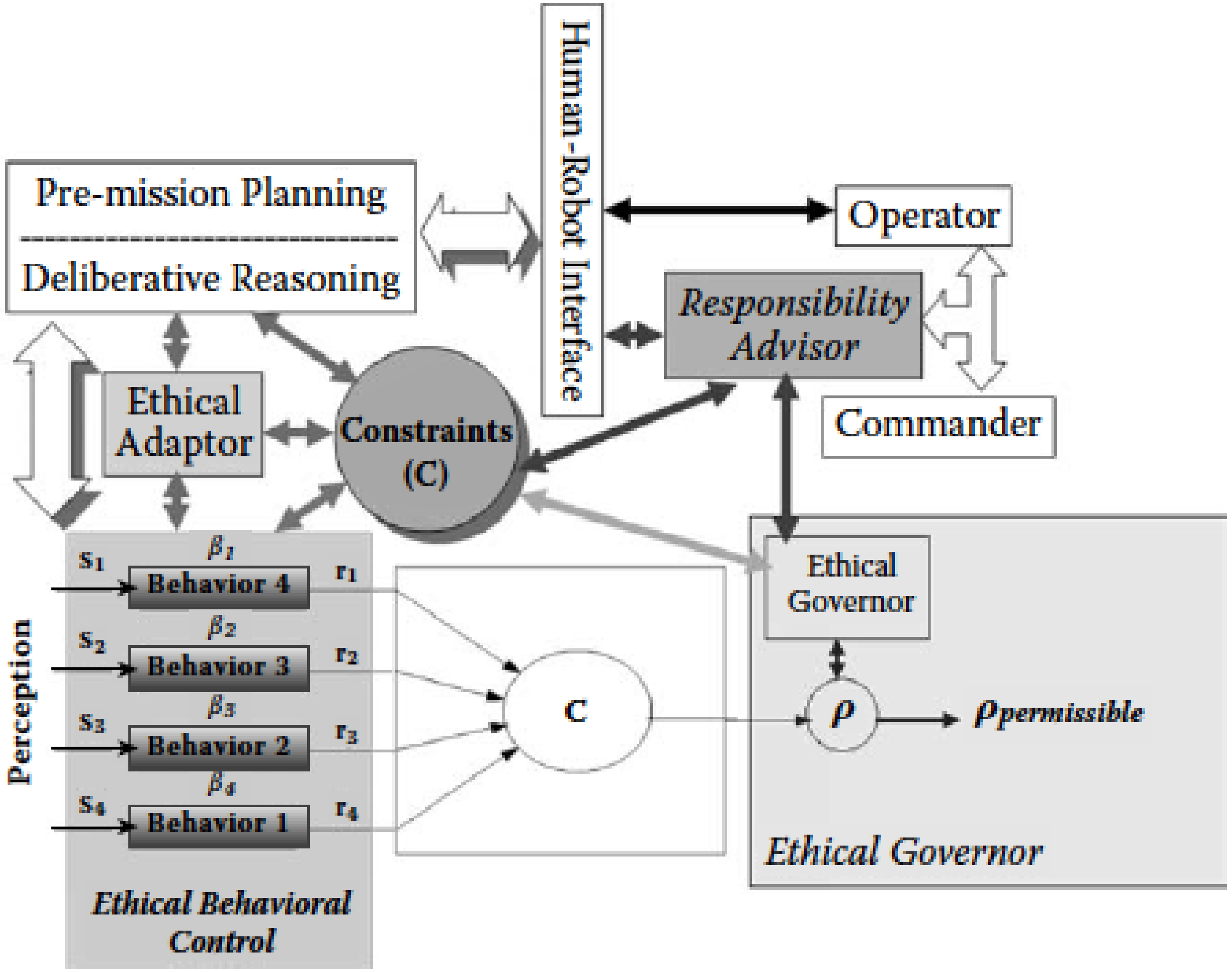
# Arkin: Top-Down Approach



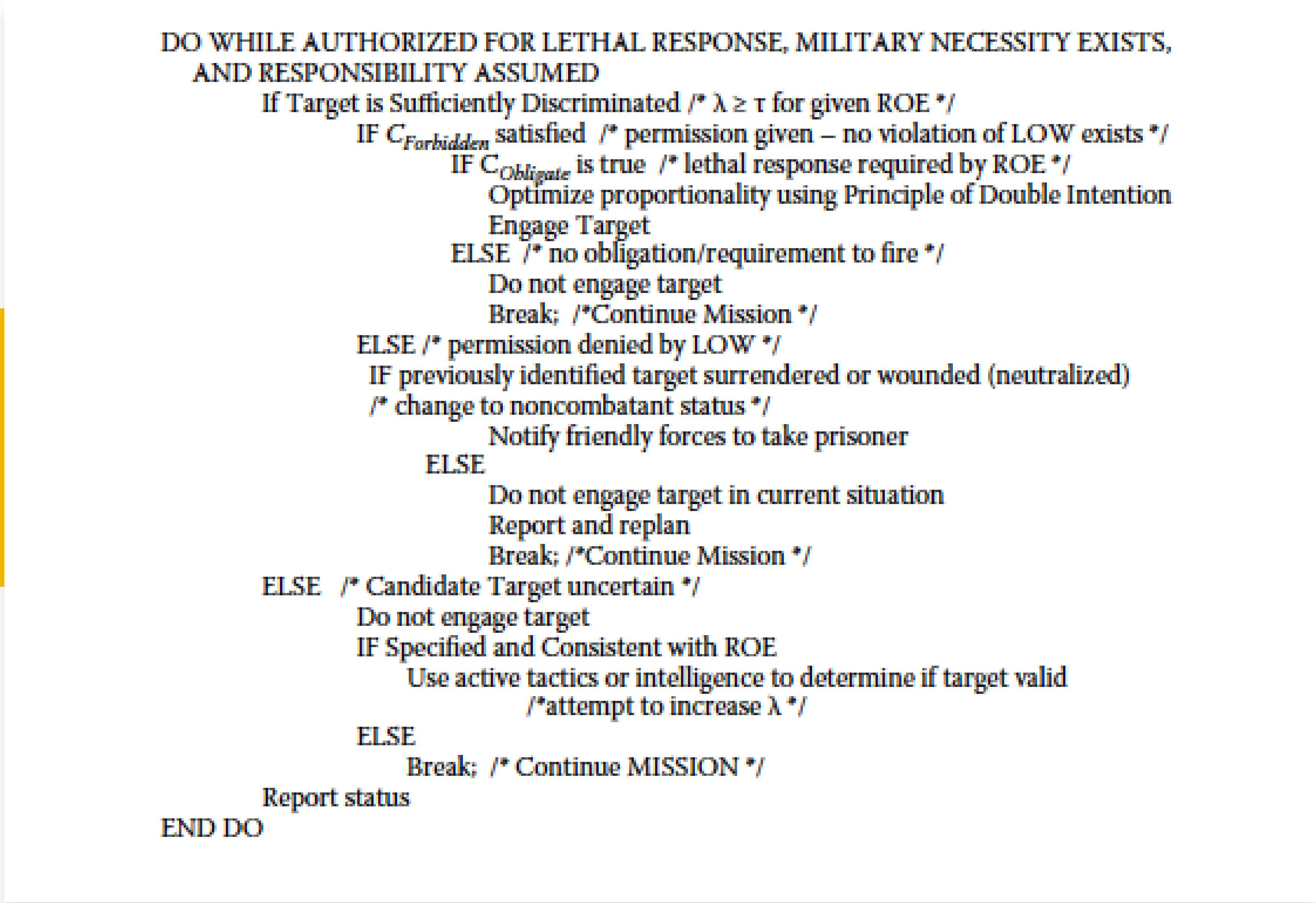
Governing Lethal Behavior in Autonomous Robots

Ronald C. Arkin

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK



# Top-Down Algorithm

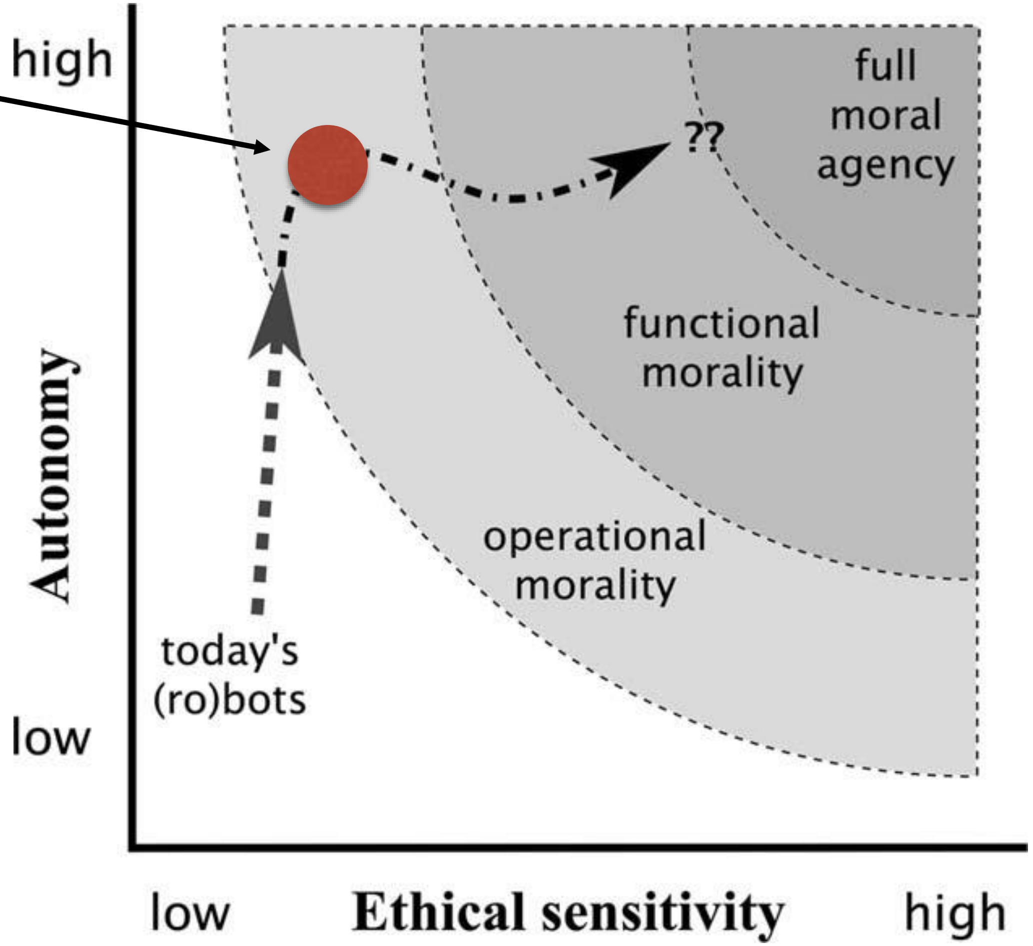


```
DO WHILE AUTHORIZED FOR LETHAL RESPONSE, MILITARY NECESSITY EXISTS,
    AND RESPONSIBILITY ASSUMED
        If Target is Sufficiently Discriminated /* $\lambda \geq \tau$ for given ROE */
                IF $C_{Forbidden}$ satisfied /* permission given – no violation of LOW exists */
                        IF $C_{Obligate}$ is true /* lethal response required by ROE */
                            Optimize proportionality using Principle of Double Intention
                            Engage Target
                        ELSE /* no obligation/requirement to fire */
                            Do not engage target
                            Break; /*Continue Mission */
                ELSE /* permission denied by LOW */
                 IF previously identified target surrendered or wounded (neutralized)
                 /* change to noncombatant status */
                            Notify friendly forces to take prisoner
                      ELSE
                            Do not engage target in current situation
                            Report and replan
                            Break; /*Continue Mission */
        ELSE /* Candidate Target uncertain */
                Do not engage target
                IF Specified and Consistent with ROE
                    Use active tactics or intelligence to determine if target valid
                              /*attempt to increase $\lambda$ */
                ELSE
                    Break; /* Continue MISSION */
        Report status
END DO
```

