MANCHESTER
1824
The University of Manchester

Verifiable Machine Ethics

Louise Dennis, University of Manchester

With help from, among others, Michael Fisher, Marija Slavkovik (and students), Matt Webster, Alan F. Winfield, Paul Bremner, Felix Lindner, Martin Mose Bentzen, Rafael Cardoso, Angelo Ferrando, Tom Evans, Daniel Ene, Cristina Perea del Olmo, Simon Kolker

All the work discussed in this talk is available as part of the MCAPL (Model-Checking Agent Programming Languages) Framework.

https://autonomy-and-verification.github.io/tools/mcapl

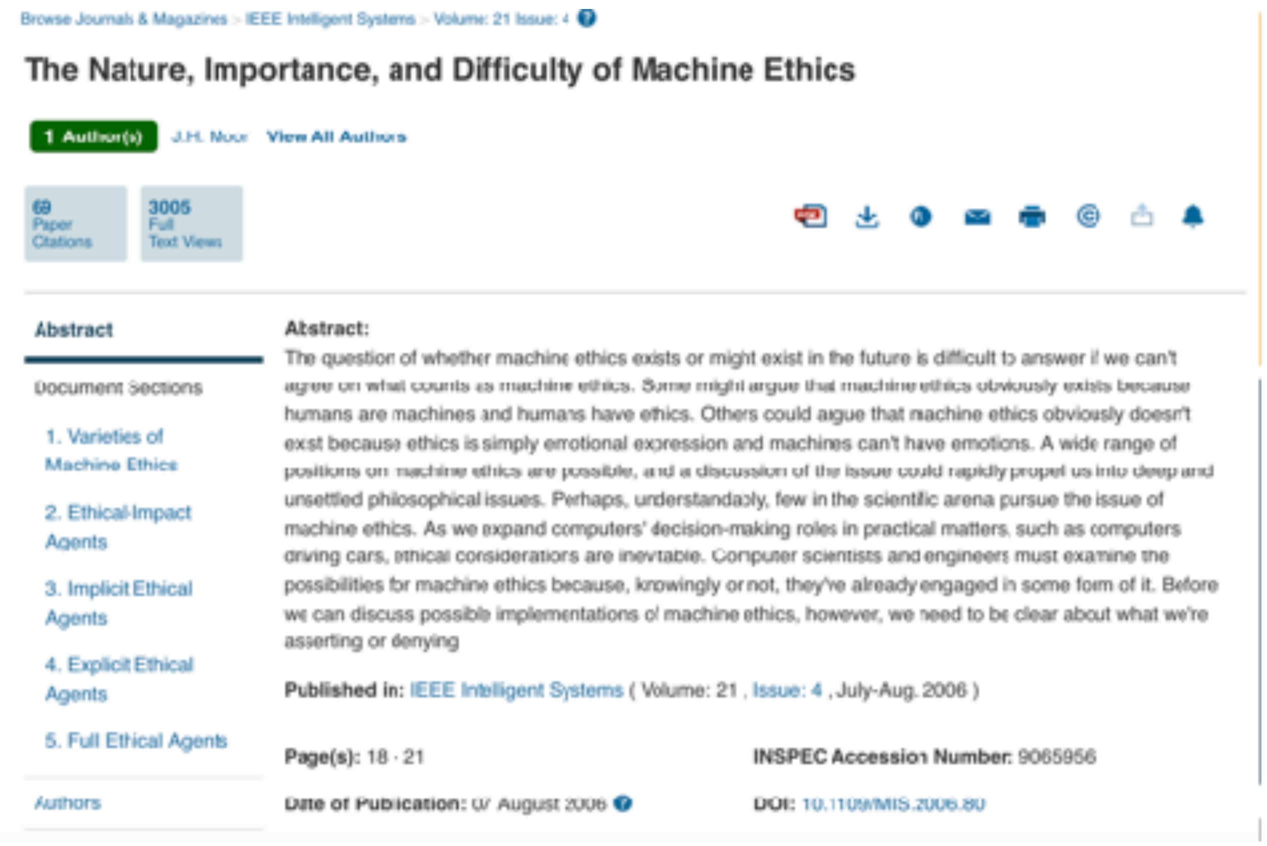
What is Machine Ethics?

How to automate moral reasoning?

Types of artificial moral agents

James H Moor. 2006. The nature, importance, and difficulty of machine ethics. IEEE intelligent systems 21, 4 (2006), 18–21.

- Ethical-impact agents
- Implicit ethical agents
- Explicit ethical agents
- Full ethical agents



Browse Journals & Magazines > IEEE Intelligent Systems > Volume: 21 Issue: 4

The Nature, Importance, and Difficulty of Machine Ethics

1 Author(s) J.H. Moor View All Authors

69 Paper Citations | 3005 Full Text Views

Abstract

Document Sections

1. Varieties of Machine Ethics
2. Ethical-Impact Agents
3. Implicit Ethical Agents
4. Explicit Ethical Agents
5. Full Ethical Agents

Authors

Abstract:
The question of whether machine ethics exists or might exist in the future is difficult to answer if we can't agree on what counts as machine ethics. Some might argue that machine ethics obviously exists because humans are machines and humans have ethics. Others could argue that machine ethics obviously doesn't exist because ethics is simply emotional expression and machines can't have emotions. A wide range of positions on machine ethics are possible, and a discussion of the issue could rapidly propel us into deep and unsettled philosophical issues. Perhaps, understandably, few in the scientific arena pursue the issue of machine ethics. As we expand computers' decision-making roles in practical matters, such as computers driving cars, ethical considerations are inevitable. Computer scientists and engineers must examine the possibilities for machine ethics because, knowingly or not, they've already engaged in some form of it. Before we can discuss possible implementations of machine ethics, however, we need to be clear about what we're asserting or denying

Published in: IEEE Intelligent Systems (Volume: 21 , Issue: 4 , July-Aug. 2006)

