

Machine ethics

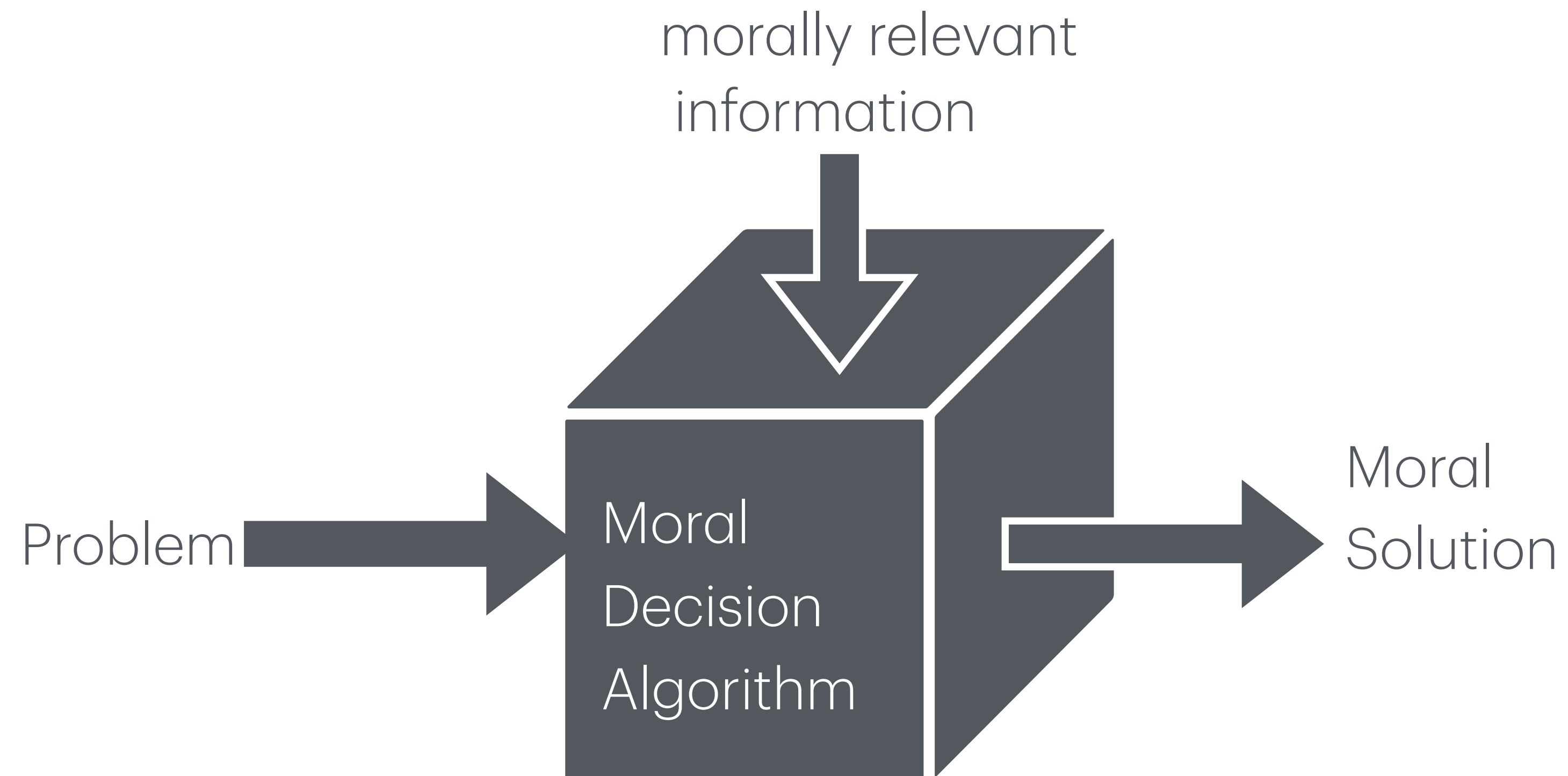
an introduction



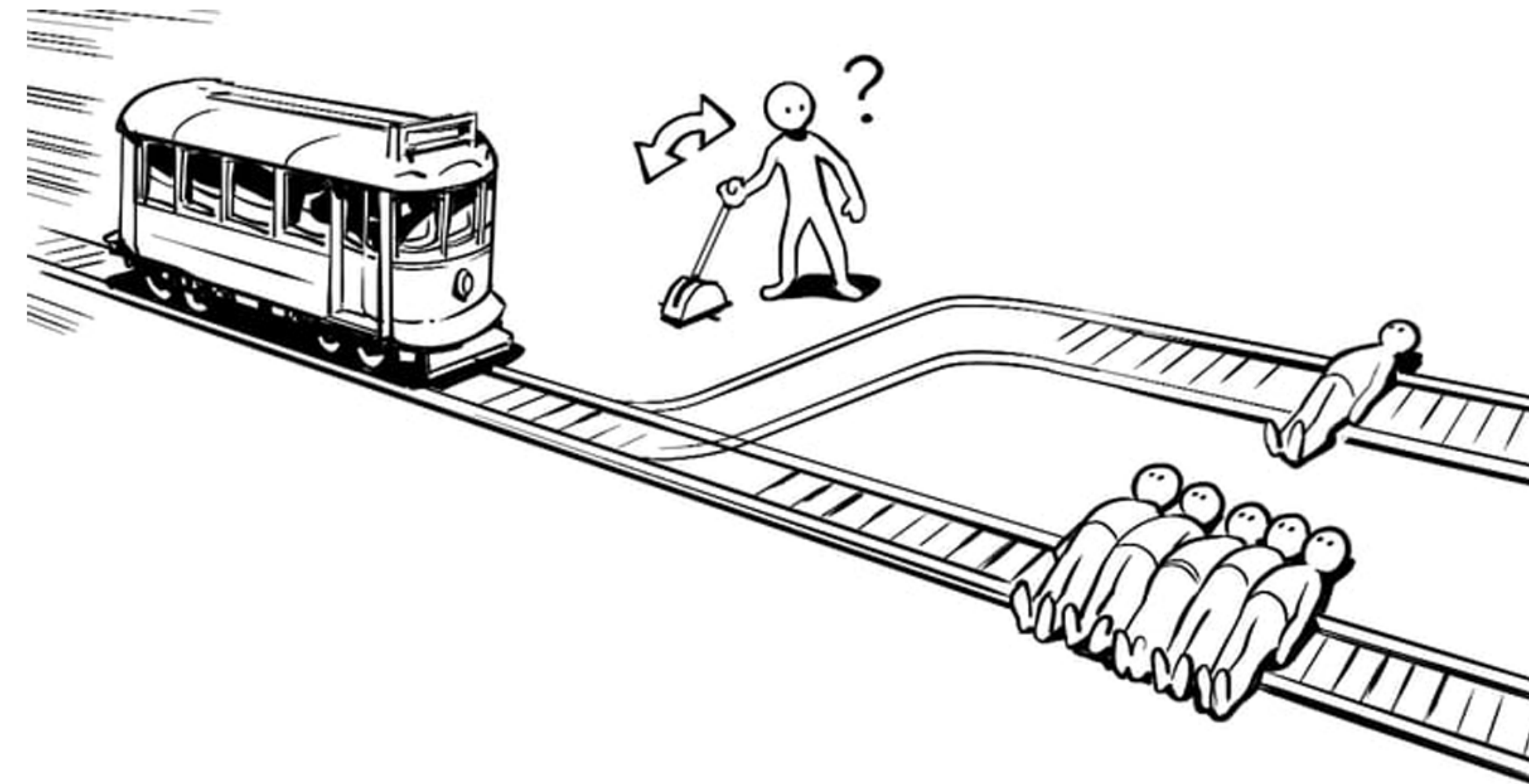
LEVERHULME
TRUST

Marija Slavkovik

The machine ethics problem



Why machine ethics?



Why machine ethics?