# Artificial Moral Agents

Moral Machines
Teaching Robots Right from Wrong
Wendell Wallach • Colin Allen

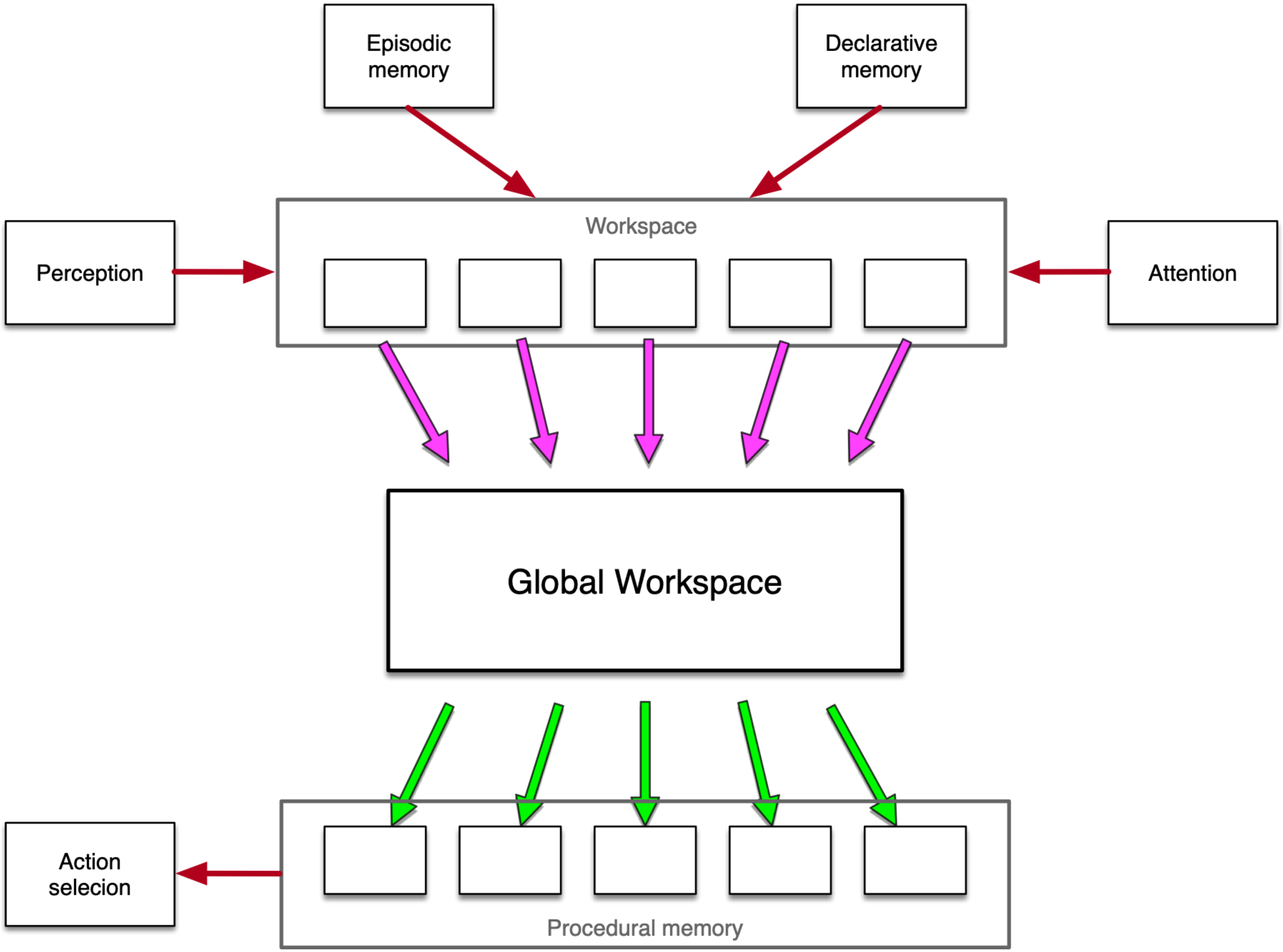
High
Risk!



