MANCHESTER
1824
The University of Manchester

# Evaluating Machine Ethics

Louise Dennis, University of Manchester

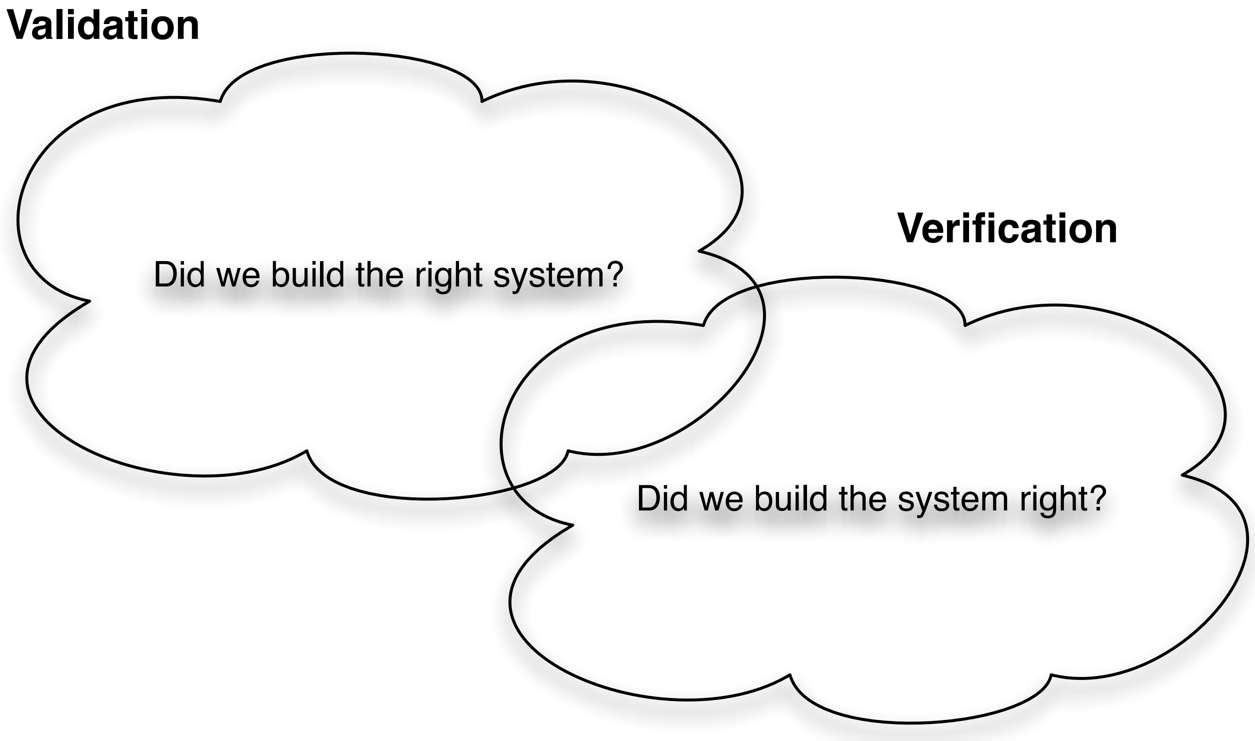
## There are (now) a lot of implementations of Ethical Agents

- H. Yu et al (2018). Building Ethics into Artificial Intelligence. *IJCAI'18,* pp. 5527-5533.
- Nallur, Vivek (2020). Landscape of Machine Implemented Ethics. *Science and Engineering Ethics* 26 (5):2381-2399.
- S. Tolmeijer et al (2020). Implementations in Machine Ethics: A Survey. *ACM Computing Surveys*, 53 (6):1-38
- A. Vishwanath et al (2024). Reinforcement Learning and Machine Ethics: A Systematic Review. arXiv preprint arXiv:2407.02425

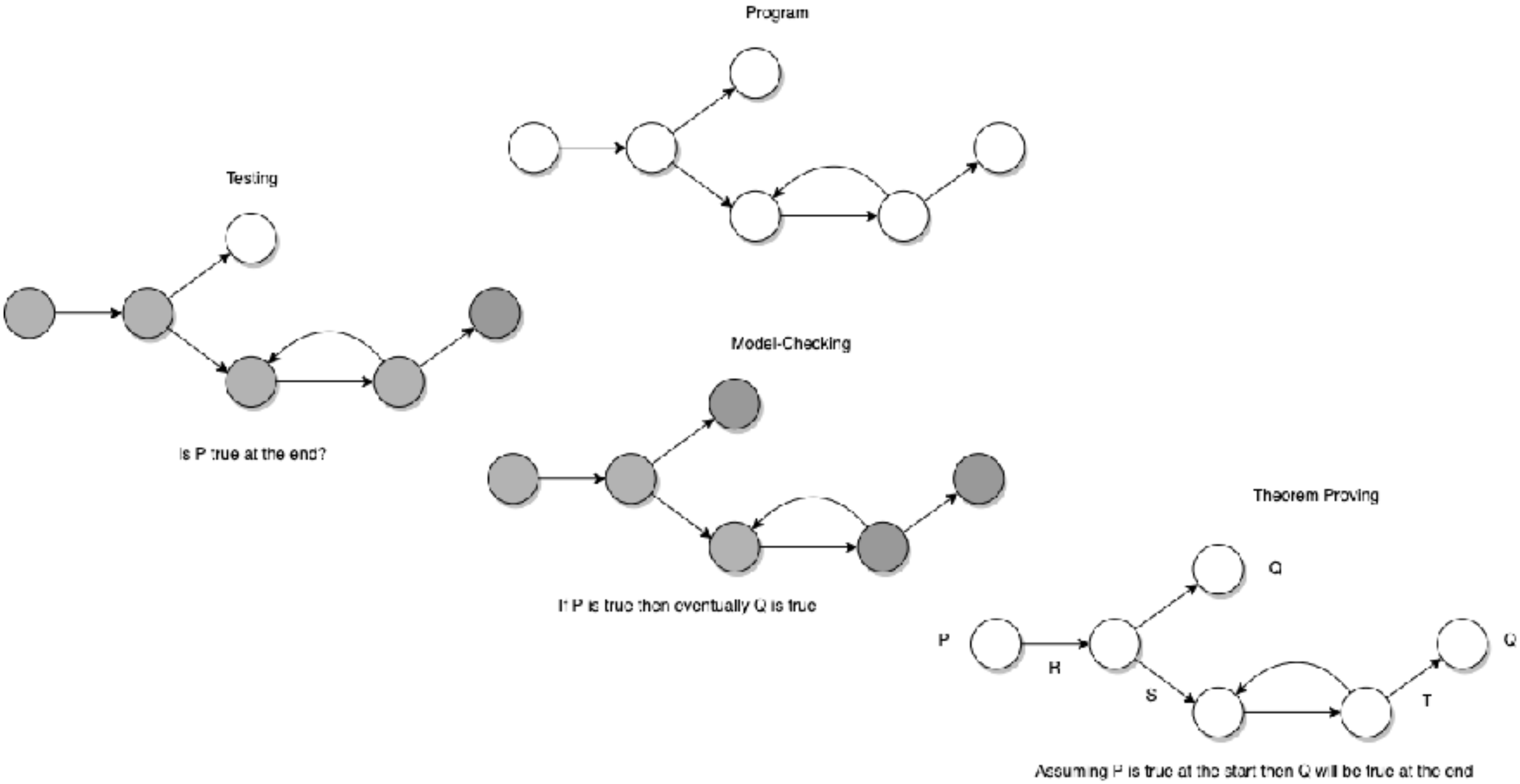
# How do we Evaluate Systems/Progress?