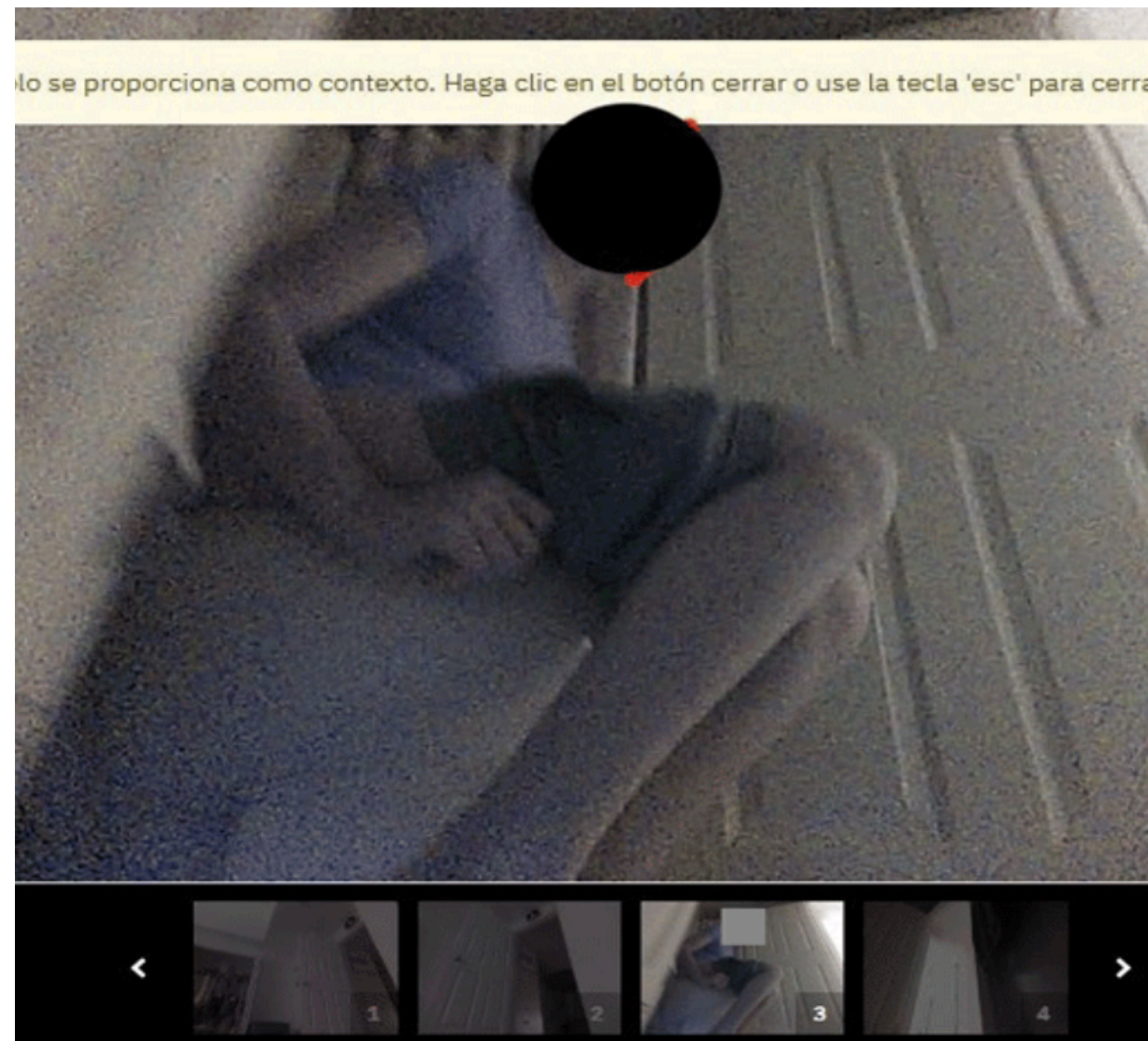


Image captured by iRobot development devices, being annotated by data labels. Faces, where visible, have been obscured with a gray box by MIT Technology

Fury over Facebook 'Napalm girl' censorship

🕒 9 September 2016 · 💬 490 Comments



| Espen Egil Hansen is editor of Aftenposten, Norway's largest newspaper.

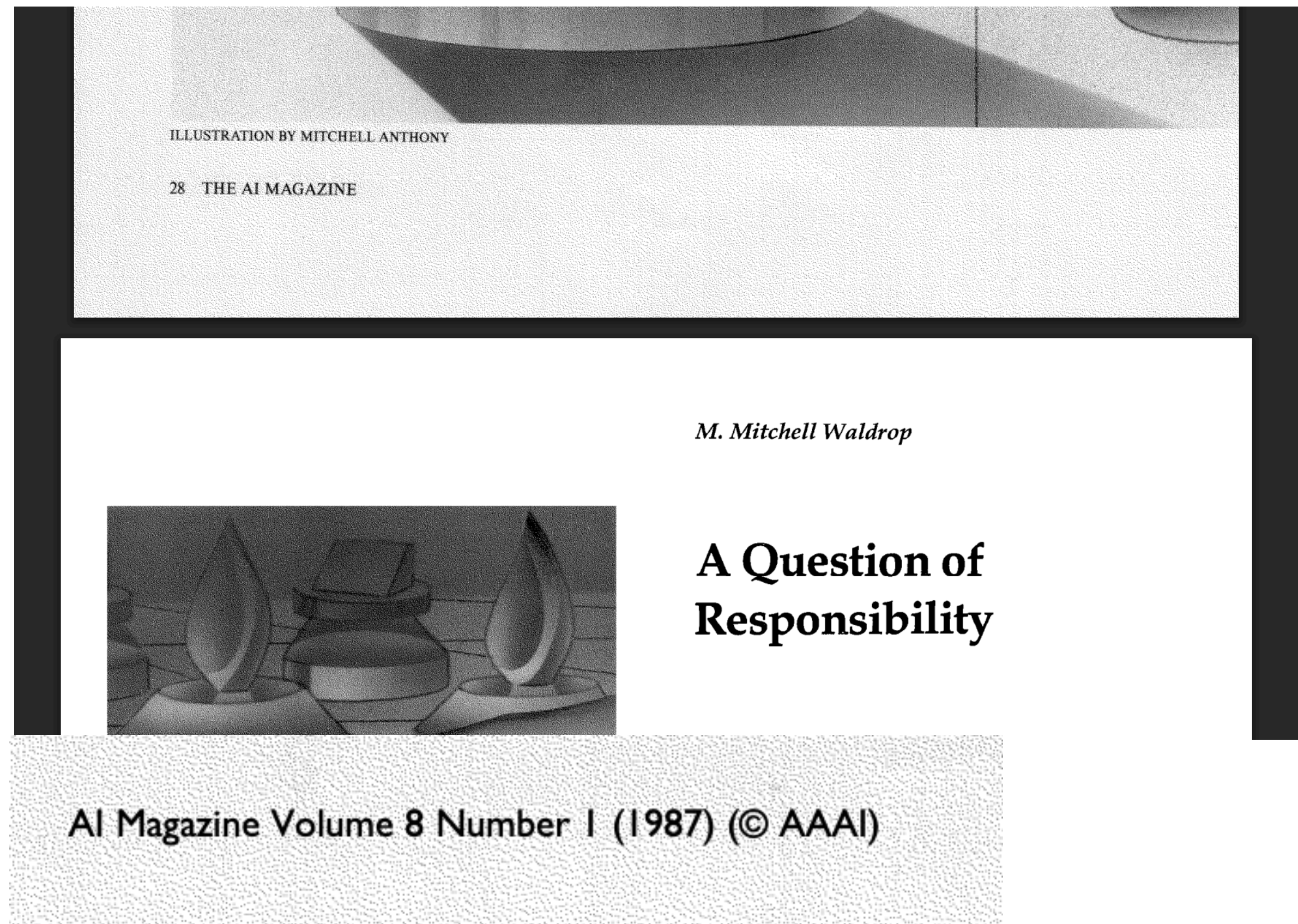
Zoe Kleinman

Technology reporter, BBC News



Machine ethics

history



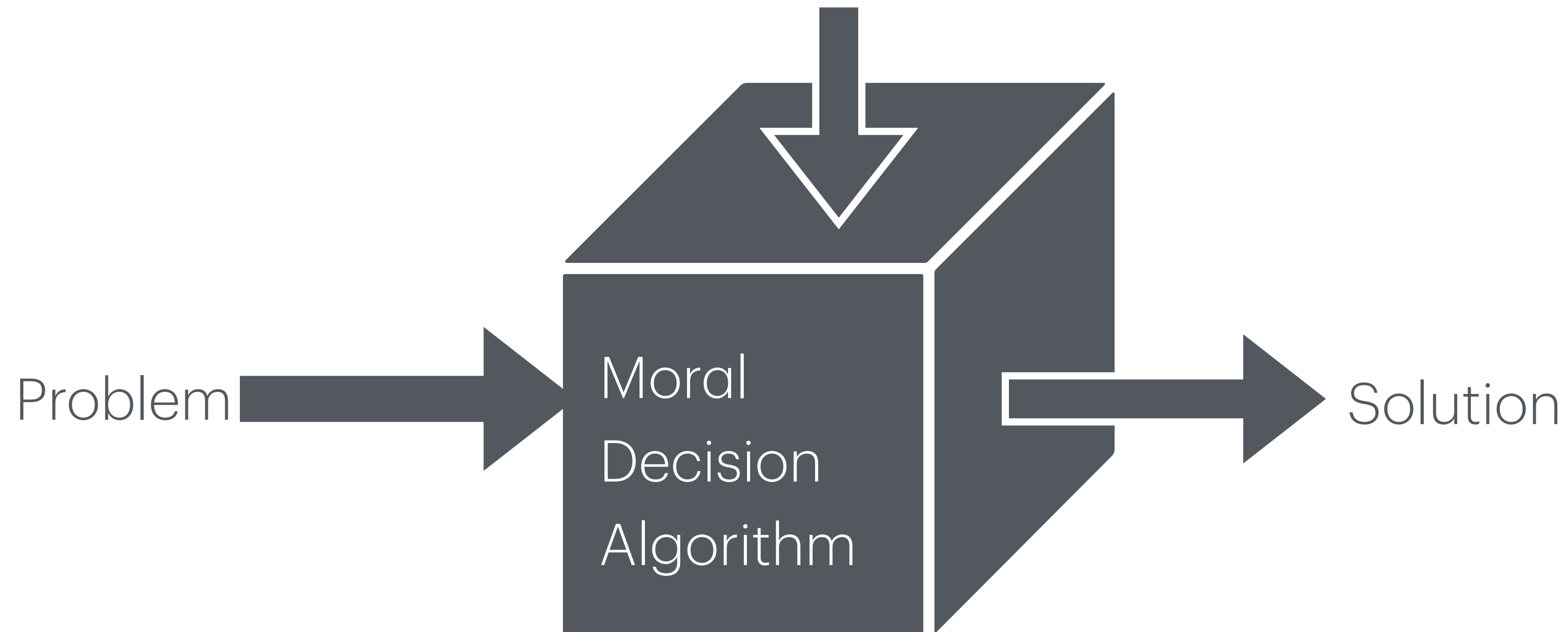
*Perhaps what we need is . . . a theory
and practice of machine ethics . . .*