Page(s): 18 - 21

INSPEC Accession Number: 9065956

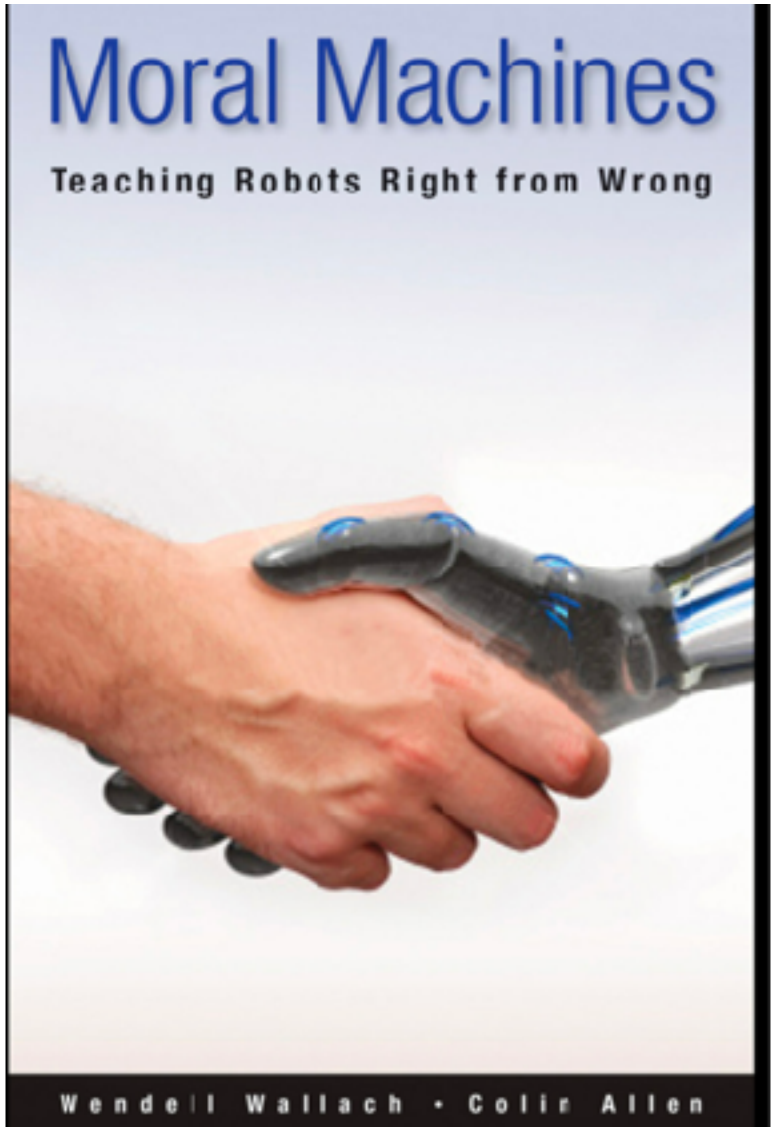
Date of Publication: 07 August 2006

DOI: 10.1109/MIS.2006.80

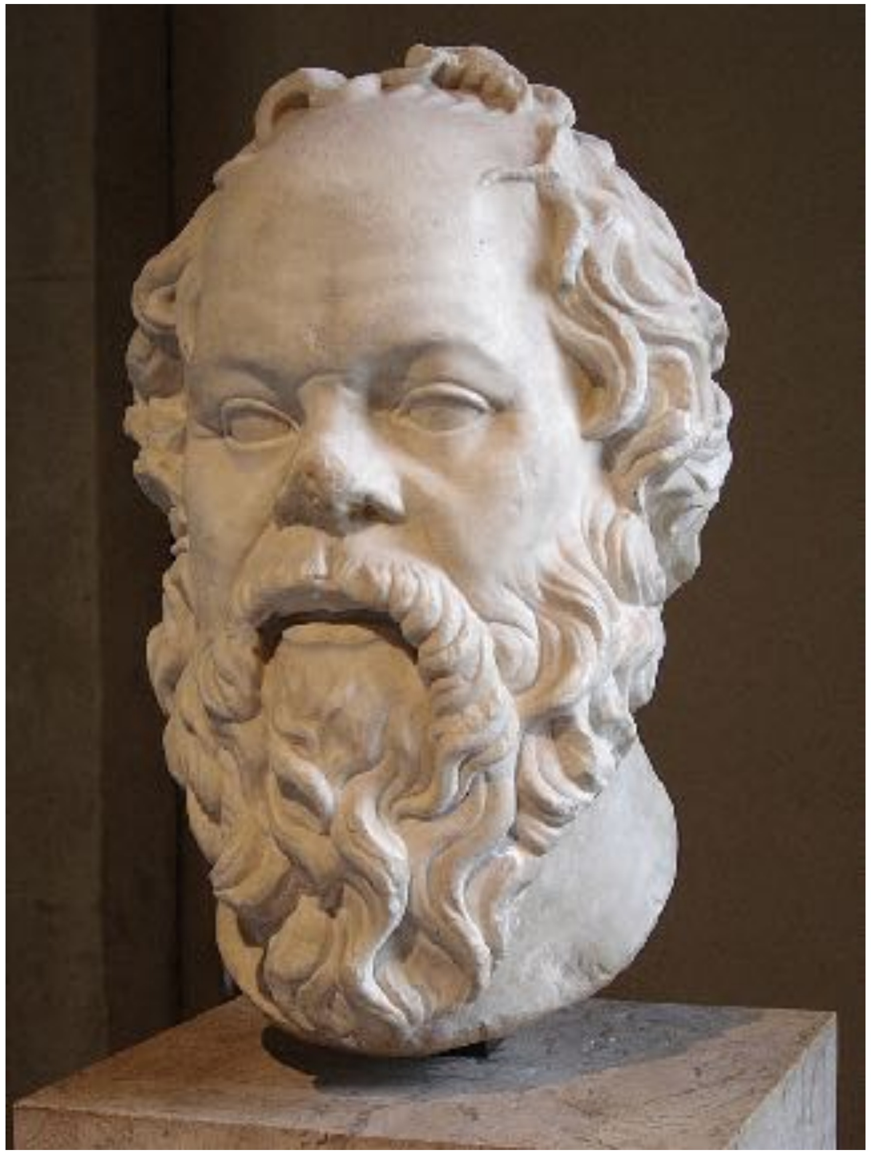
MANCHESTER
1824
The University of Manchester

Top-Down vs. Bottom-Up (in Machine Ethics)