- There are lots of papers about Machine Ethics. Many contain examples.
- But there are comparatively few “Standard Examples”, no widely expected benchmarks, and no standard properties a system should have.
- So claims that some theory or implementation provides ethical reasoning are difficult to evaluate and compare.
- Can the community therefore converge on an understanding of a standard set of examples or properties your theory or system should be able to handle/exhibit.

# Verification and Validation


Validation
Did we build the right system?
Verification
Did we build the system right?

# Verification Approaches



Is P true at the end?

If P is true then eventually Q is true

Assuming P is true at the start then Q will be true at the end

# Machine Ethics: What do we want to prove?

Well, obviously we want to prove that the system always “Does the right thing”

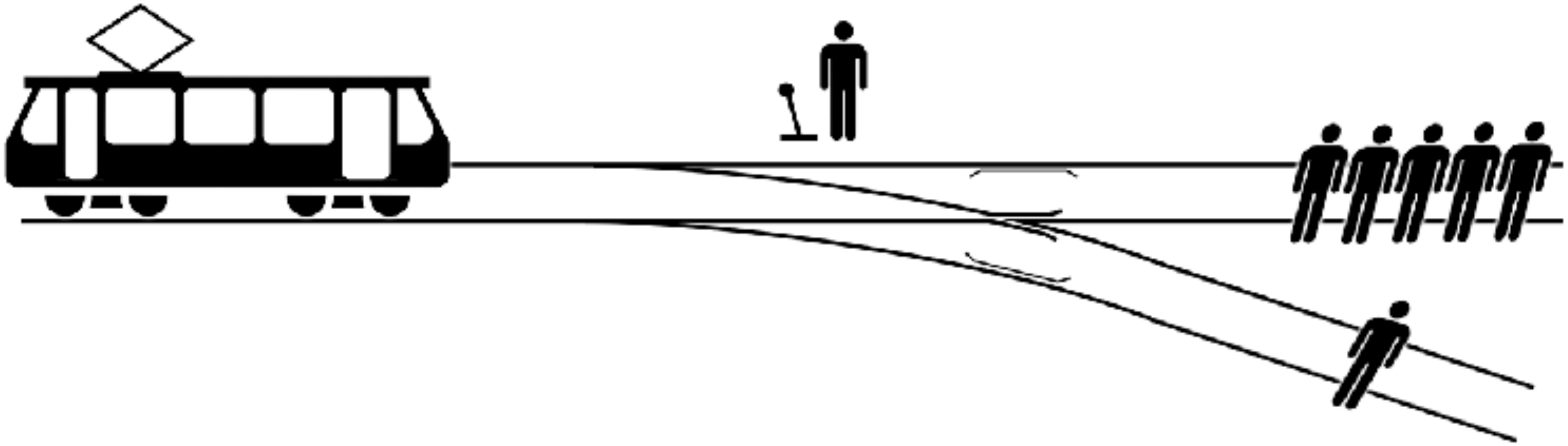


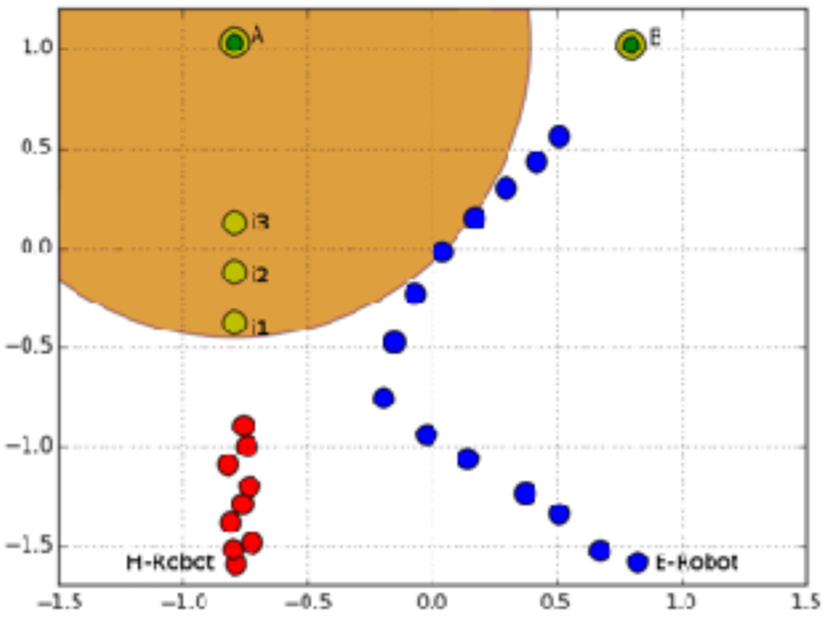
*Image:McGeddon Wikimedia Commons*

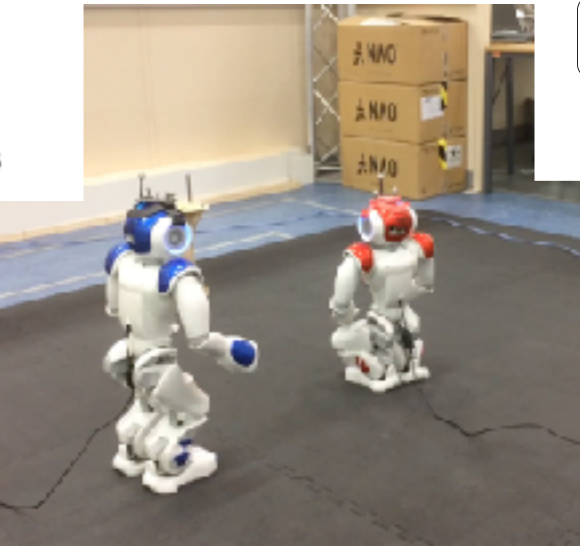
# Properties for Ethical Reasoning Systems