The machine ethics problem

Who decides?

What standards should artificial moral agents follow?

What is moral?



What does it mean to be an artificial moral agent?

How can we tell that one is/has competence?

Is ethics for machine agents the same as ethics for people agents?

AI Alignment

From Wikipedia, the free encyclopedia

In the field of [artificial intelligence](#) (AI), **AI alignment** research aims to steer AI systems toward a person's or group's intended goals, preferences, and ethical principles. An AI system is considered *aligned* if it advances its intended objectives. A *misaligned* AI system may pursue some objectives, but not the intended ones.^[1]

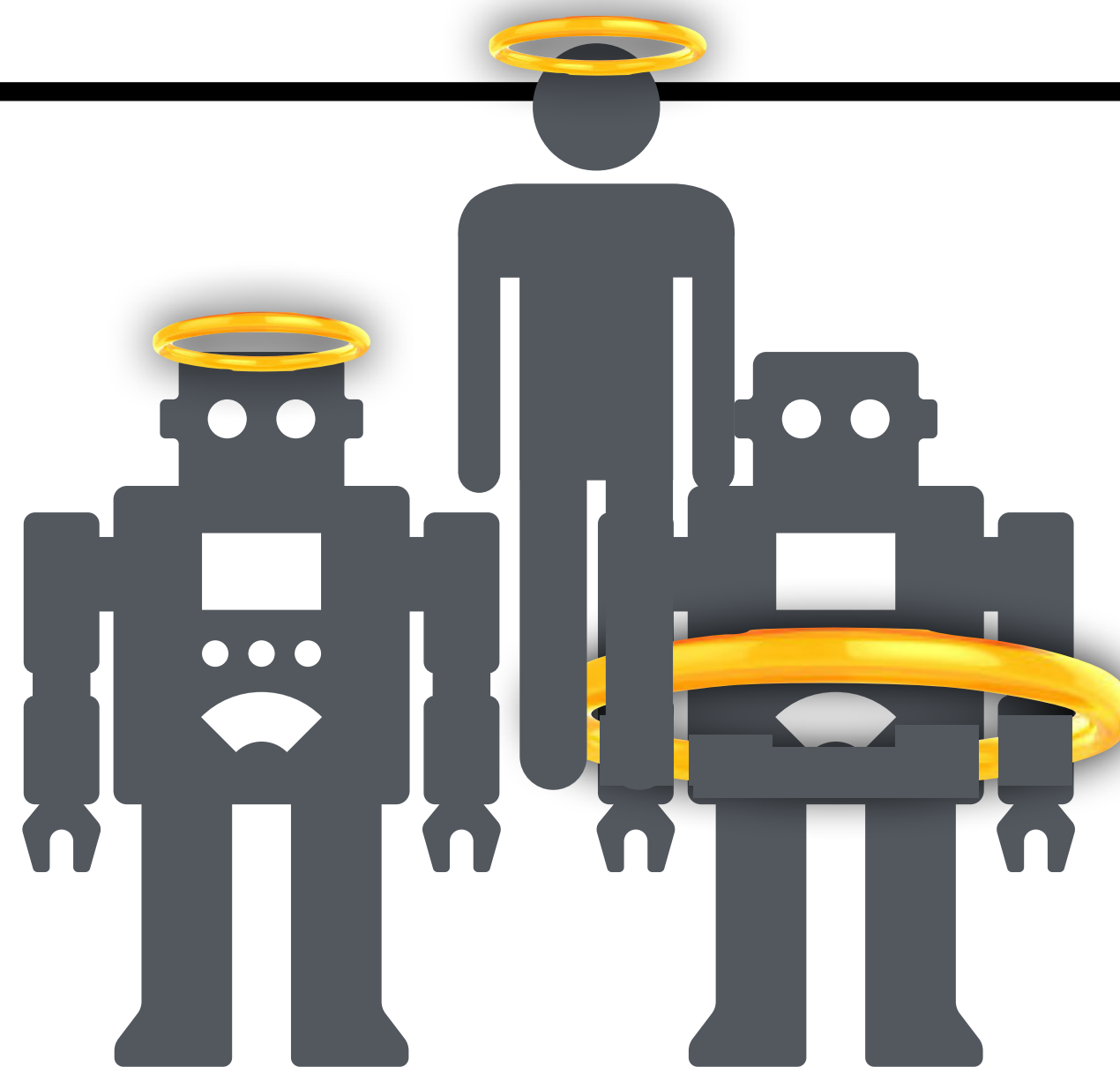
Machine Ethics

Machine ethics is “concerned with the **(ethical)** behaviour of machines towards human users and other machines”

The landscape

AI Ethics

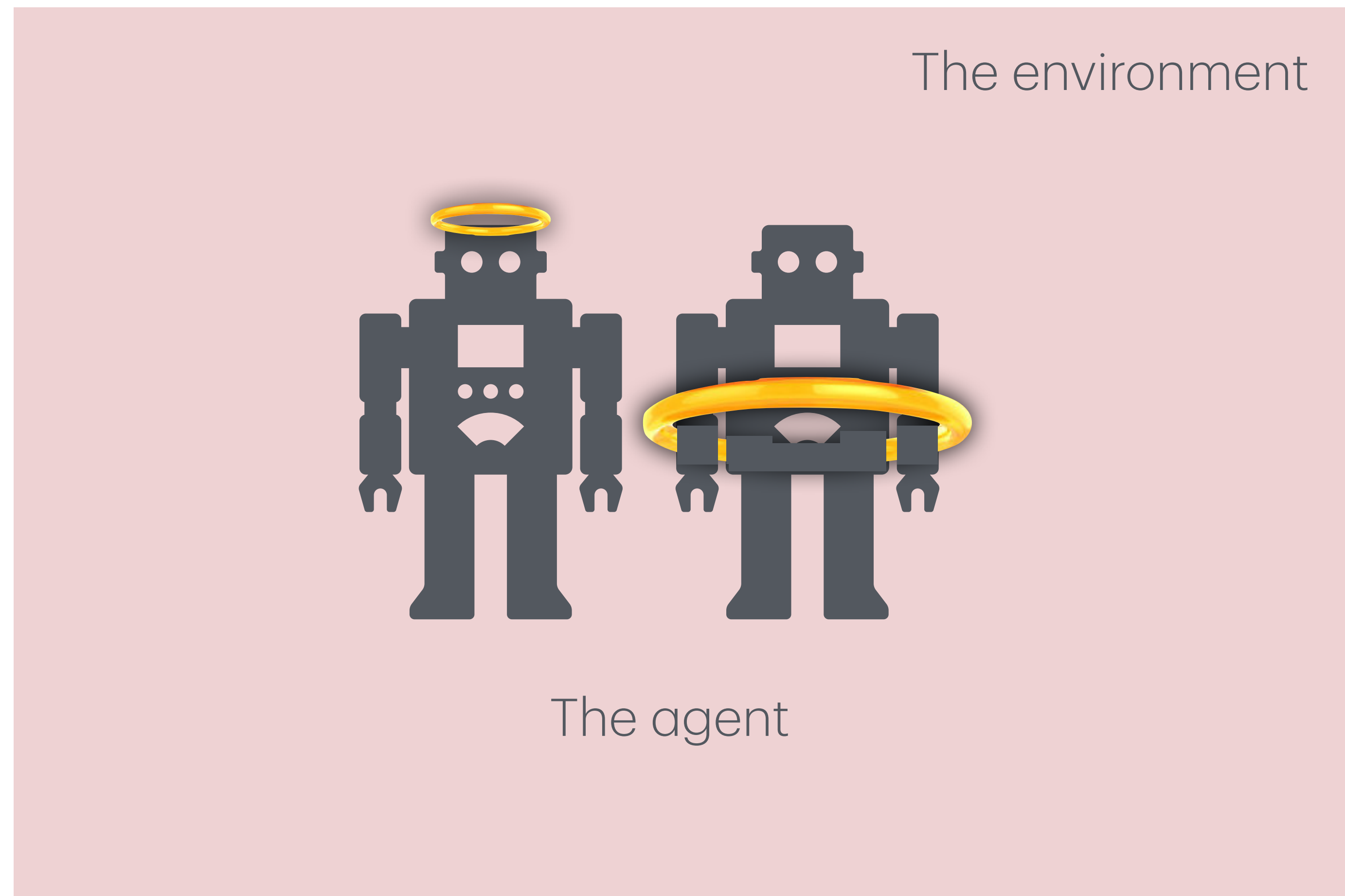
- Accountability
- Explainability
- Transparency
- Fairness
- Bias
- Trustworthiness