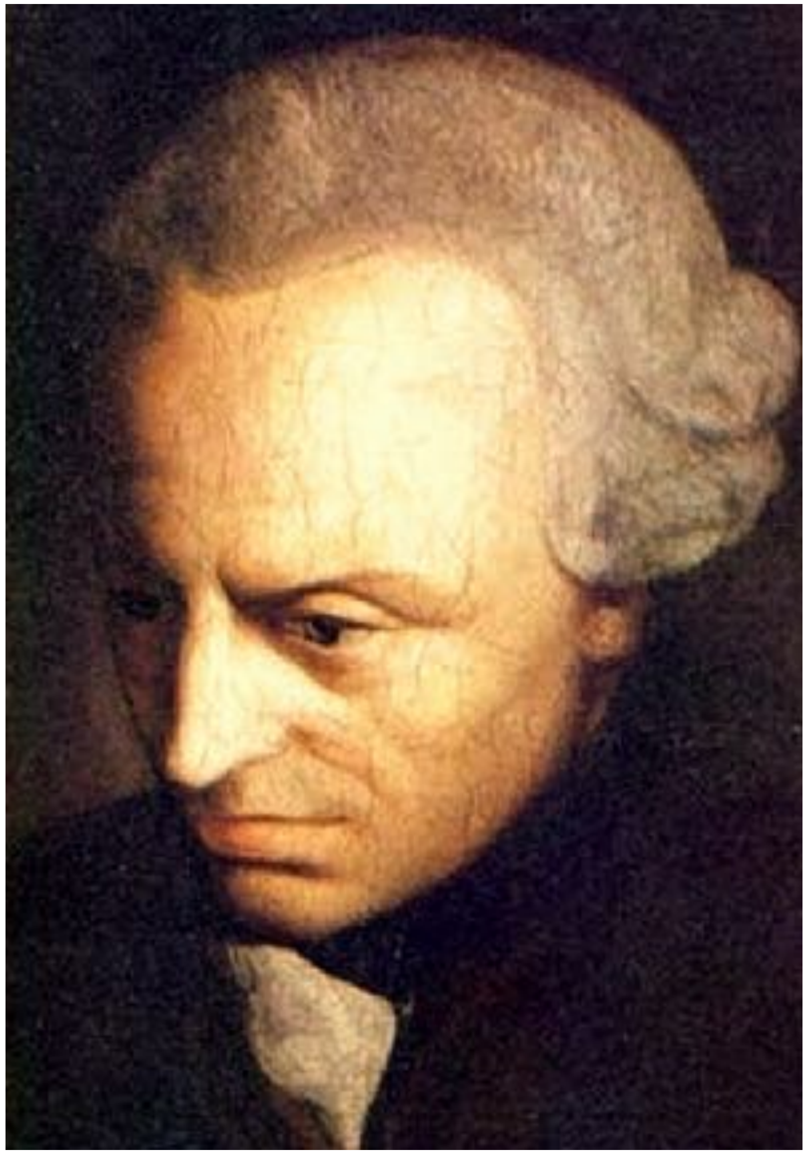
- Top Down: given an ethical theory, how can we implement it?
- Bottom Up: learning ethical behaviour from data.


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

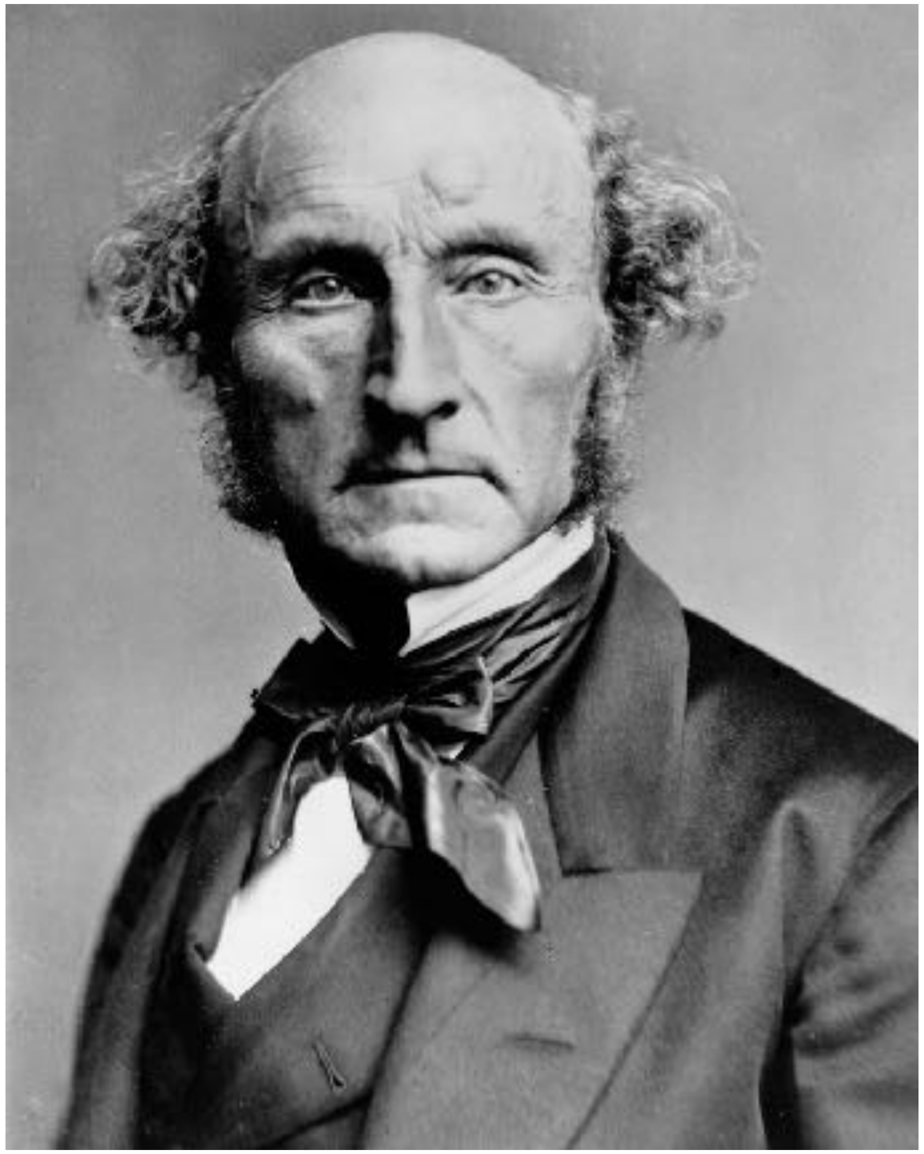
There are a lot of Systems of Ethical Reasoning…



Socrates
Photo Credit: Eric Gaba



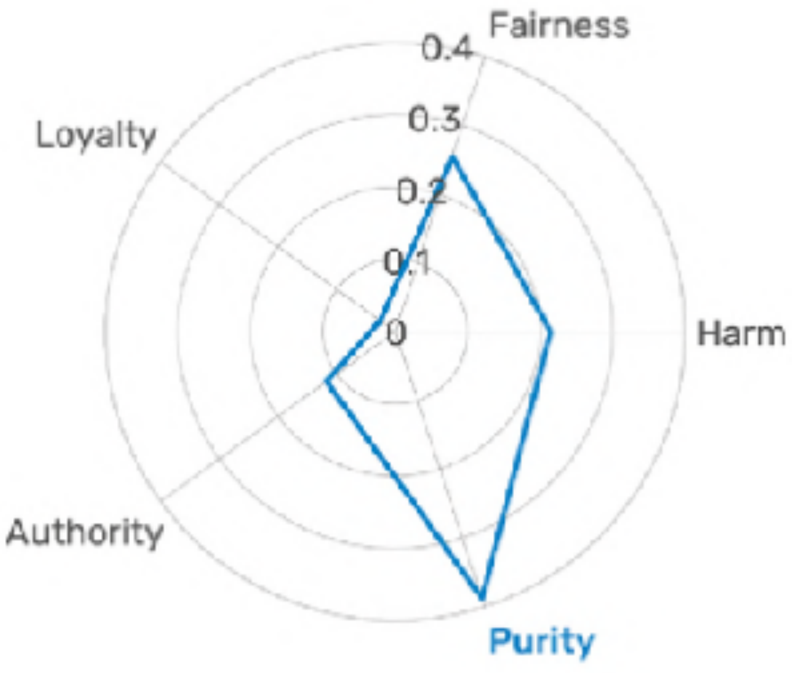
Emmanuel Kant
Unknown Painter
Public Domain