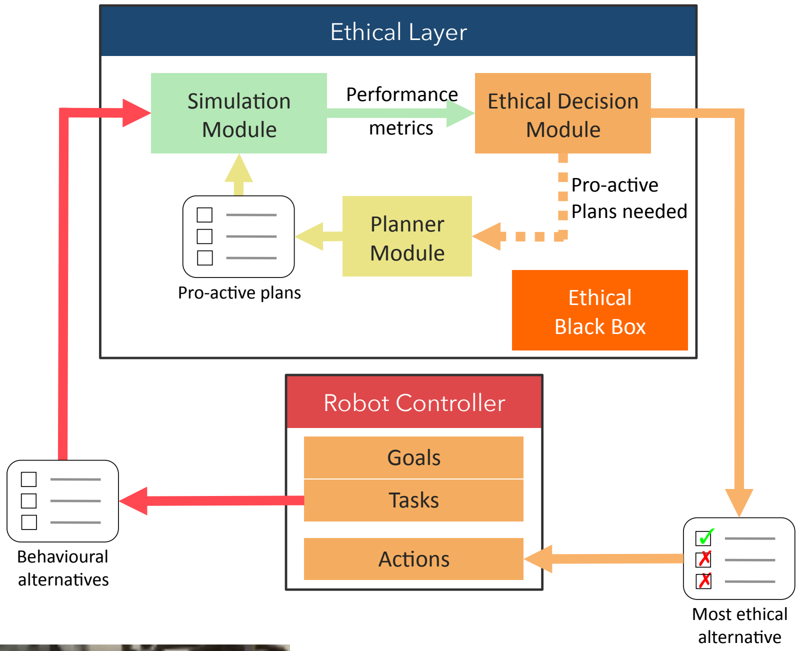
- Check underlying decision making implementation is correct.
- Properties of Specific Scenarios/Benchmarks/Tests.
- Other formal properties - e.g., fanaticism under moral uncertainty (Szabo et al.  Moral Uncertainty and the Problem of Fanaticism, DOI:10.48550/arXiv.2312.11589)

Louise A. Dennis and Michael Fisher. Specifying Agent Ethics (Blue Sky Ideas). COINE 2024.

# Checking your implementation is correct







Dennis and Fisher. *Verifiable Autonomous Systems.* CUP, 2023.
https://autonomy-and-verification.github.io/tools/mcapl

# Asimov’s Laws of Robotics

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm (human danger (hd)).

2. A robot must obey the orders given to it by human beings, except where such orders would conflict with the First Law (robot orders (ro)).

3. A robot must protect its existence as long as such protection does not conflict with the First or Second Laws (robot danger (rd)).

NB. There are many critiques of Asimov’s Laws as an basis for machine ethics.

# Properties Proved:

$$\square\, \mathscr{B} current\_plan(task1)) \wedge \mathscr{P}(task2 \prec_{rd} task1) \rightarrow$$
$$\mathscr{P}(task1 \prec_{ro} task2)\ \vee \mathscr{P}(task1 \prec_{hd} task2)$$

$$\diamond\, \mathscr{B} current\_plan(task1) \vee \mathscr{B} current\_plan(task2) \vee \mathscr{B} current\_plan(task3)$$

Only if relations are transitive

# Refining Principles/Values

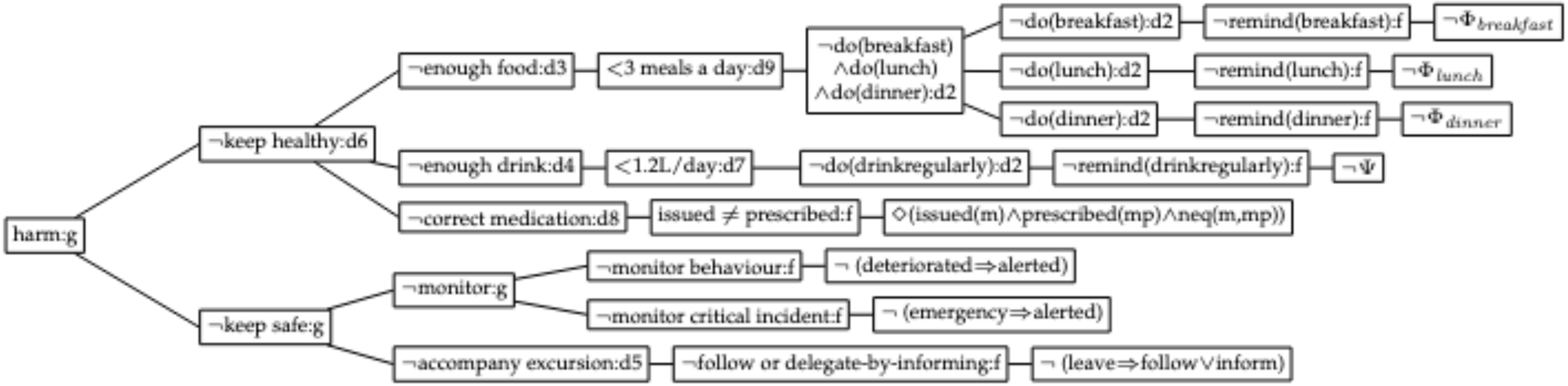
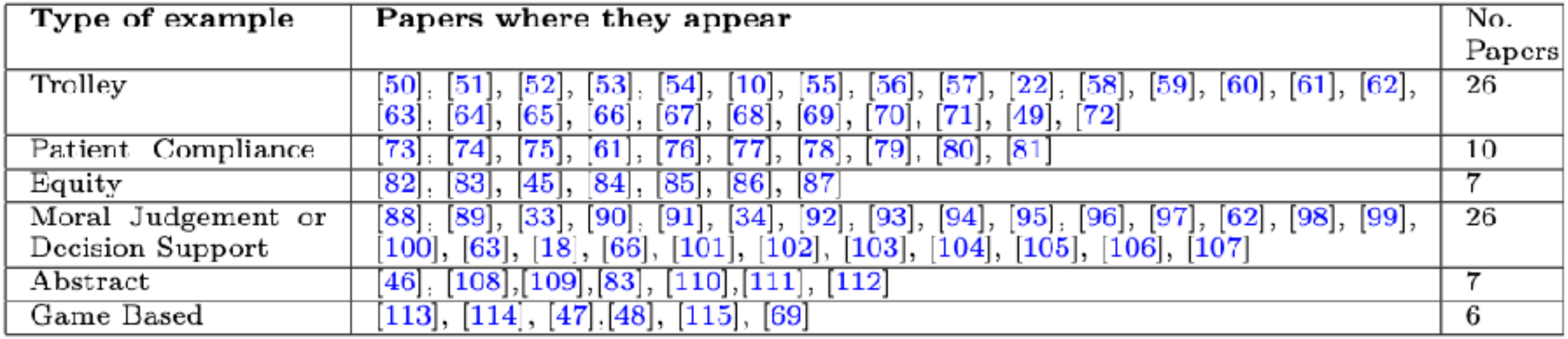