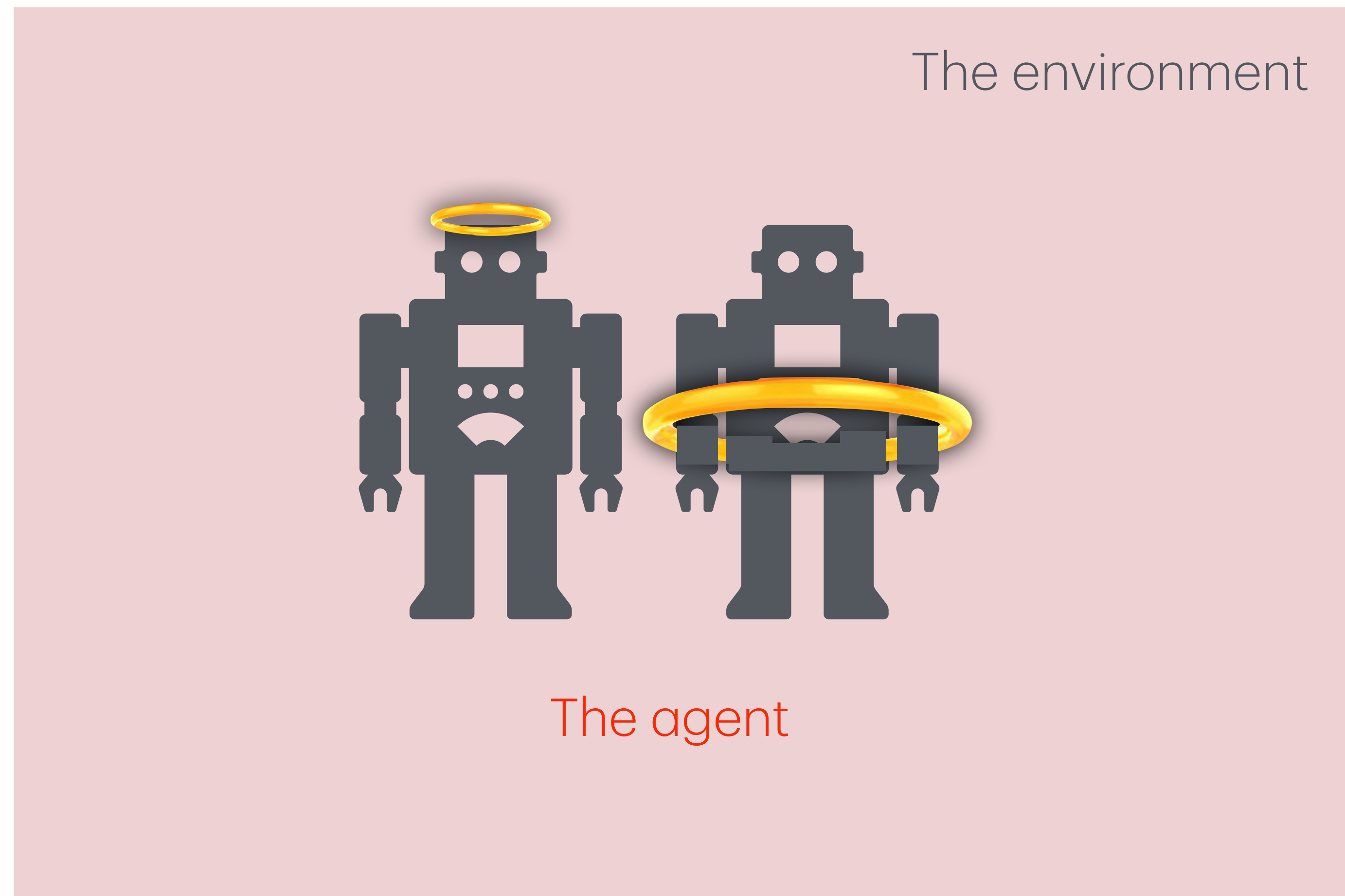
researchers in autonomous agents and multi-agent systems working on the scientific and technological aspects of social coordination, organizational theory, normative MAS, artificial or electronic institutions, norm/policy-aware and ethical agents

Normative reasoning

What can we engineer?



What can we engineer?



What is an artificial moral agent?

(O) An action is said to be morally qualifiable if and only if it can cause moral good or evil. An agent is said to be a moral agent if and only if it is capable of morally qualifiable action.

[Home](#) > [Minds and Machines](#) > Article

On the Morality of Artificial Agents

Published: August 2004

Volume 14, pages 349–379, (2004) [Cite this article](#)

✓ Access provided by Università degli Studi di Bologna – Area Biblioteche e Servizi allo Studio

[Download PDF](#) 

[Luciano Floridi](#) & [J.W. Sanders](#)