John Stuart Mill
London Stereoscopic Society
Public Domain

Values

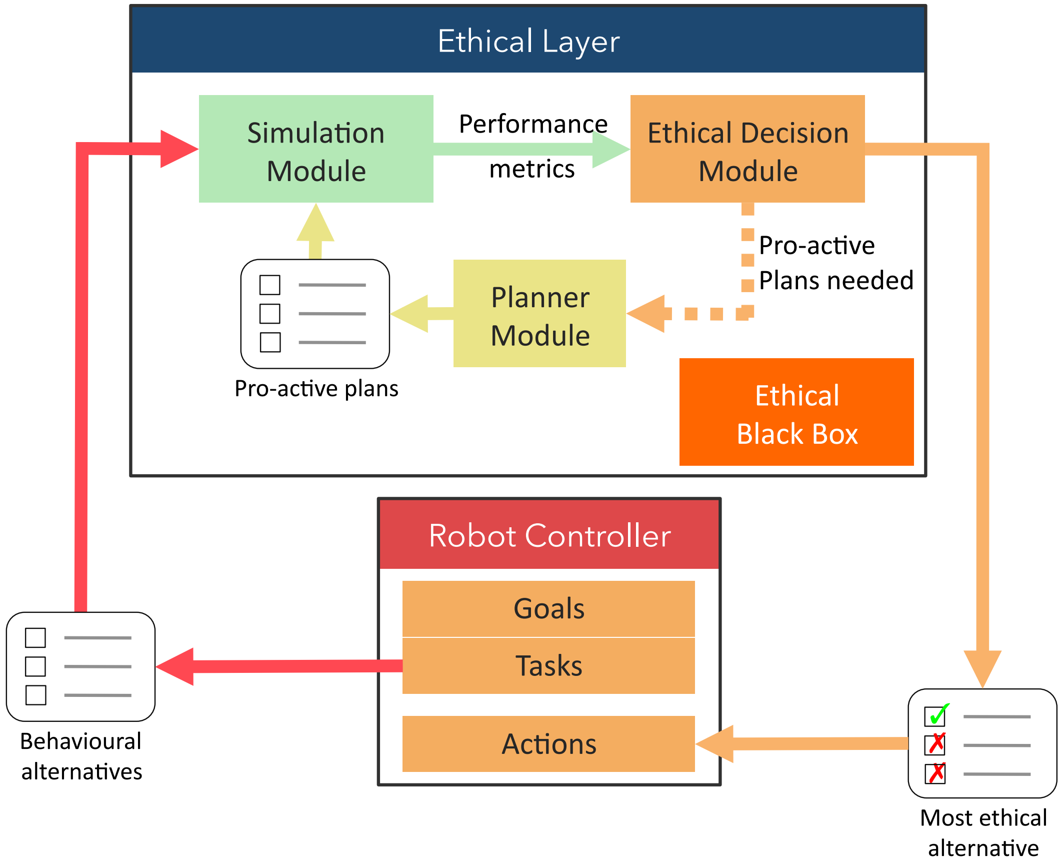
In practice I'm seeing a lot of systems that take *values* (principles/duties) as a starting point and attempt to evaluate actions/outcomes in terms of those values and them somehow rank/prioritise the values.



Hidalgo, C. A., Orghian, D., Canals, J. A., De Almeida, F., and Martín, N. (2021). *How humans judge machines*. MIT Press.

Is everything ethics?

- *Constraint/Governor-Based Ethical Systems* assume that not all system reasoning directly involves ethics. Therefore ethics is placed in some sub-system that guides or constrains the actions of the rest of the system.
- *Global Ethical Systems* assume that ethical reasoning is involved in all system reasoning - that, in fact, all decisions are ethical decision.



Paul Bremner, Louise A. Dennis, Michael Fisher and Alan F. Winfield. On Proactive, Transparent and Verifiable Ethical Reasoning for Robots. *Proceedings of the IEEE. Special Issue on Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems*. 107(3), pp:541-561.

Our Approach

- We’ve taken an explicitly ethical top-down approach, implementing a variety of ethical theories in a variety of applications.
- We’ve looked at both constraint-based and global systems.
- A driver behind our approach has been verifiability and assurance.

Ethical Reasoning as a Fall Back

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.

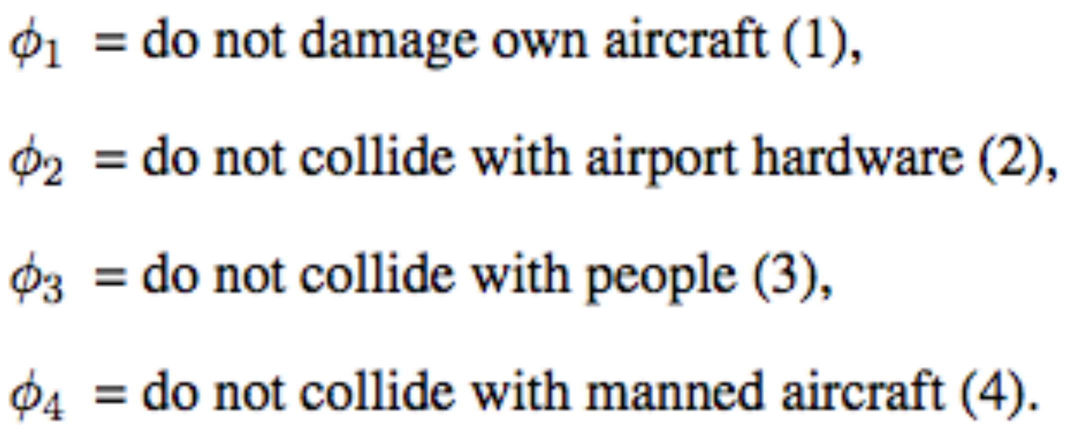
Extension of work on implementing the rules of the air done by Fisher and Webster in conjunction with Daresbury Labs