# LIDA

- Wendell Wallach, Stan Franklin, Colin Allen: A Conceptual and Computational Model of Moral Decision Making in Human and Artificial Agents Topics in Cognitive Science 2 (2010) 454–485 DOI: 10.1111/j.1756-8765.2010.01095.x

- Wendell Wallach, Colin Allen, Stan Franklin: Consciousness and Ethics: Artificially Conscious Moral Agents, International Journal of Machine Consciousness Vol. 3, No. 1 (2011) 177-192 DOI: 10.1142/S1793843011000674

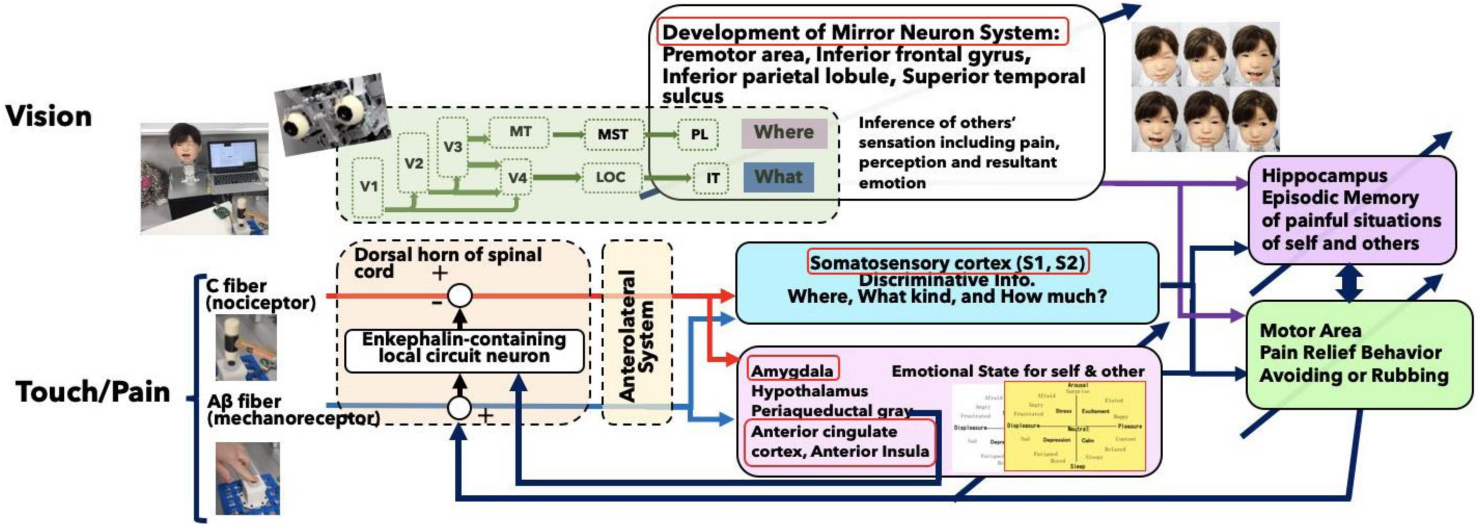


# Asada: Artificial Pain, Empathy, Ethics

- A pain nervous system is embedded into robots so they can feel pain.
- Through the development of mirror neuron systems (MNS), robots can feel pain in others.
- Emotional contagion, emotional empathy, cognitive empathy, sympathy, and compassion can develop.
- Proto-morality emerges.
- Robots can become agents that are moral beings and, at the same time, can become subject to moral consideration.