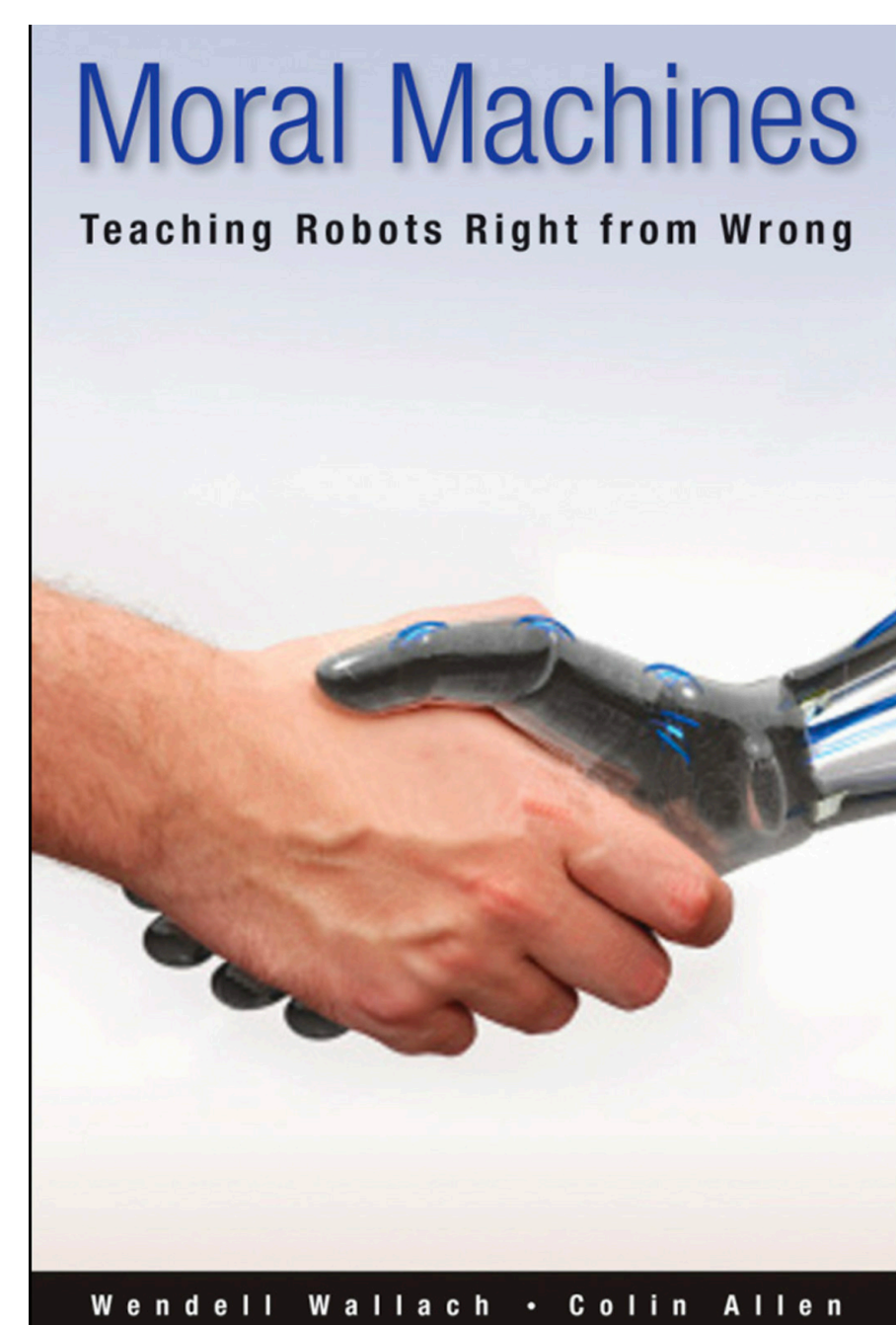
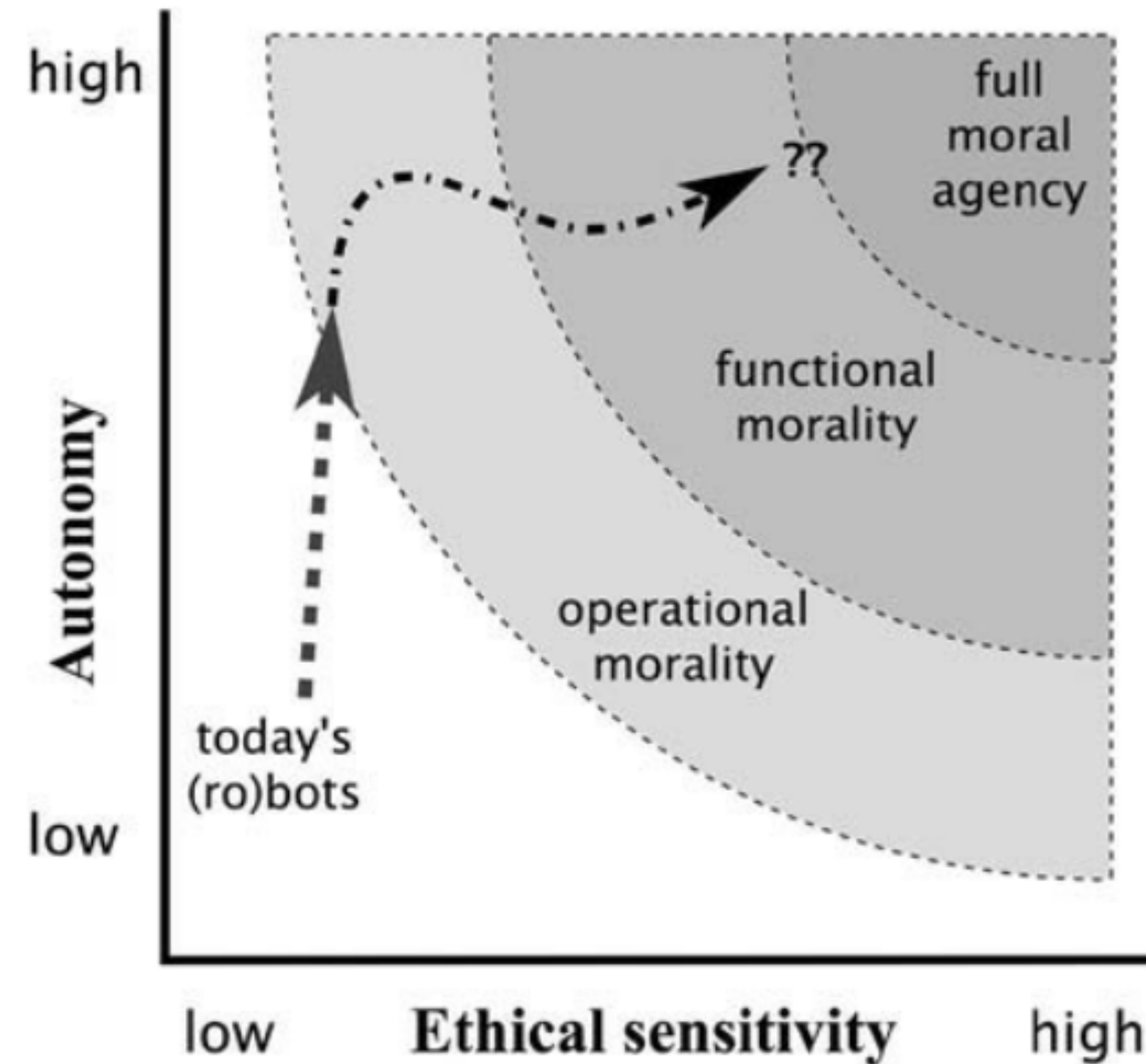
Ethical impact agents

- Agent is not capable of morally qualifiable action
- Agent is an instrument in enabling morally qualifiable actions



Artificial moral agents

- Implicit ethical agents
- Explicit 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](#)