Implementation of Prima Facie Duties

- We have a set of ethical concerns which we rank: killing is worse than stealing is worse than lying.
- A plan, P1, is worse than another, P2, if
  - P1 violates an ethical concern and P2 doesn't
  - The worst concern violated by P2 and not by P1 is less serious than the worst concern violated by P1 and not P2
  - The worst concerns are equally bad, but P1 violates more concerns than P2 does

A Scenario

- Turn Left (damages the aircraft and airport hardware)
- Turn Right (damage the aircraft and risks colliding with people)
- Continue (risks collision with a manned aircraft)



ϕ_1 = do not damage own aircraft (1),

ϕ_2 = do not collide with airport hardware (2),

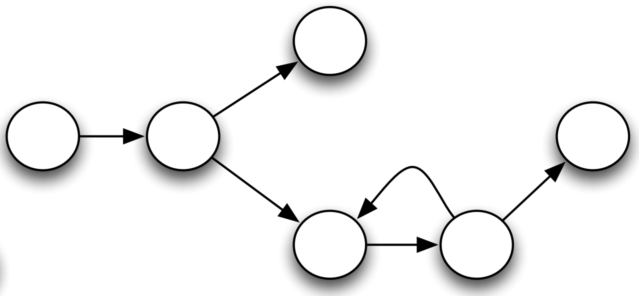
ϕ_3 = do not collide with people (3),

ϕ_4 = do not collide with manned aircraft (4).

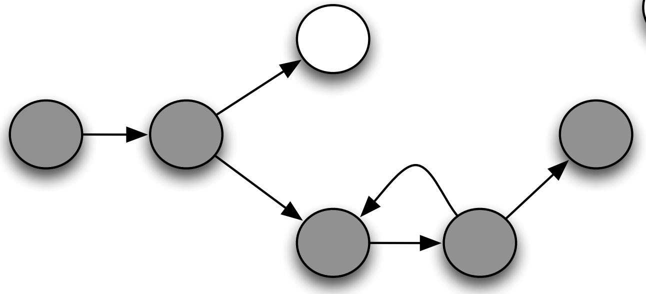
The Aircraft Turns Left

A Diversion: What is Verification?

Program

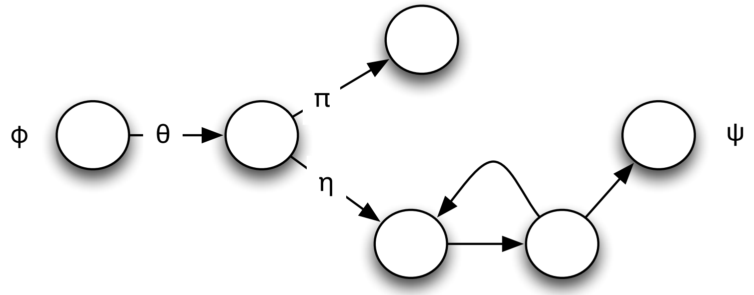


Testing



Is ϕ true at the end?