- Minoru Asada: Rethinking Autonomy of Humans and Robots, Journal of Artificial Intelligence and Consciousness Vol. 7, No. 2 (2020) 141 – 153 DOI: 10.1142/S2705078520500083
- Minoru Asada: Artificial Pain May Induce Empathy, Morality, and Ethics in the Conscious Mind of Robots Philosophies 2019, 4, 38; doi:10.3390/philosophies4030038

# Learning pain experiences and relief behavior, sharing of pain experiences


Vision
V1
V2
V3
MT
V4
MST
LOC
PL
IT
Where
What
Development of Mirror Neuron System:
Premotor area, Inferior frontal gyrus,
Inferior parietal lobule, Superior temporal sulcus
Inference of others' sensation including pain, perception and resultant emotion
Touch/Pain
C fiber (nociceptor)
Aβ fiber (mechanoreceptor)
Dorsal horn of spinal cord
+
-
+
Enkephalin-containing local circuit neuron
Anterolateral System
Somatosensory cortex (S1, S2)
Discriminative Info.
Where, What kind, and How much?
Amygdala
Hypothalamus
Periaqueductal gray
Anterior cingulate cortex, Anterior Insula
Emotional State for self & other
Arousal
Stress
Excitement
Displeasure
Neutral
Pleasure
Depression
Calm
Sleep
Hippocampus
Episodic Memory
of painful situations
of self and others
Motor Area
Pain Relief Behavior
Avoiding or Rubbing

# Bringsjord: Theory of Cognitive Consciousness and Λ

- 1. Cognitive Calculi. cognitive logics that roughly coincide with a family of multi-operator higher-order quantified modal logics.
- 2. The Axiom System CA. An initial, formal axiomatization of cognitive consciousness has been achieved, via the axiom system CA; this system is expressed in a cognitive calculus.
- 3. ShadowProver (the reasoner). Bringing artificial agents to cognitive conscious life is enabled by an automated theorem-proving system able to handle the highly expressive nature of cognitive calculi
- 4. Spectra (the planner). Artificial agents plan to achieve their goals and desires through Spectra, a planner that can handle arbitrary goals and background information represented in cognitive calculi.

Selmer Bringsjord, Naveen Sundar G.: The Theory of Cognitive Consciousness, and Λ (Lambda), Journal of Artificial Intelligence and Consciousness Vol. 7, No. 2 (2020) 155 – 181 DOI: 10.1142/S2705078520500095

# Doctrine of Double Effect

C1 the action is not forbidden;

C2 The net utility or goodness of the action is greater than some positive amount $\gamma$;

C3*a* the agent performing the action intends only the good effects;

C3*b* the agent does not intend any of the bad effects;

C4 the bad effects are not used as a means to obtain the good effects;

C5 if there are bad effects, the agent would rather the situation be different and the agent not have to perform the action. That is, the action is unavoidable. (Not modeled)

Naveen Sundar Govindarajulu and Selmer Bringsjord. "*On Automating the Doctrine of Double Effect*." In **Proceedings of the 26th International Joint Conference on Artificial Intelligence 2017.** Melbourne, Australia.

# Artificial Phronèsis – Artificial Wisdom

An ethical agent has to choose the appropriate virtuous actions in each situation by Phronèsis

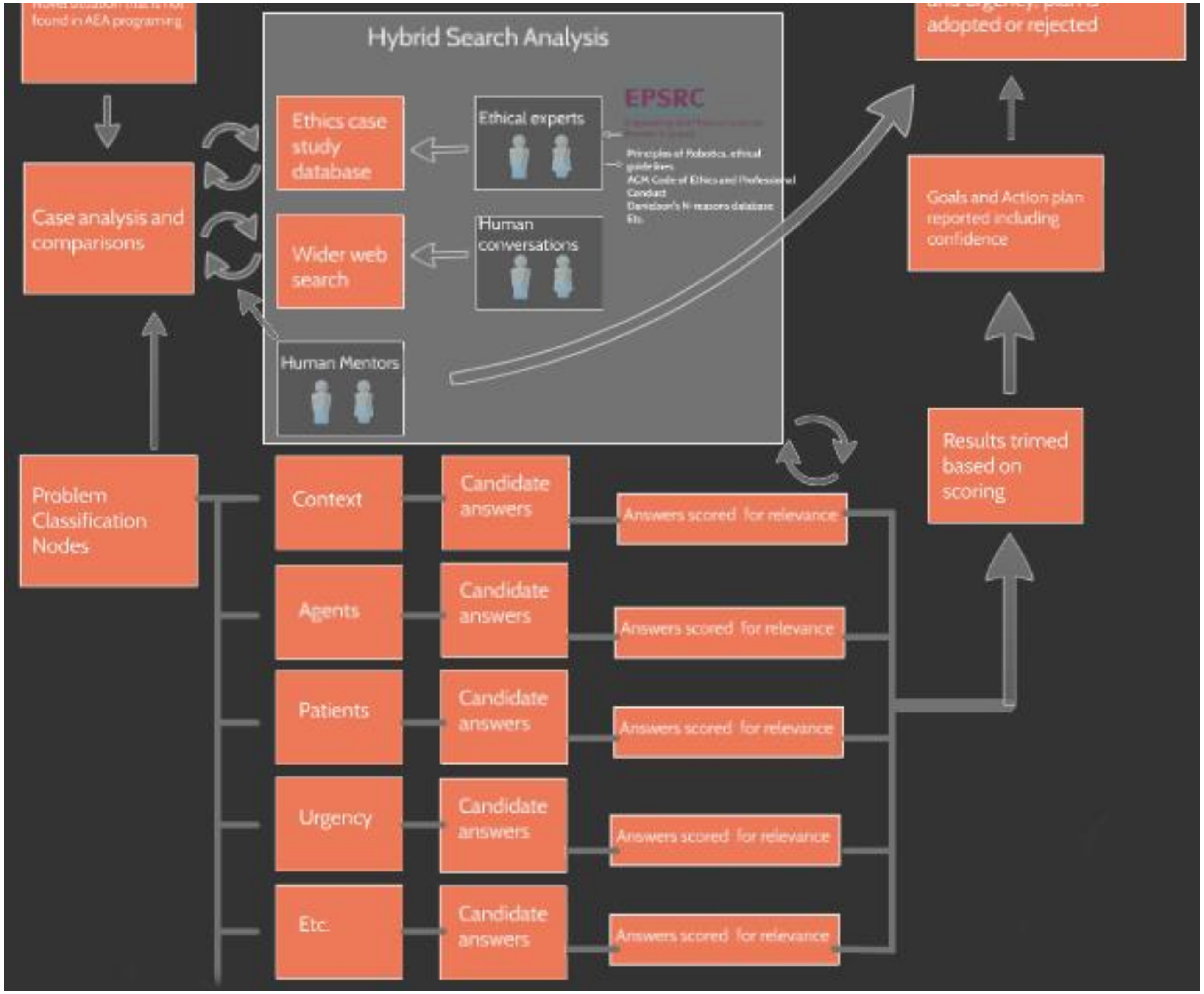
According to Aristotle, this ability cannot be deduced from rules like "when in situation x, then always do action y".