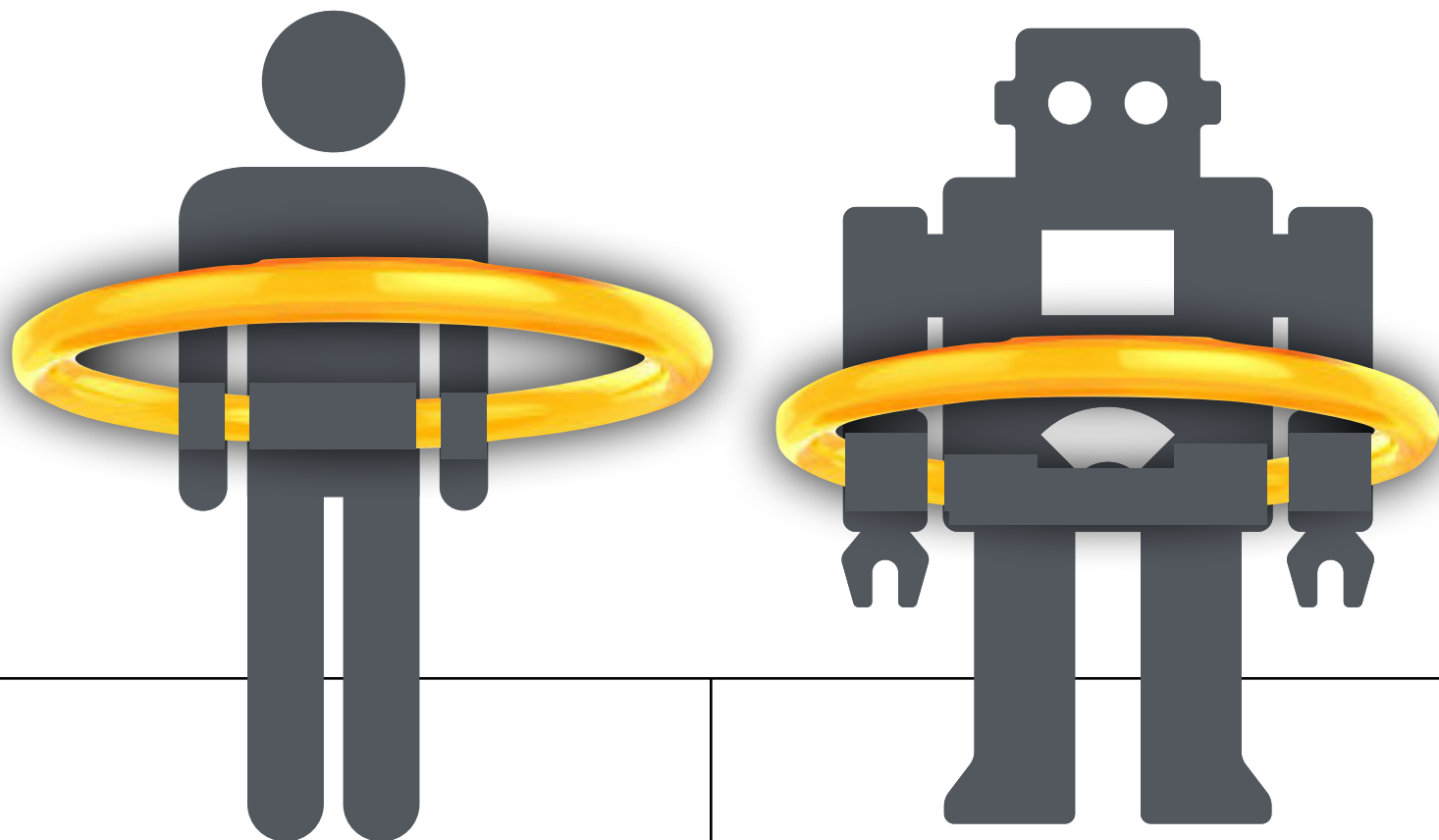
69
Paper
Citations


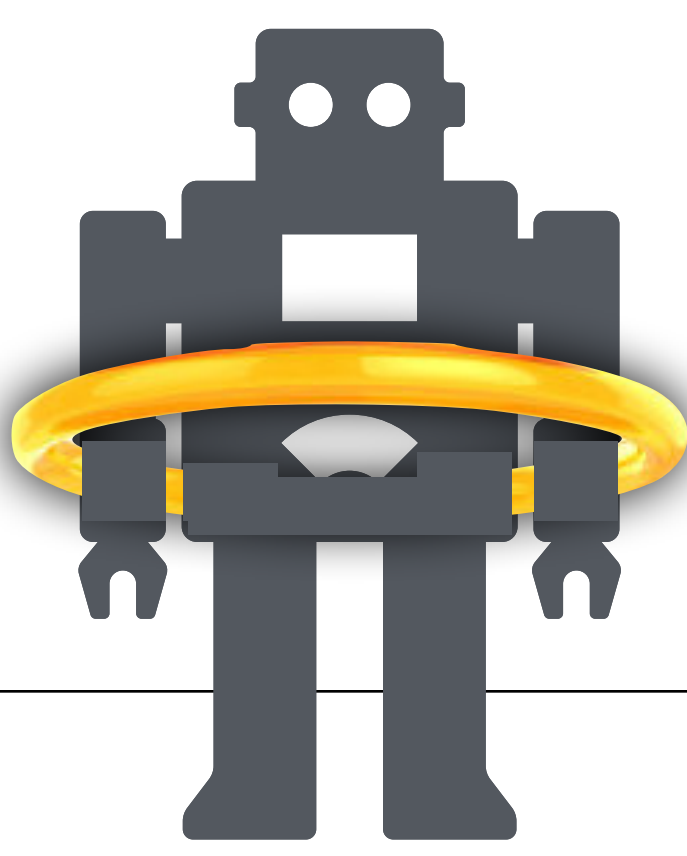
3005
Full
Text Views



The difference between implicit and explicit

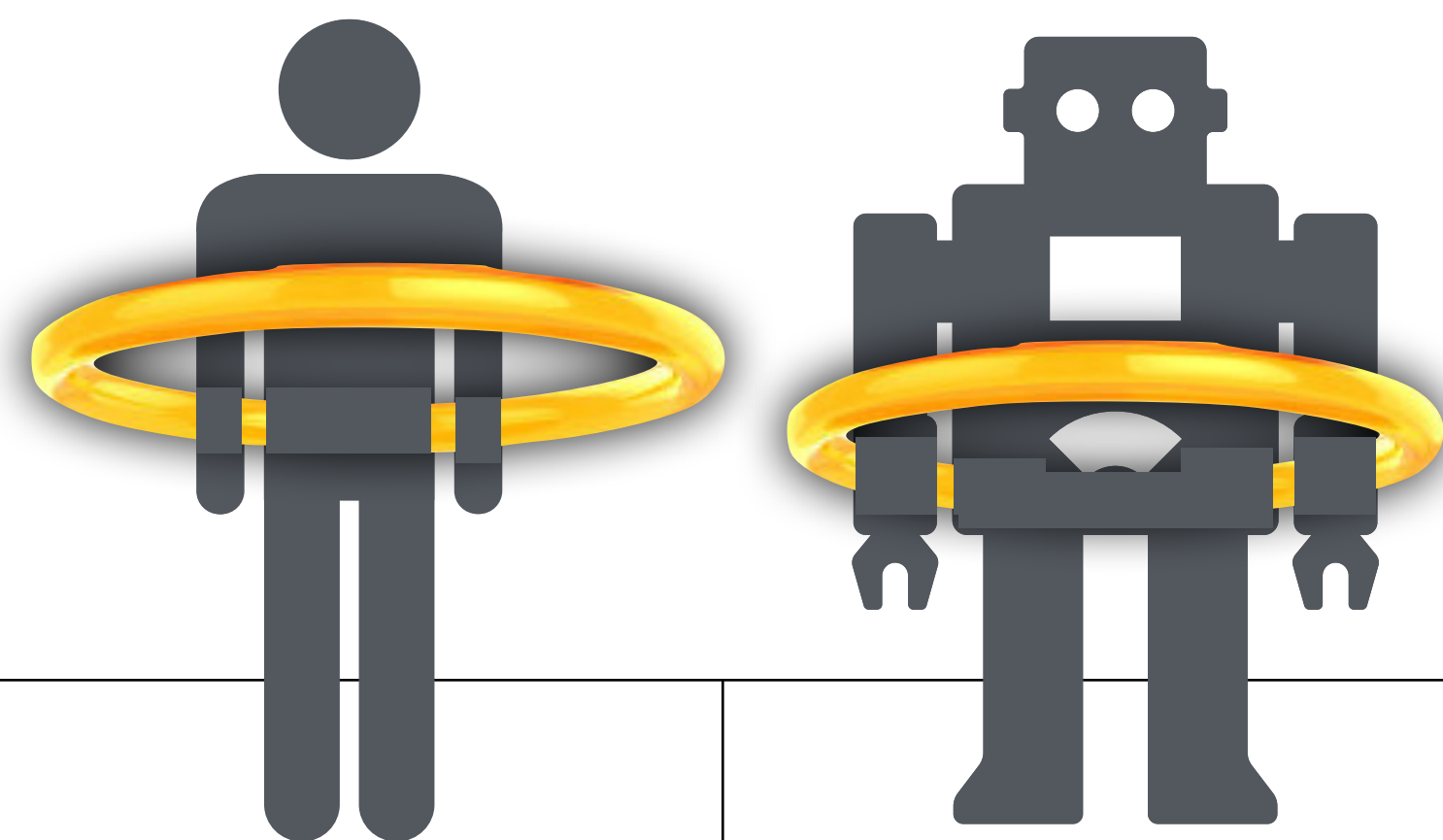
(as computer scientists)


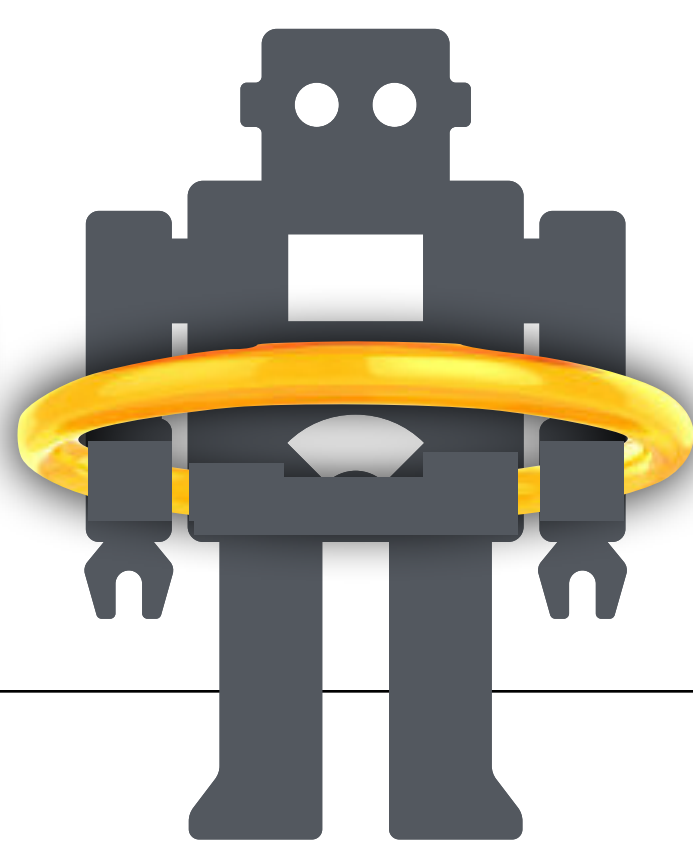


		
explicitly ethical	discerns right/wrong after guidance	can morally evaluate options & situations
implicitly ethical	e.g., in prison	actions constrained to good
source of morally relevant information	grew up among humans	?
verification of moral behaviour	exists among humans	?