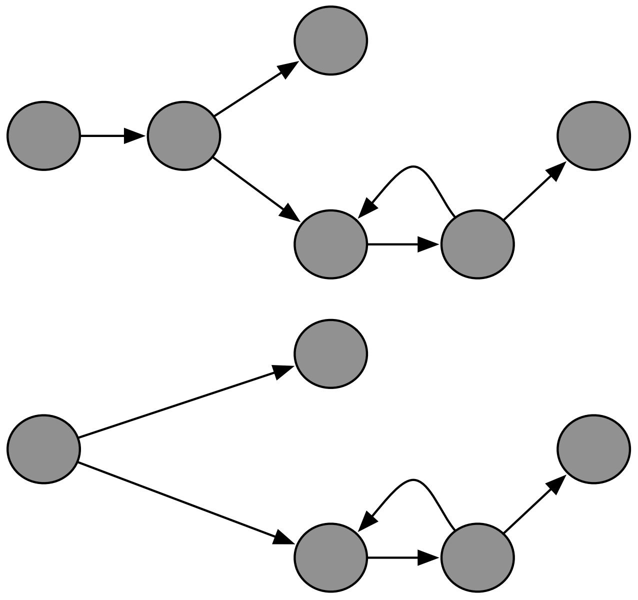
Theorem Proving



If ϕ is true at the start then ψ is true at the end

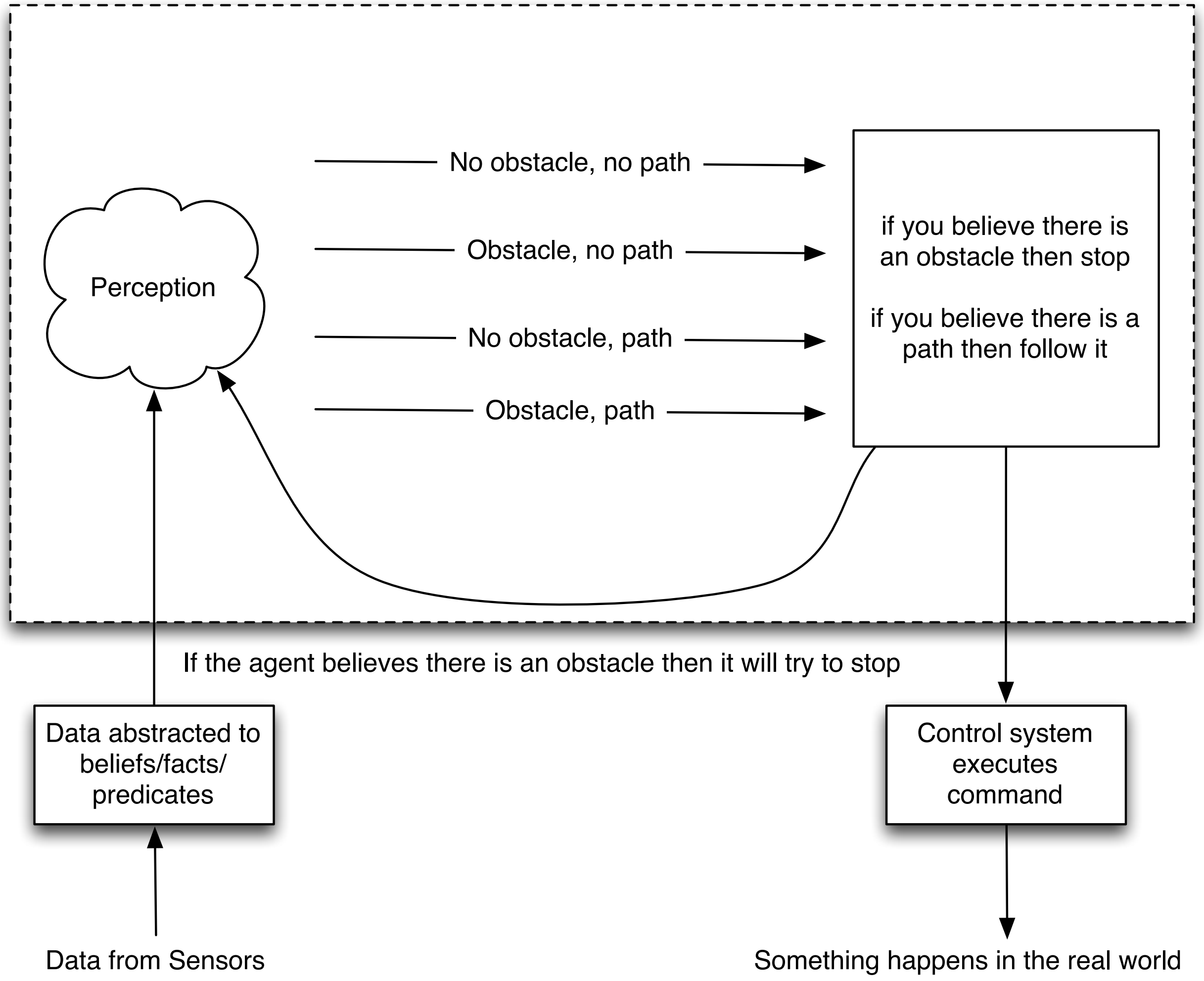
Model Checking

Program Model Checking



If ϕ is true then eventually ψ is true

Model-Checking Autonomous Systems


No obstacle, no path
Perception
Obstacle, no path
No obstacle, path
Obstacle, path
if you believe there is an obstacle then stop
if you believe there is a path then follow it
If the agent believes there is an obstacle then it will try to stop
Data abstracted to beliefs/facts/ predicates
Control system executes command
Data from Sensors
Something happens in the real world

Consider outputs of decision maker given all possible inputs

Verifying the Aircraft Example: How did we branch the search space?

- Anonymous plans but explored all combinations of violated concerns. Checked that the aircraft always selected least unethical choice.
- Fixed set of plans with fixed consequences (e.g., landing on a road will damage infrastructure) but varied which plans were available. Checked that the aircraft only landed on a road if no field were available to land in.
- Fixed set of plans and consequences but varied whether they succeeded. Checked the aircraft always selected least unethical choice.

Machine Ethics: What do we want to prove?

- Well, obviously we want to prove that the system always "Does the right thing"

- Most of these systems have a set of rules or utilities (an *ethical encoding*) and a decision mechanism. In theory "stakeholders" can sign off the encoding (the rules, or the utilities) that they capture the stakeholder's values.

- So what is there to prove?

The Smart Home that would not evacuate

- Utilities:
  - lights_on = -1,
  - people_leave_house = -1,
  - people_are_safe = 10
  - people_can_see = 0, 2 (depending on context)

- Mechanisms:
  - $turn\_lights\_on \rightarrow lights\_on$
  - $lights\_on \vee daylight \rightarrow people\_can\_see$
  - $evacuation\_attempt \wedge people\_can\_see \rightarrow people\_leave\_house$
  - $people\_leave\_house \vee \neg danger\_in\_house \rightarrow people\_are\_safe$
  - $fire \rightarrow danger\_in\_house$

- Principle of Double Effect: net balance of consequences of an an action must be positive and no negative consequences can be intended.

Louise. A. Dennis, Martin Mose Bentzen, Felix Lindner and Michael Fisher. Verifiable Machine Ethics in Changing Contexts. In: 35th AAAI Conference on Artificial Intelligence (AAAI 2021).

Properties for Ethical Reasoning Systems

- Check underlying decision making implementation is correct.
  - Broadly speaking we want to prove that the "least worst" option according to the theory is always the one chosen. In some theories this is easier to specify than in others.
- Sanity Checking properties.
  - Overriding safety concerns
  - Legal constraints
- Scenario probing
  - Explore specific case studies and settings to check that the "correct" choice is made in those case studies and settings.

Open Questions

- Practicality: Both of reasoning and knowledge engineering.
- Identifying the Stakeholders.
- Reasoning over sequences of actions, multiple agents (causality).
- Moral Uncertainty (Resolving pathological edge cases).
- Situational Awareness — getting the information necessary to start ethical reasoning.
- Benchmarking.

Thank You

Other Work

- **Probabilistic model checking used to assess risk of violations:** Dennis et al. Towards Verifiably Ethical Robot Behaviour. Proceedings of the AAAI Workshop on Artificial Intelligence and Ethics (1st International Workshop on AI and Ethics).
- **Framework for multiple ``Evidential Reasoners":** Cardoso et al. Implementing Ethical Governors in BDI - EMAS 2021
- **Defeasible Logic as a way to simplify Ethical "Rules":** Dennis and Perea del Olmo. A Defeasible Logic Implementation of Ethical Reasoning - CME 2021
- **Approaches to Benchmarking:** Bjørgen et al. Cake, death, and trolleys: dilemmas as benchmarks of ethical decision-making. AAAI/ACM Conference on Artificial Intelligence, Ethics and Society 2018
- **Multi-Principle Approach which incorporates Uncertainty:** Simon Kolker et al. Uncertain Machine Ethical Decisions using Hypothetical Retrospection. COINE 2023.