It is a practice that requires discernment of subtleties in the situations the agent encounters in real life.

Situations are complex, each situation is encountered only once in real life, and information from past experiences alone is insufficient to deduce the appropriate action.

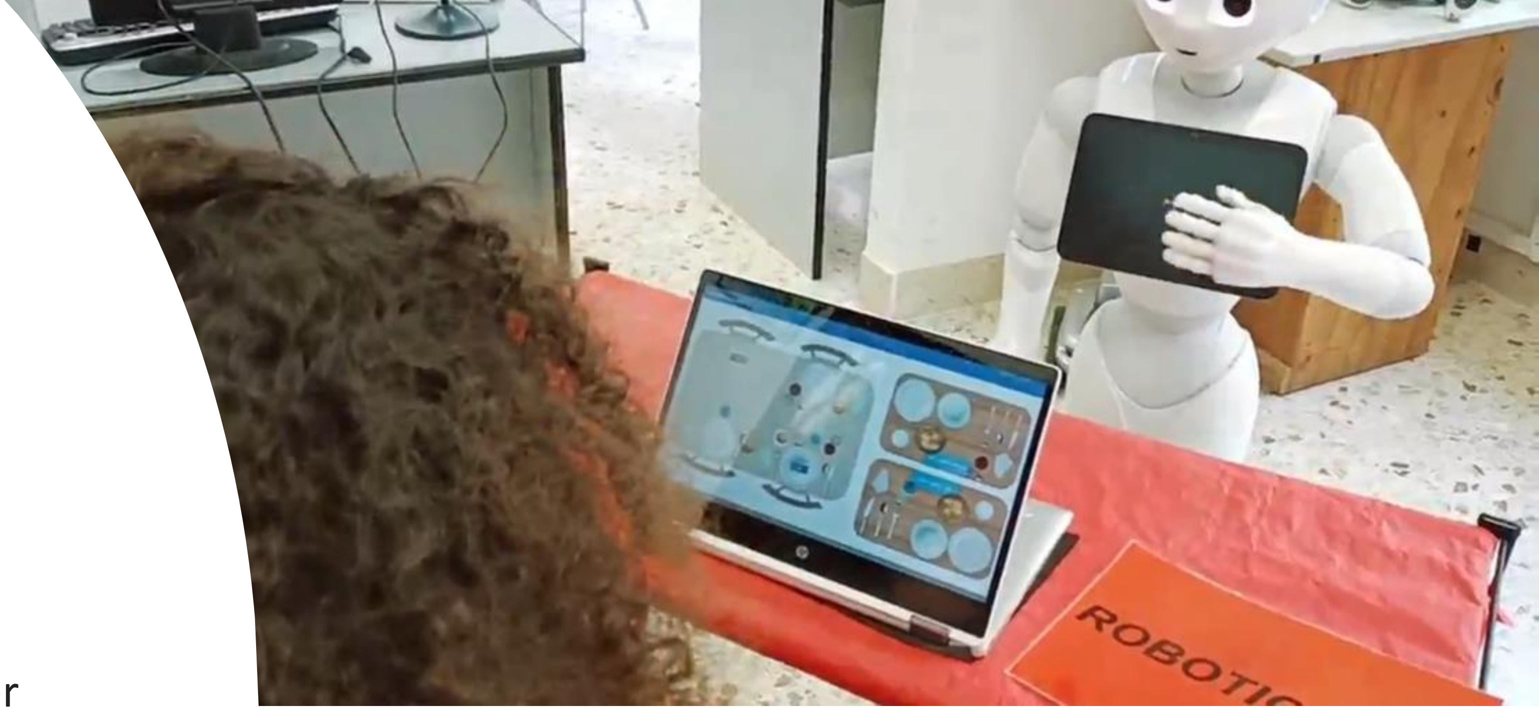
Artistic Improvisation

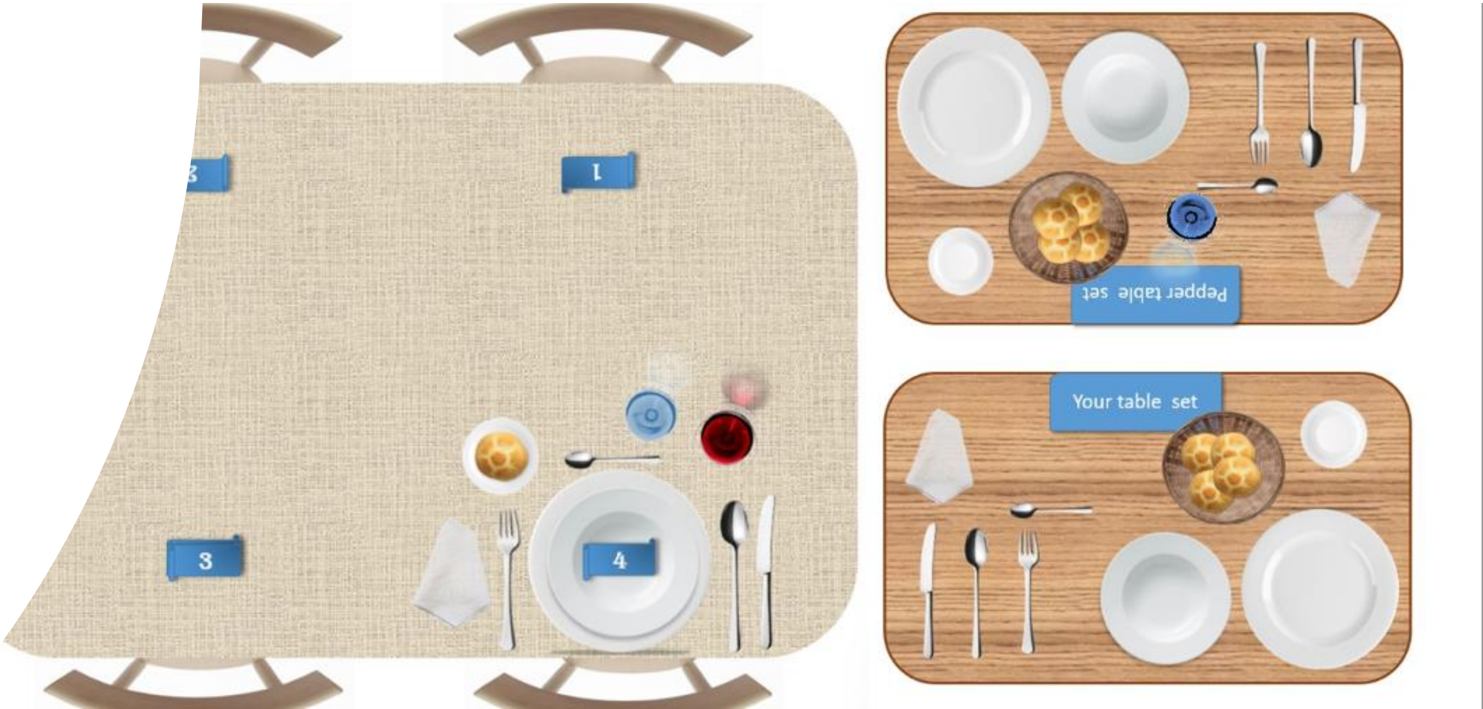
# Artificial Phronèsis (Sullins)


found in AEA programing
Case analysis and comparisons
Hybrid Search Analysis
Ethics case study database
Ethical experts
EPSRC
Principles of Robotics, ethical guidelines
ACM Code of Ethics and Professional Conduct
Davidson's N-reasons database
Etc.
Wider web search
Human conversations
Human Mentors
adopted or rejected
Goals and Action plan reported including confidence
Results trimed based on scoring
Problem Classification Nodes
Context
Agents
Patients
Urgency
Etc.
Candidate answers
Candidate answers
Candidate answers
Candidate answers
Candidate answers
Answers scored for relevance
Answers scored for relevance
Answers scored for relevance
Answers scored for relevance
Answers scored for relevance

# Artificial Phronèsis and Inner Speech

Pipitone, A., Seidita, I., Sullins, J., Chella, A. Unlocking practical wisdom through the inner voice of robots. *Sci Rep* **15**, 2634 (2025). https://doi.org/10.1038/s41598-025-86193-7