The difference between implicit and explicit

(as computer scientists)

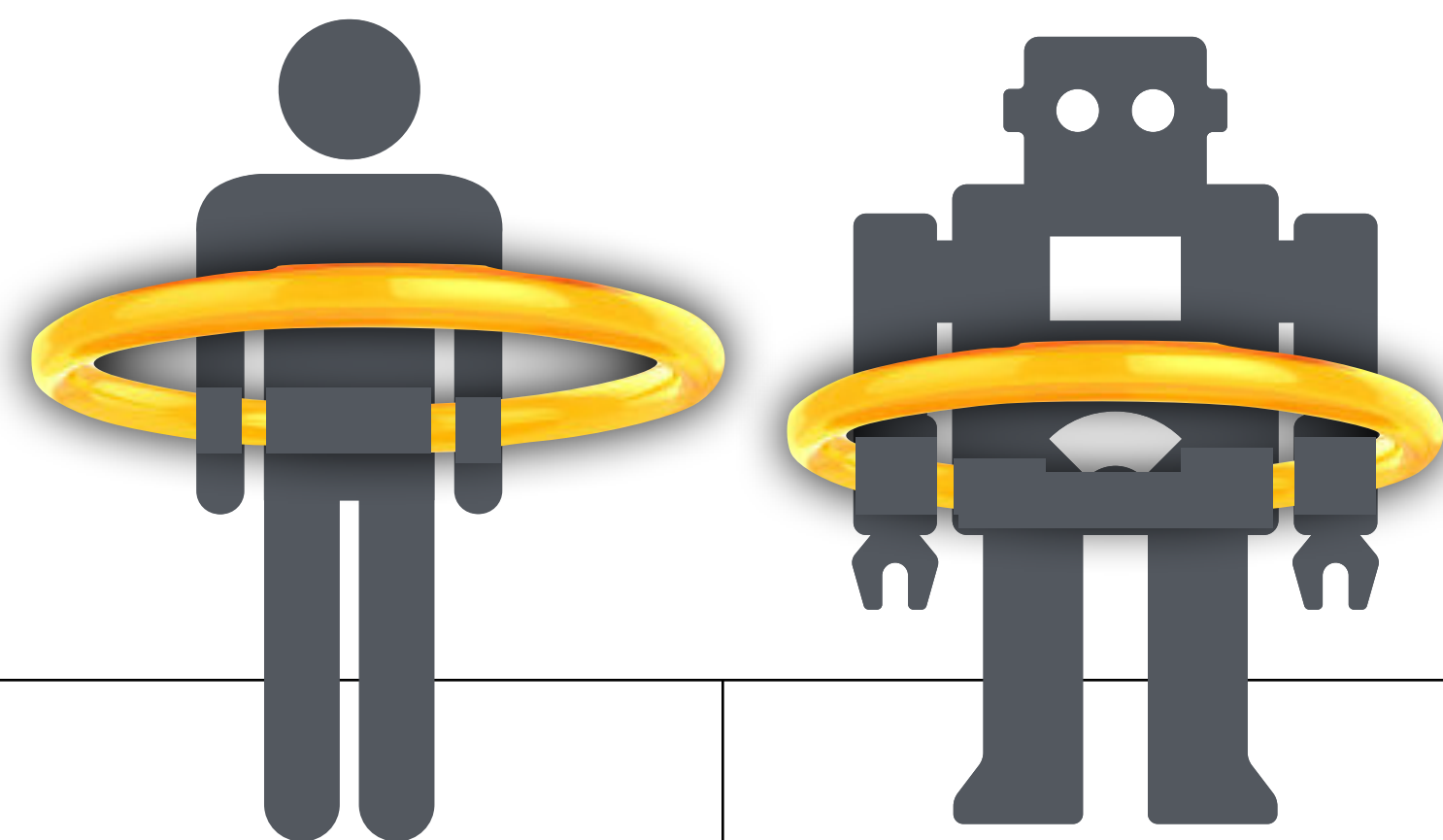



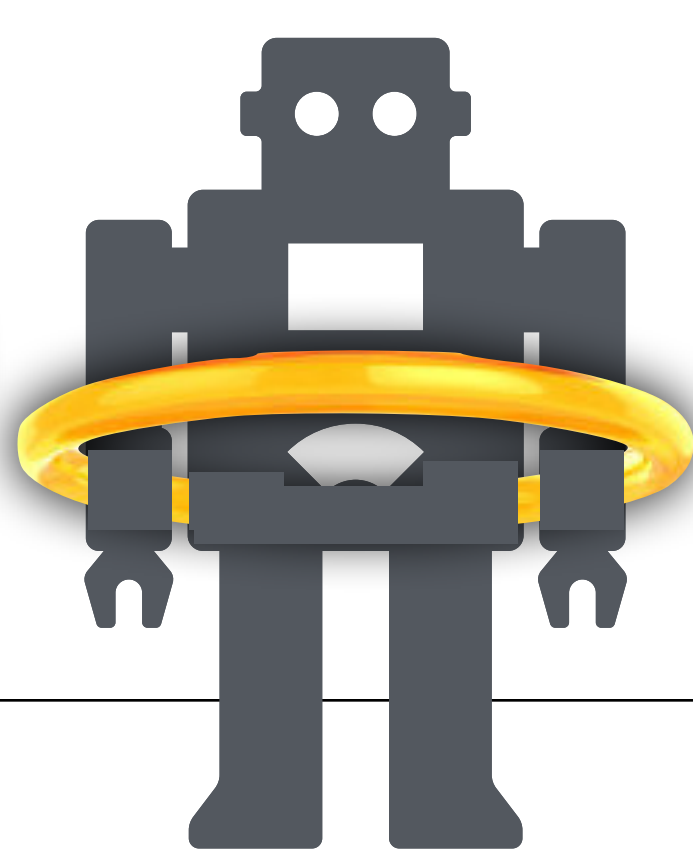
		
explicitly ethical	discerns right/wrong after guidance	can morally evaluate options & situation
implicitly ethical	e.g., in prison	actions constrained to good
source of morally relevant information	grew up among humans	?
verification of moral behaviour	exists among humans	?

A moral theory from moral philosophy

The difference between implicit and explicit

(as computer scientists)



		
explicitly ethical	discerns right/wrong after guidance	can morally evaluate options & situation
implicitly ethical	e.g., in prison	actions constrained to good
source of morally relevant information	grew up among humans	?
verification of moral behaviour	exists among humans	?

Recently in RL: from example

What do we automate?

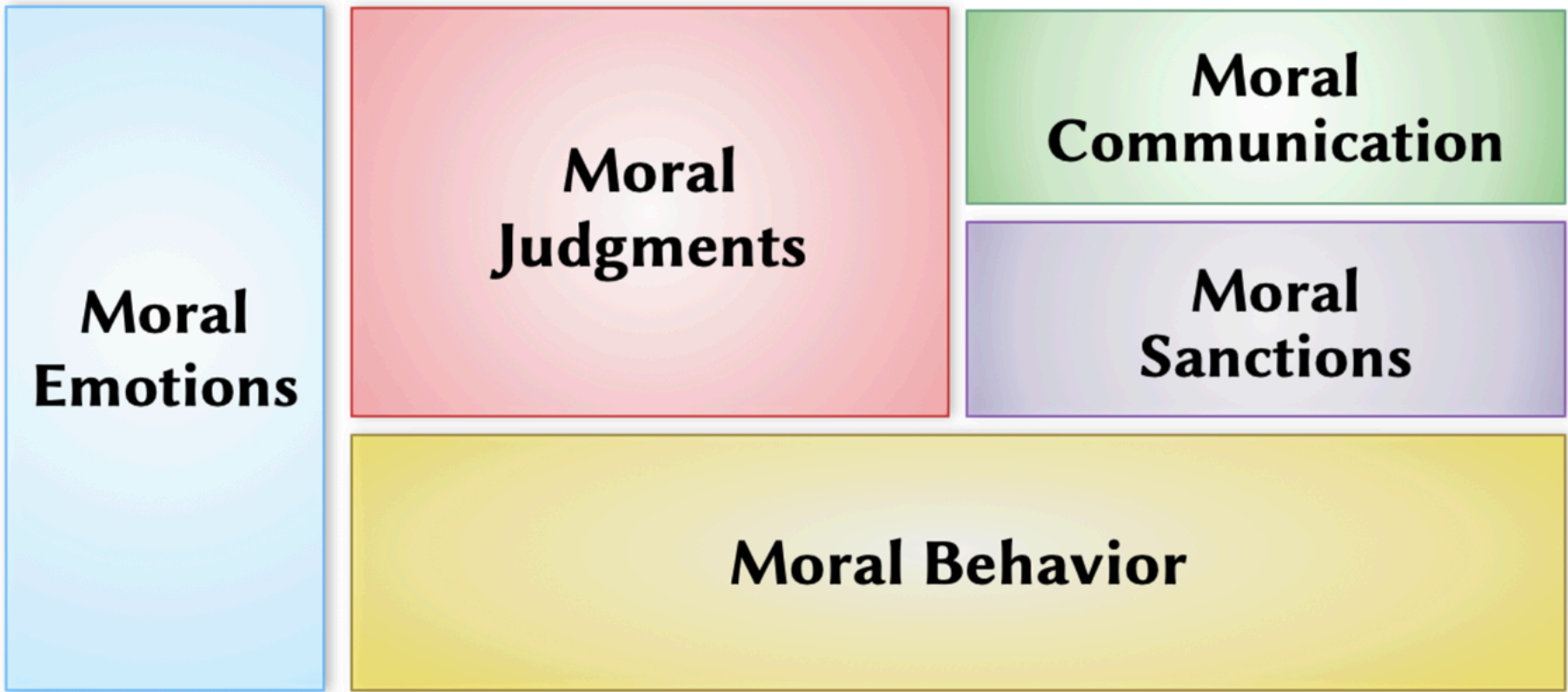


Figure 31.1. Five major moral phenomena: moral behavior (including moral decision making), moral judgments, moral emotions, moral sanctions, and moral communication.