MANCHESTER
1824
The University of Manchester

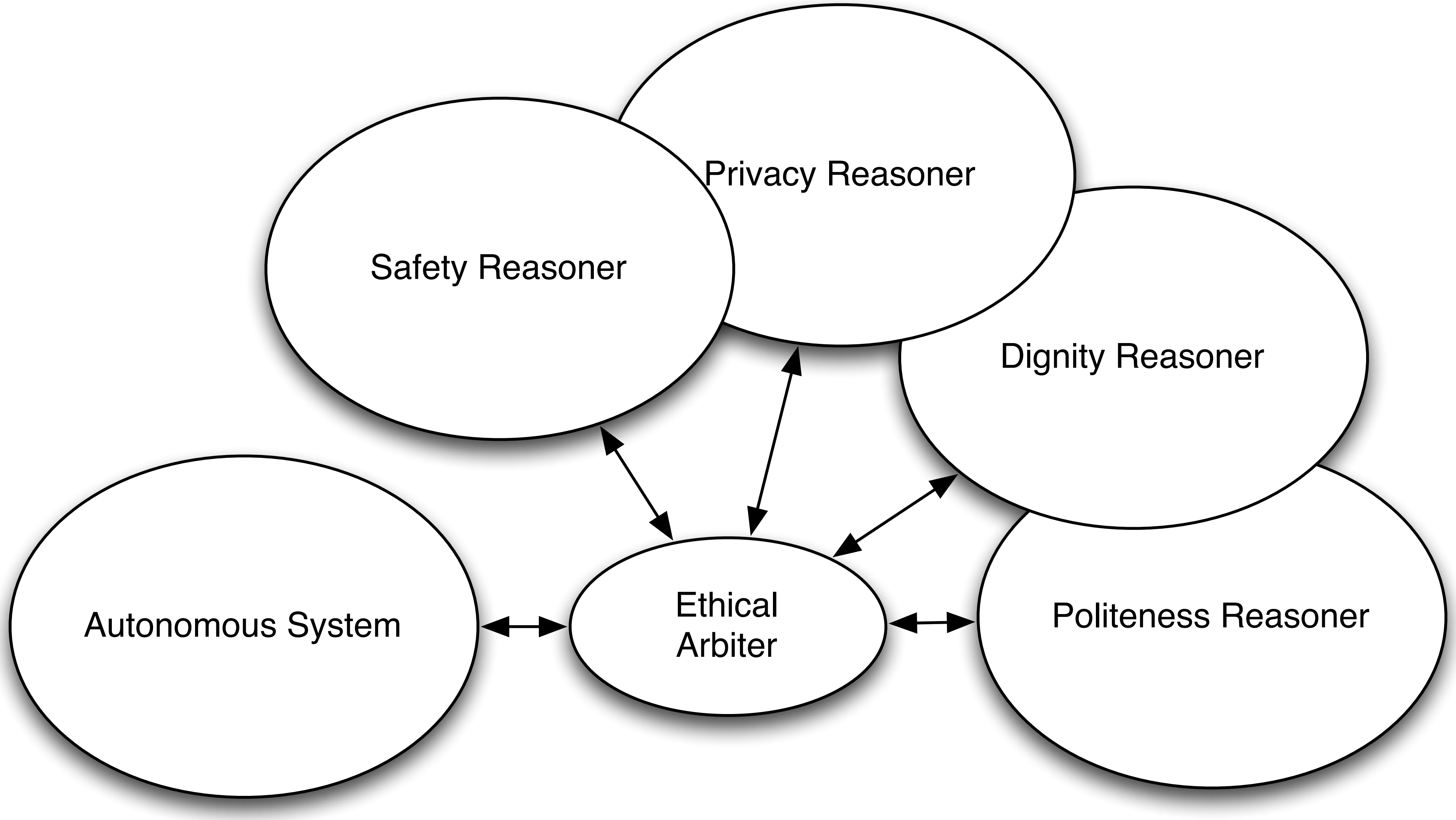
Thank You

Looking Forward

- Ordinary people don’t use philosophical ethical frameworks (much) and nevertheless function as moral agents. Are philosophical frameworks the correct approach for practical ethical reasoning? We hope to explore the concept of responsibilities as an alternative.
- How does reasoning about risk and uncertainty interact with all these approaches?