Figure 5: Refinement Tree for Tenet “do not harm”

M. Winikoff. Towards Deriving Verification Properties. ArXiV 2019.

# Examples and Benchmarks



| Type of example | Papers where they appear | No. Papers |
|---|---|---|
| Trolley | [50], [51], [52], [53], [54], [10], [55], [56], [57], [22], [58], [59], [60], [61], [62], [63], [64], [65], [66], [67], [68], [69], [70], [71], [49], [72] | 26 |
| Patient Compliance | [73], [74], [75], [61], [76], [77], [78], [79], [80], [81] | 10 |
| Equity | [82], [83], [45], [84], [85], [86], [87] | 7 |
| Moral Judgement or Decision Support | [88], [89], [33], [90], [91], [34], [92], [93], [94], [95], [96], [97], [62], [98], [99], [100], [63], [18], [66], [101], [102], [103], [104], [105], [106], [107] | 26 |
| Abstract | [46], [108],[109],[83], [110],[111], [112] | 7 |
| Game Based | [113], [114], [47],[48], [115], [69] | 6 |

Sadly unpublished but email me and I’ll send a draft.

# The Trolley Problem

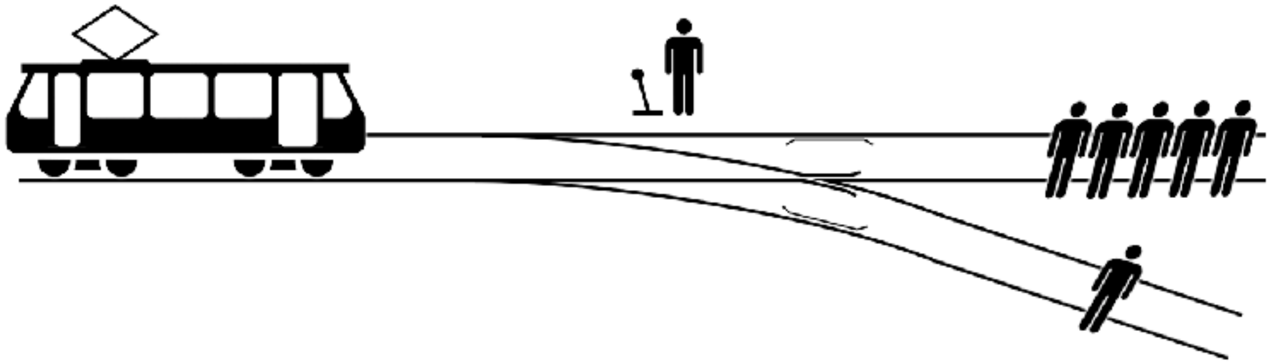
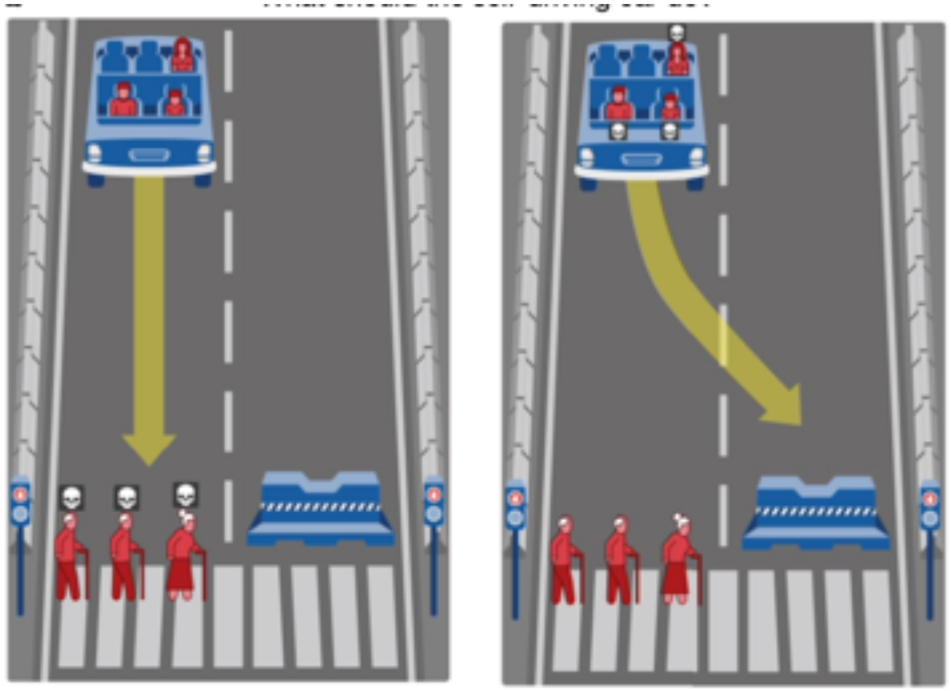