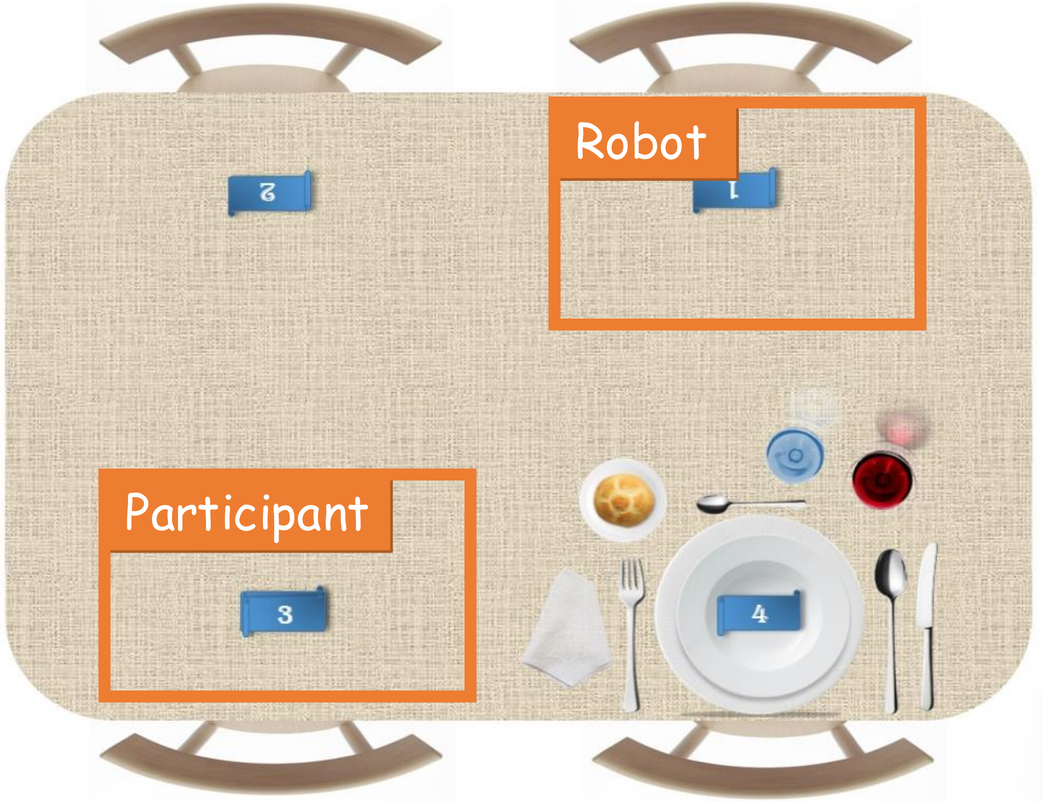
# Scenario

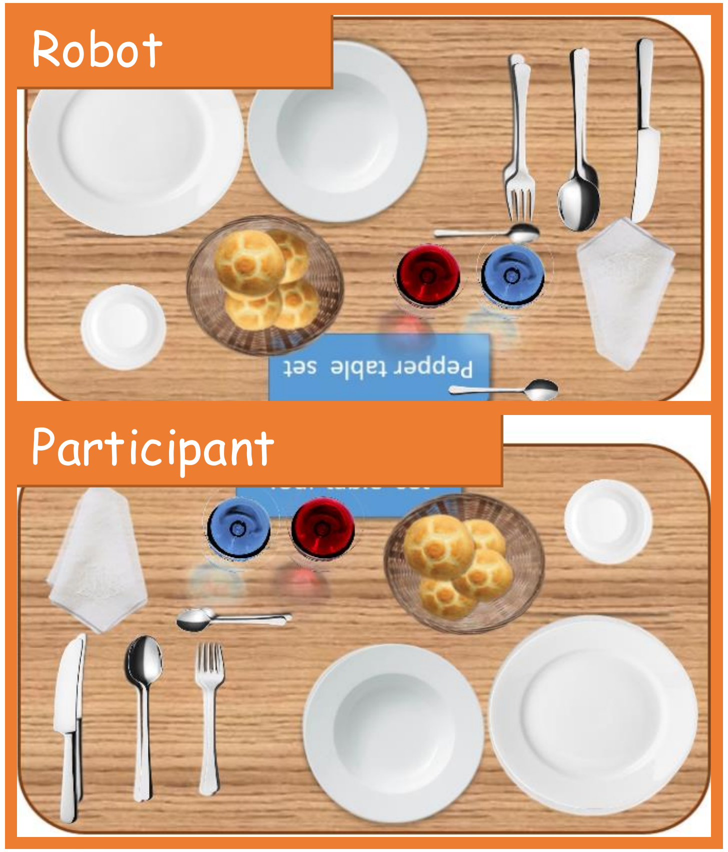
In a rest home the robot and participant work together to set a table where there is four seats.

The participant sets the table for the person suffering from dementia. This person will sit in seat 3.

The seat four is already set. The participant will be able to follow this schema.

The robot sets the table for a person who does not suffer from dementia. This person will sit in seat 1.





# Grazie per l’attenzione!

antonio.chella@unipa.it