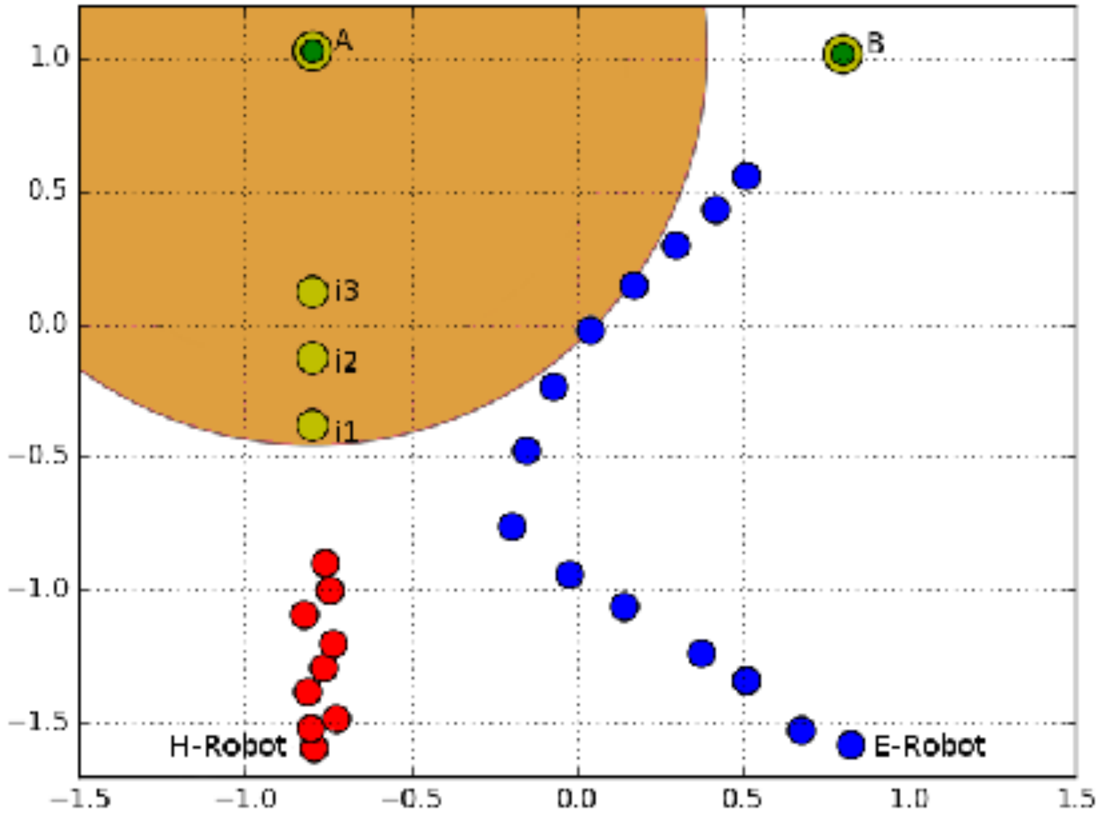
Other Work

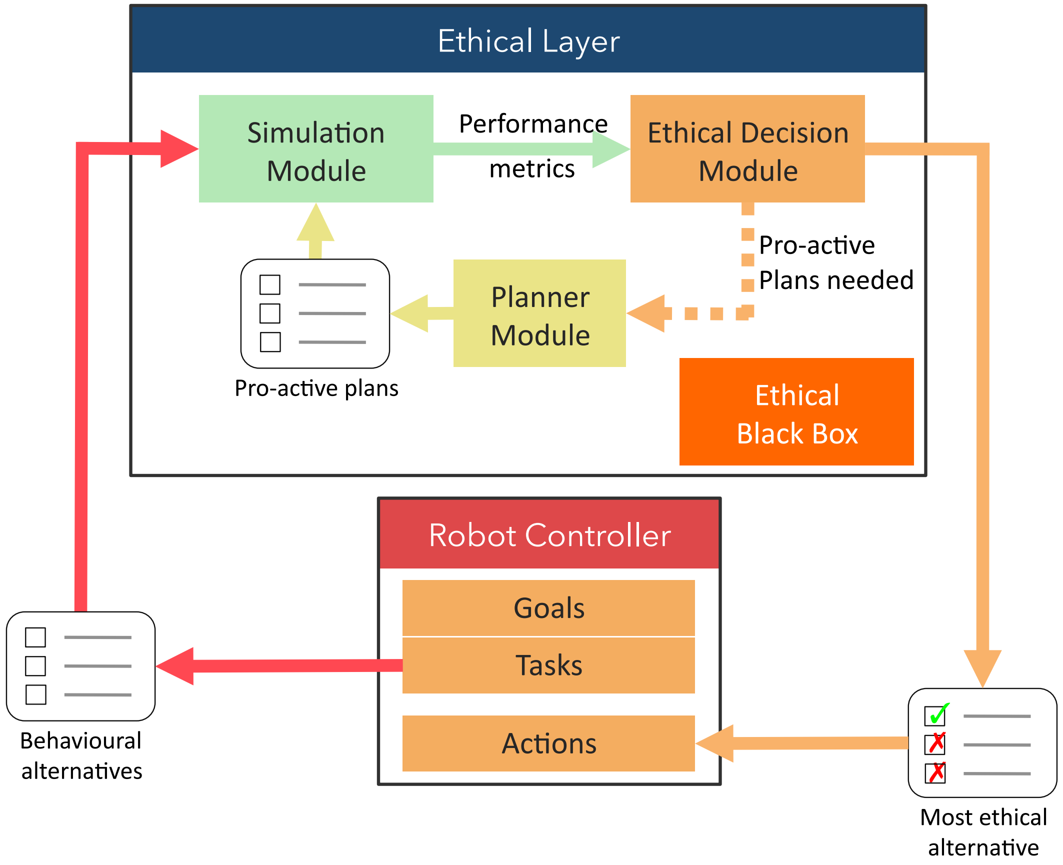
Cardoso et al. 2021. Implementing Ethical Governors in BDI - EMAS 2021



Linking verified version to the actual robot.

Paul Bremner, Louise A. Dennis, Michael Fisher and Alan F. Winfield. On Proactive, Transparent and Verifiable Ethical Reasoning for Robots. *Proceedings of the IEEE. Special Issue on Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems*. 107(3), pp:541-561. DOI: 10.1109/JPROC.2019.2898267

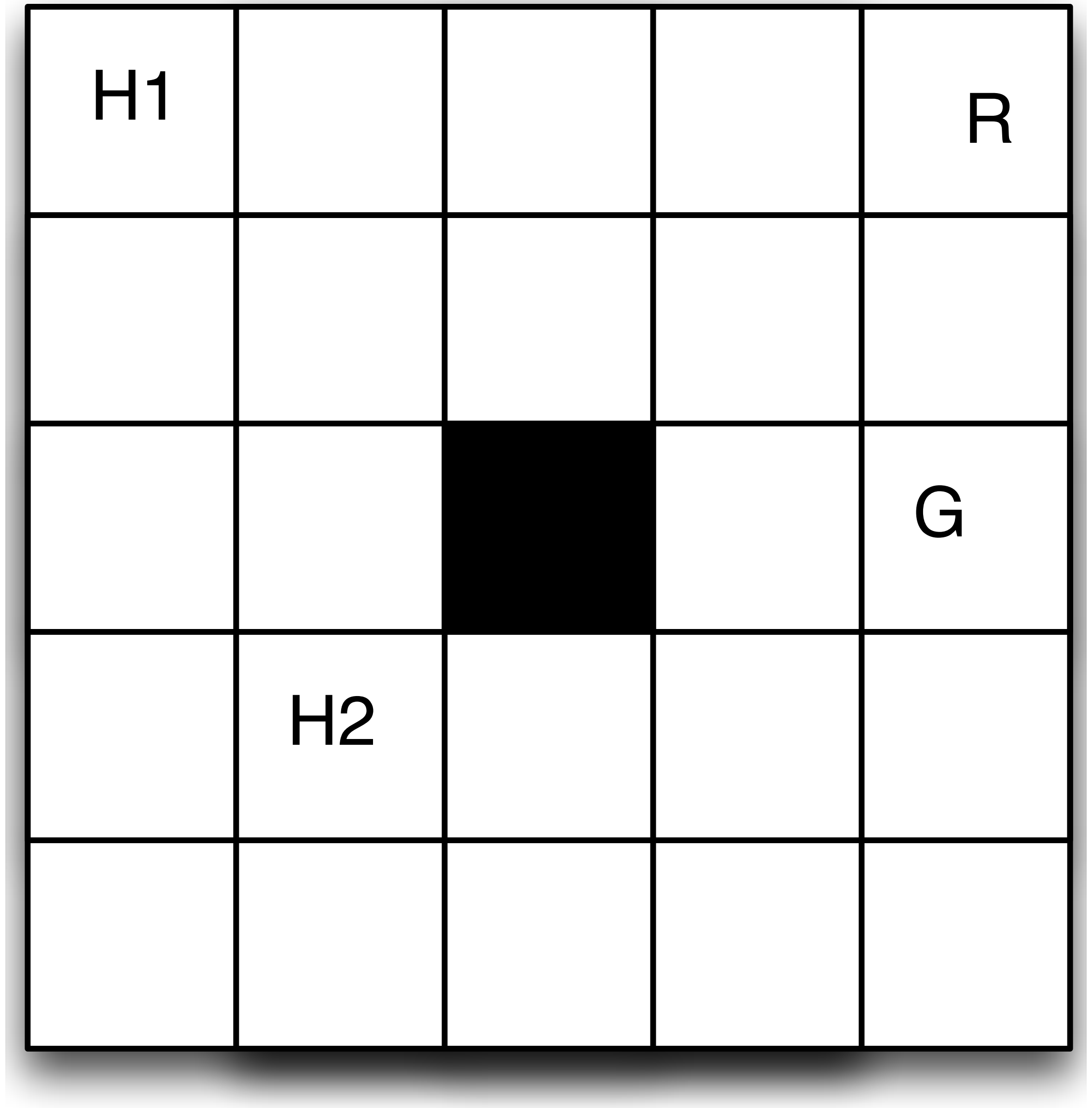




Scenario Probing can also allow some forms of risk evaluation

Louise A. Dennis, Michael Fisher, and Alan Winfield. Towards Verifiably Ethical Robot Behaviour. *Proceedings of the AAAI Workshop on Artificial Intelligence and Ethics (1st International Workshop on AI and Ethics)*.

- If the robot can always find a safe path to the human when it believes the human is in danger, then the human doesn't fall in the hole.
- Also used PRISM to calculated the probability of the human falling in the hole.



H1				R
		■		G
	H2			