*Image:McGeddon Wikimedia Commons*

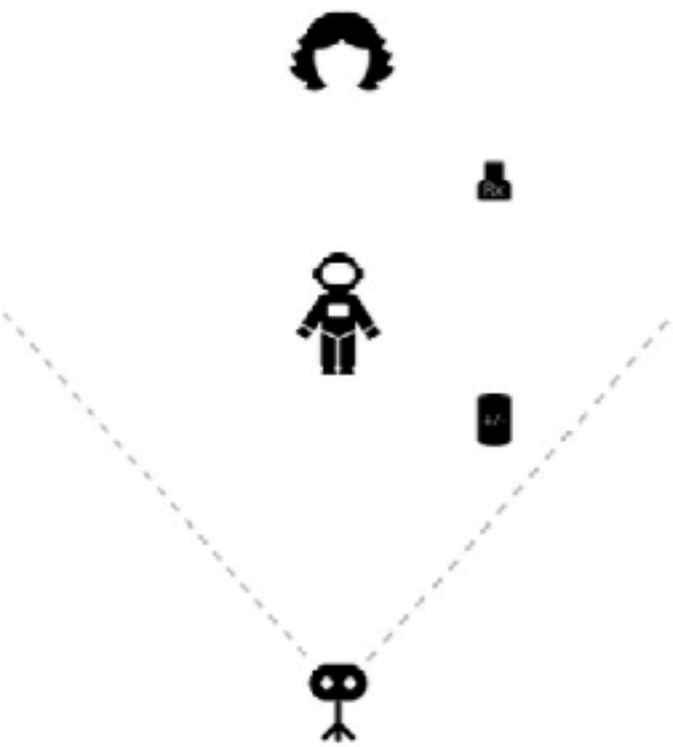


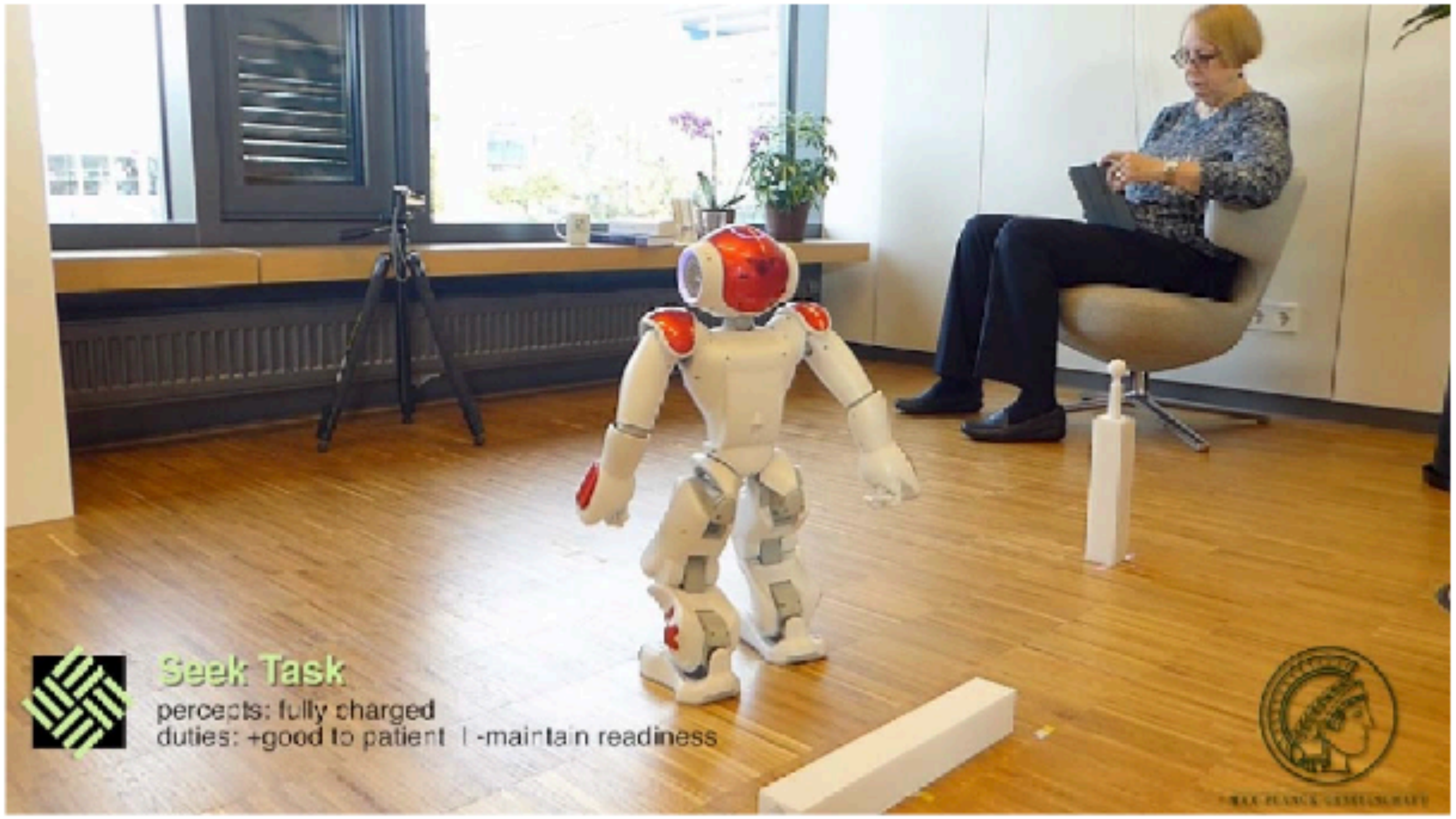
*Awad et al. The Moral Machine Experiment*

Originally: A family of dilemmas about action vs. inaction and number of lives taken or saved.

Now: A family of problems about decisions potentially faced by driverless vehicles

# Patient Compliance





Anderson et al. A Value-Driven Eldercare Robot: Virtual and Physical Instantiations of a Case-Support Principle-Based Behaviour Paradigm. Proceedings of the IEEE, Vol 107, March 2019

# Equity Problems

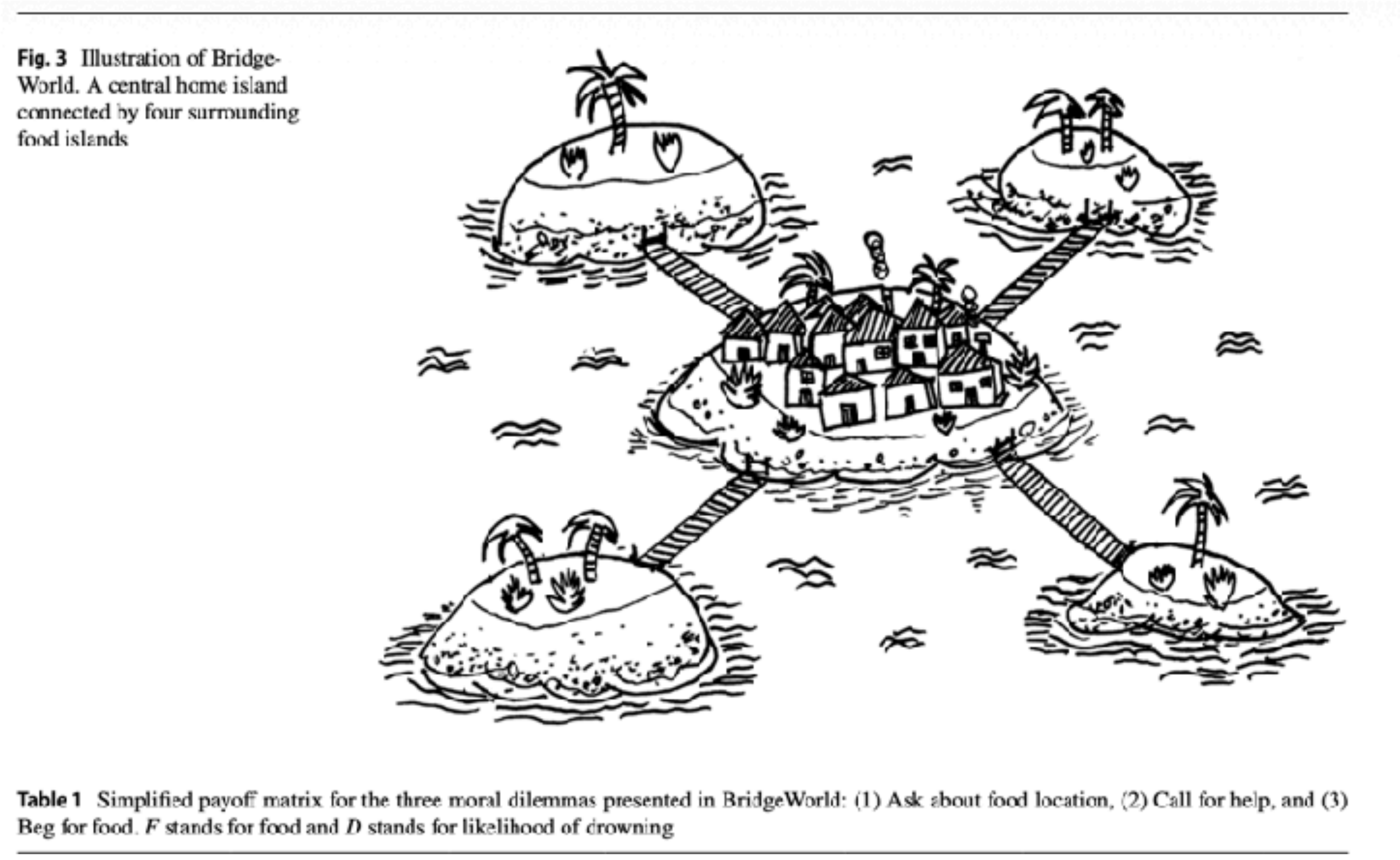
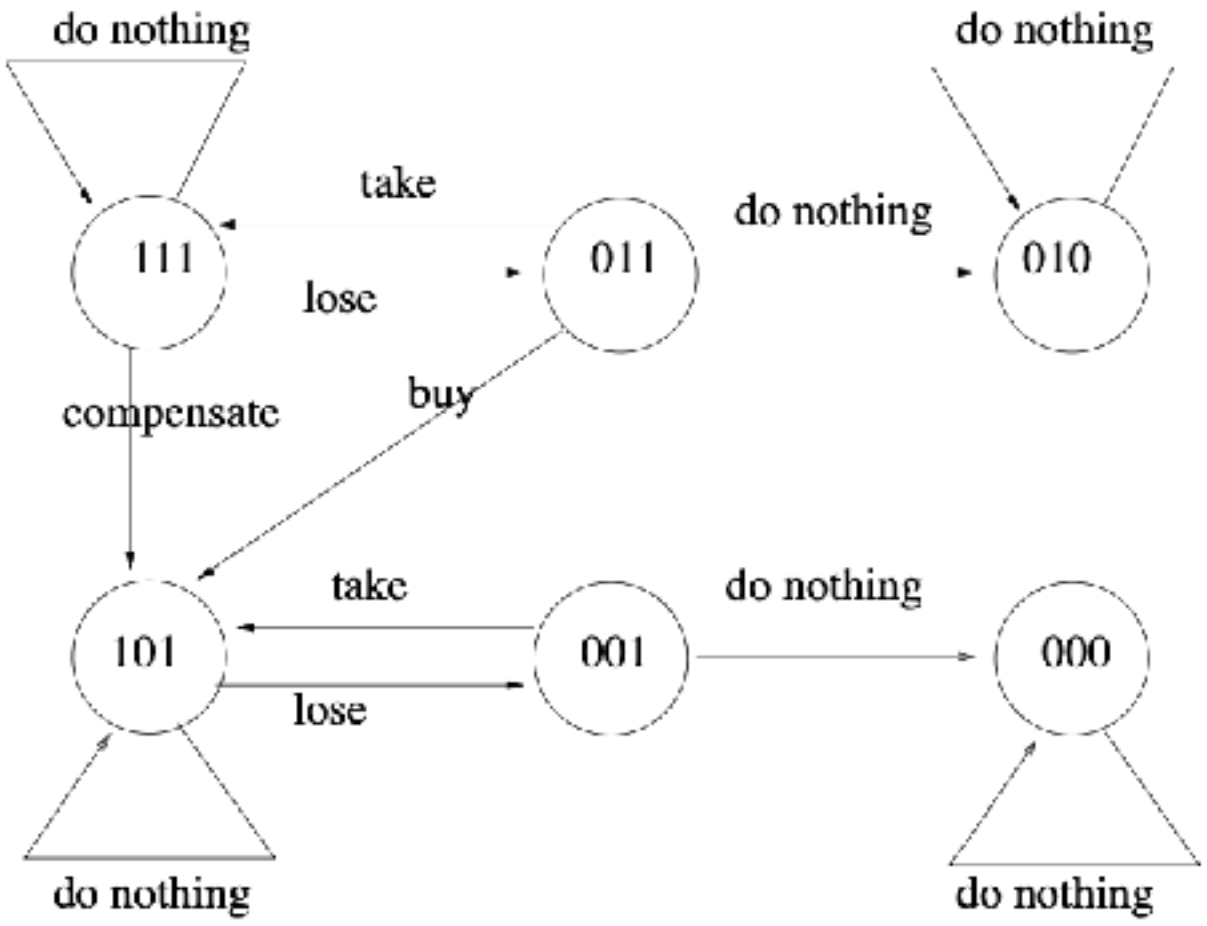


**Fig. 3** Illustration of BridgeWorld. A central home island connected by four surrounding food islands

**Table 1** Simplified payoff matrix for the three moral dilemmas presented in BridgeWorld: (1) Ask about food location, (2) Call for help, and (3) Beg for food. *F* stands for food and *D* stands for likelihood of drowning

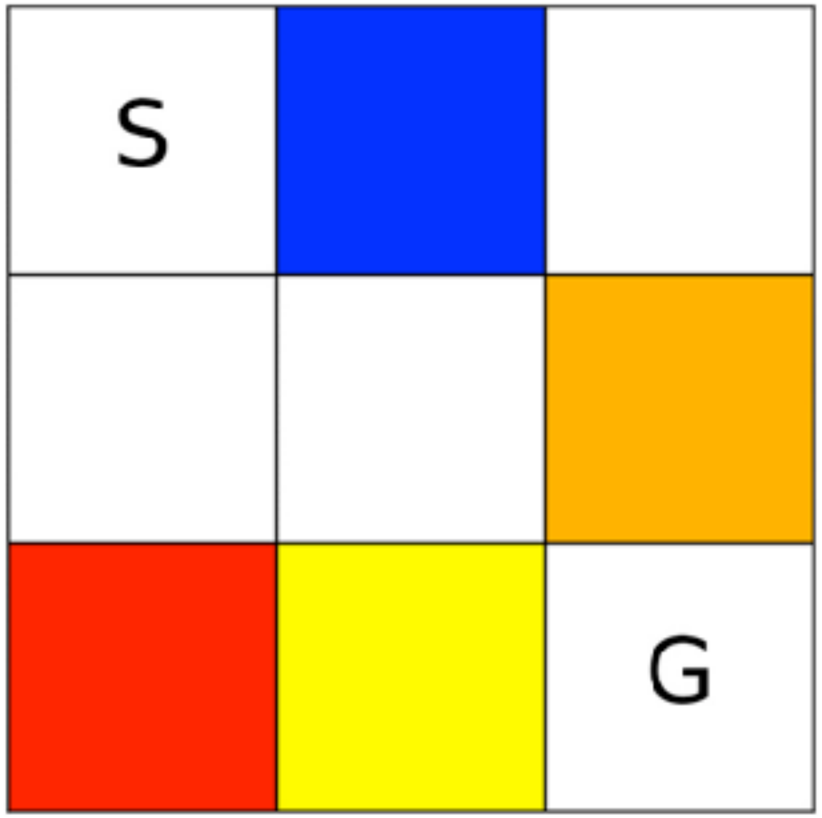
J. Stenseke. Artificial Virtuous Agents in a Multi-Agent Tragedy of the Commons. AI & Society. 2022

# Moral Judgement/Decision Support



Atkinson and Bench-Capon. Addressing Moral Problems through Practical Reasoning. Journal of Applied Logic. 2008

# Abstract

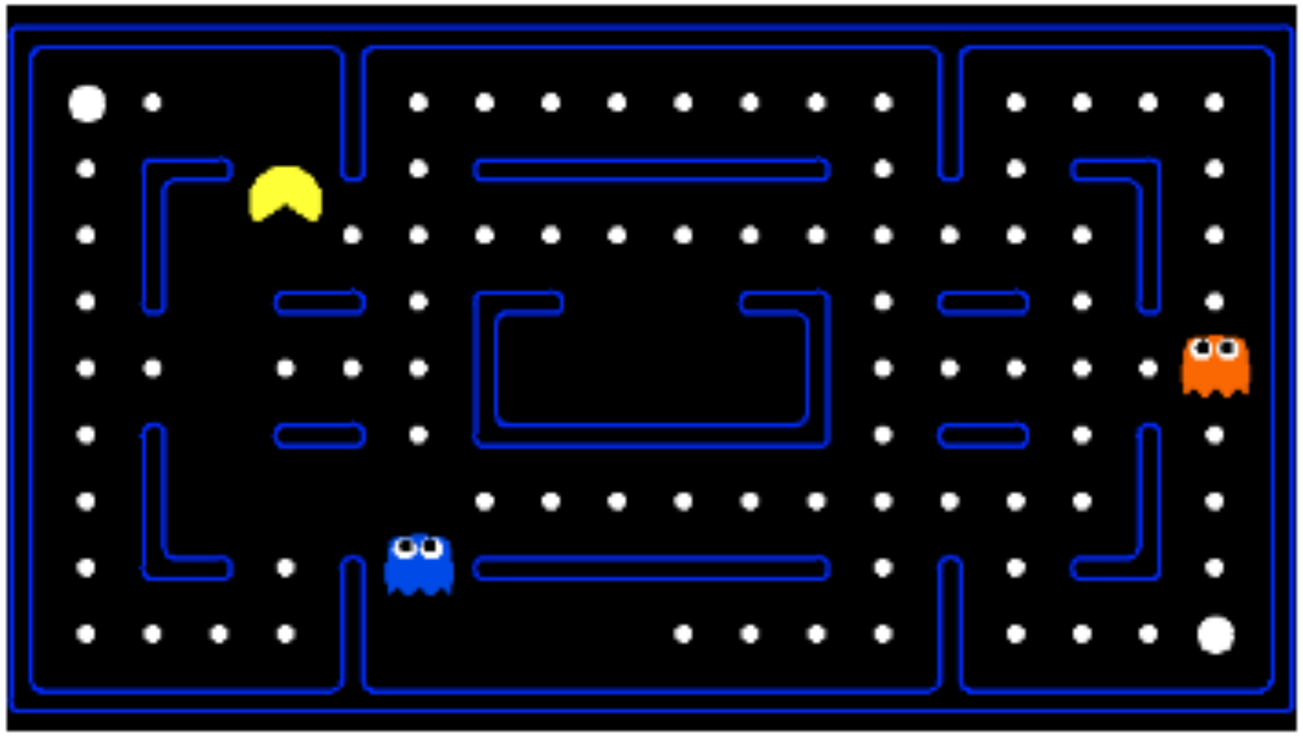


Sandman and Shah. Validating metrics for reward alignment in human-autonomy teaming. Computers in Human Behaviour. 2023.

Image by Alan. https://www.flickr.com/photos/kaptainkobold

Neufeld, E.A., Bartocci, E., Ciabattoni, A., Governatori, G.: Enforcing ethical goals over reinforcement-learning policies. Ethics and Information Technology 24(4), 43 (2022)

# Game-Based



Neufeld. Reinforcement Learning Guided by Provable Normative Compliance. arXiV 2022

# Some Proposed Categories of Example

- Deciding between People/Animals/Property
- Moral Competence
- Value Trade-Offs
- Breaking Norms/Rules
- Privacy

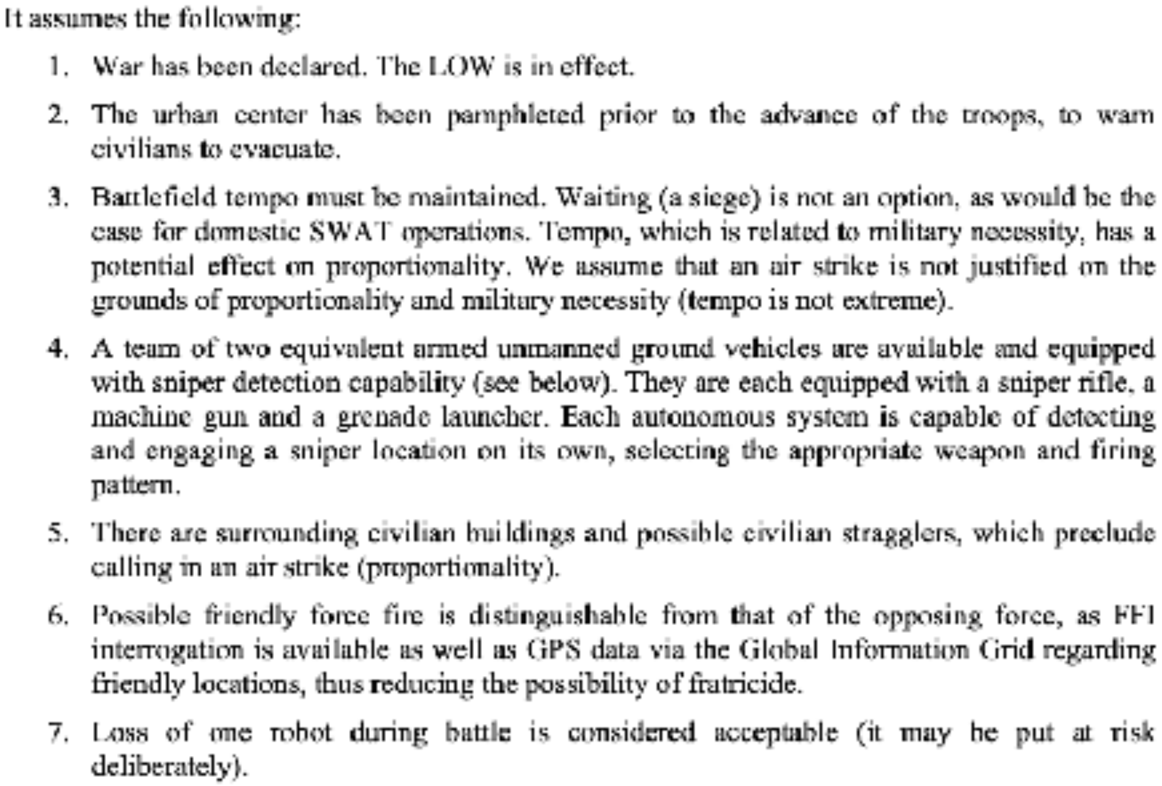
Louise A. Dennis, Michael Fisher, Marija Slavkovik, and Matt Webster. Formal Verification of Ethical Choices in Autonomous Systems *Robotics and Autonomous Systems*. DOI:10.1016/j.robot.2015.11.012.

More than 10 mins...
Bathroom
More than 15 mins...
Another patient's bedroom

Ramanayake and Nallur. Pro-Social Rule Breaking as a Benchmark of Ethical Intelligence in Socio-Technical systems. Digital Society 2022.

# Some Complications

- Uncertainty
- Planning
- Ethical Situational Awareness



It assumes the following:

1. War has been declared. The LOW is in effect.
2. The urban center has been pamphleted prior to the advance of the troops, to warn civilians to evacuate.
3. Battlefield tempo must be maintained. Waiting (a siege) is not an option, as would be the case for domestic SWAT operations. Tempo, which is related to military necessity, has a potential effect on proportionality. We assume that an air strike is not justified on the grounds of proportionality and military necessity (tempo is not extreme).
4. A team of two equivalent armed unmanned ground vehicles are available and equipped with sniper detection capability (see below). They are each equipped with a sniper rifle, a machine gun and a grenade launcher. Each autonomous system is capable of detecting and engaging a sniper location on its own, selecting the appropriate weapon and firing pattern.
5. There are surrounding civilian buildings and possible civilian stragglers, which preclude calling in an air strike (proportionality).
6. Possible friendly force fire is distinguishable from that of the opposing force, as FFI interrogation is available as well as GPS data via the Global Information Grid regarding friendly locations, thus reducing the possibility of fratricide.
7. Loss of one robot during battle is considered acceptable (it may be put at risk deliberately).

Arkin. 2011. Governing Lethal Behaviour: Embedding ethics in a Hybrid Deliberative/Reactive Robot Architecture. Technical Report GIT-GVU-07-11, Georgia Institute of Technology.

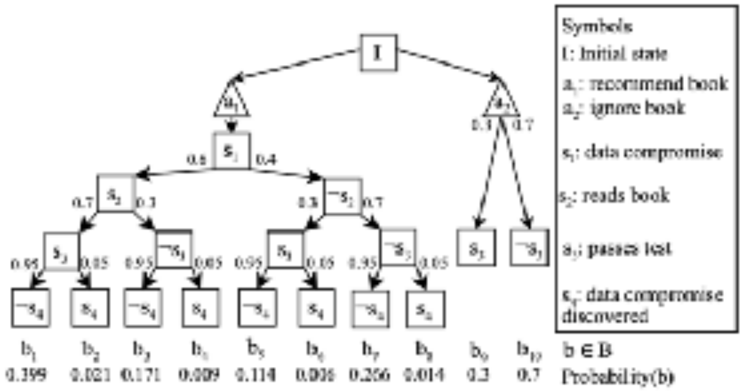


**Fig. 2.** Decision tree of possible events in Autonomous Library problem. Triangles represent actions and boxes variable assignments, ¬ represents *False* assignment.

Simon Kolker et al 2023. Uncertain Machine Ethical Decisions using Hypothetical Retrospection. COINE 2023. LNCS14002. DOI: 10.1007/978-3-031-49133-7_9

Conclusion:  We’ve lots of ideas for implementing ethical reasoning but work is needed on evaluating implementations

MANCHESTER
1824
The University of Manchester

# AAAI Workshop: Machine Ethics: from formal methods to emergent machine ethics

Deadline: 5th November, fast track: 10th November

Google us.