Bello, P., & Malle, B. F. (2023). Computational approaches to morality. In R. Sun (Ed.), *Cambridge Handbook of Computational Cognitive Sciences* (pp. 1037-1063). Cambridge University Press.

What do we automate?

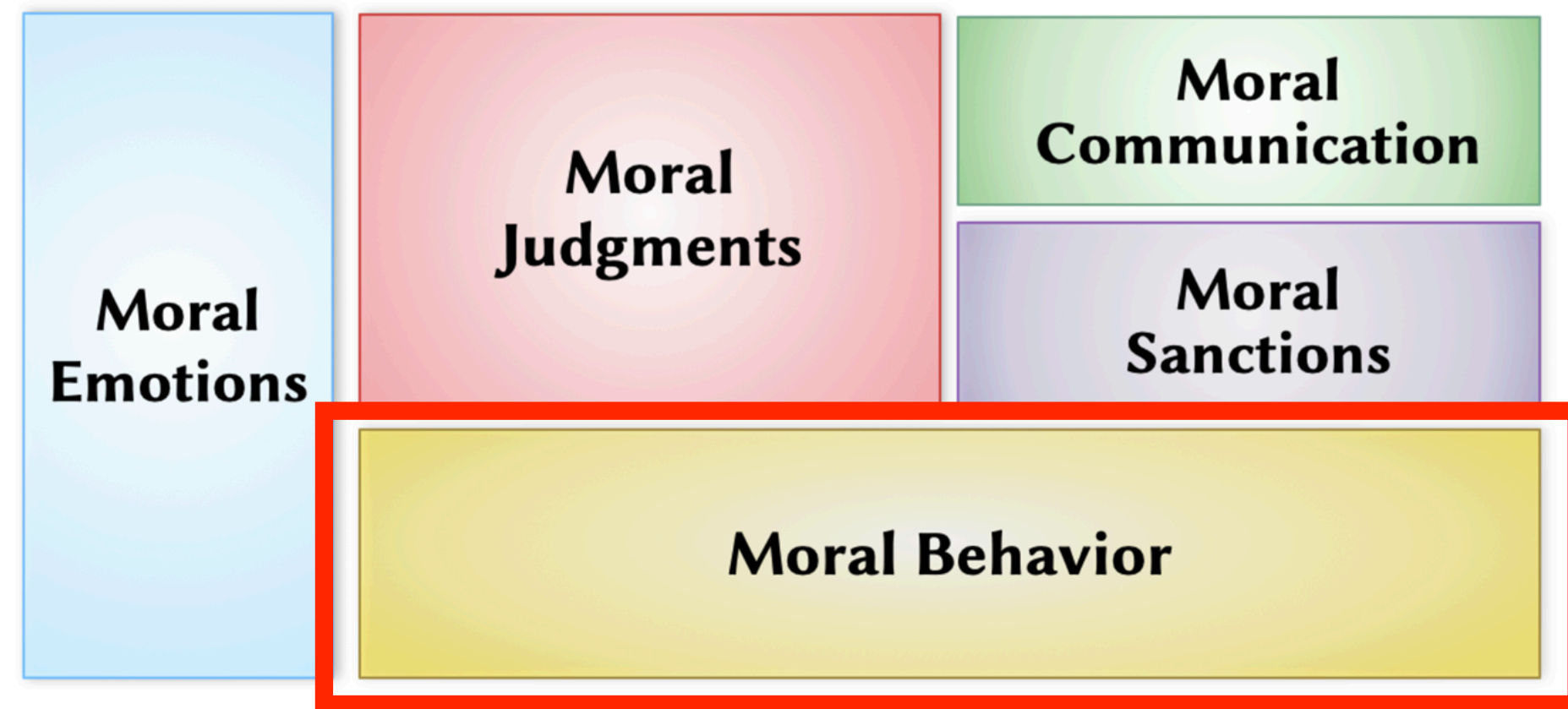


Figure 31.1. Five major moral phenomena: moral behavior (including moral decision making), moral judgments, moral emotions, moral sanctions, and moral communication.

Bello, P., & Malle, B. F. (2023). Computational approaches to morality. In R. Sun (Ed.), *Cambridge Handbook of Computational Cognitive Sciences* (pp. 1037-1063). Cambridge University Press.

31

Computational Approaches to Morality

Paul Bello and Bertram F. Malle

What do we automate?

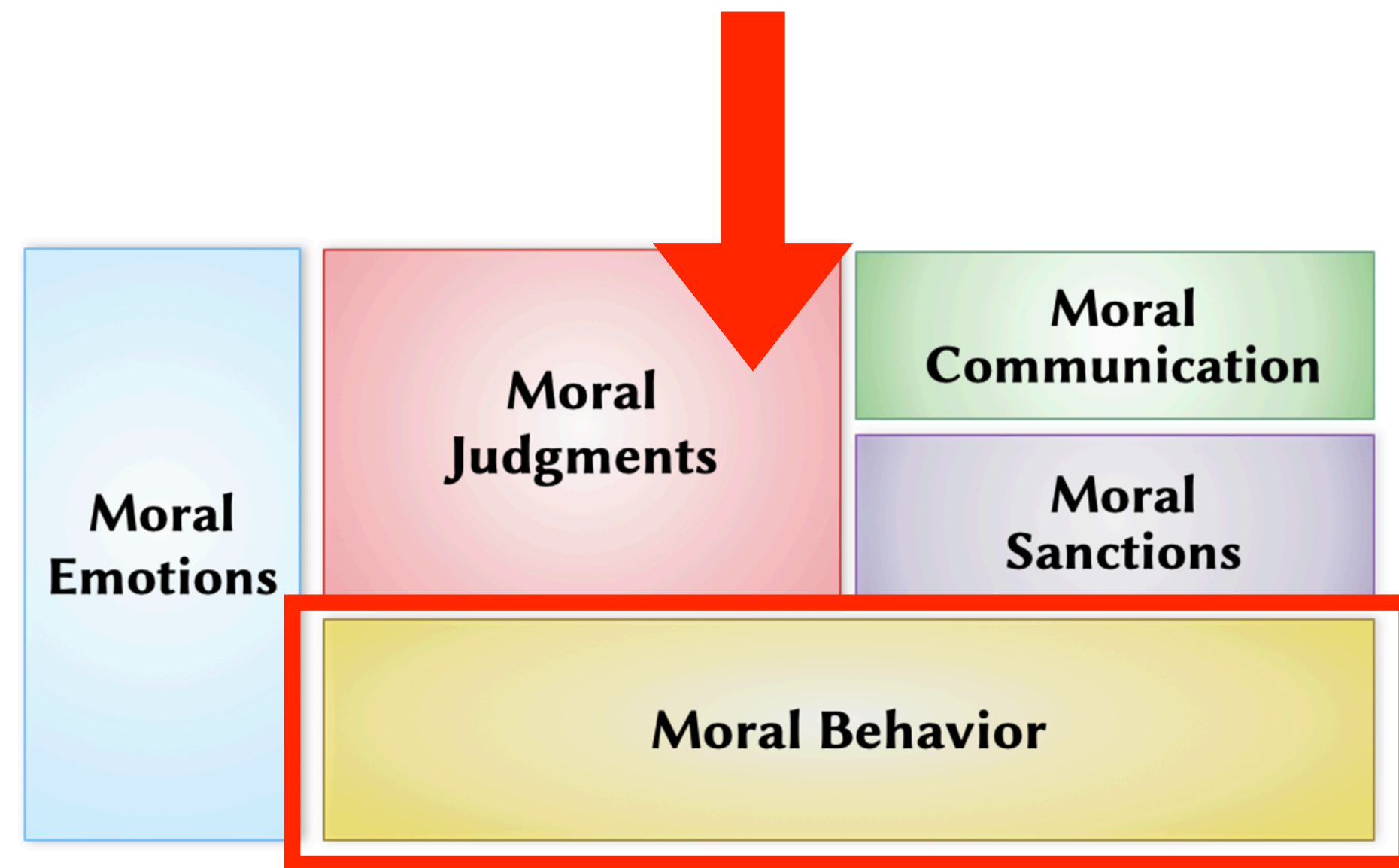


Figure 31.1. Five major moral phenomena: moral behavior (including moral decision making), moral judgments, moral emotions, moral sanctions, and moral communication.

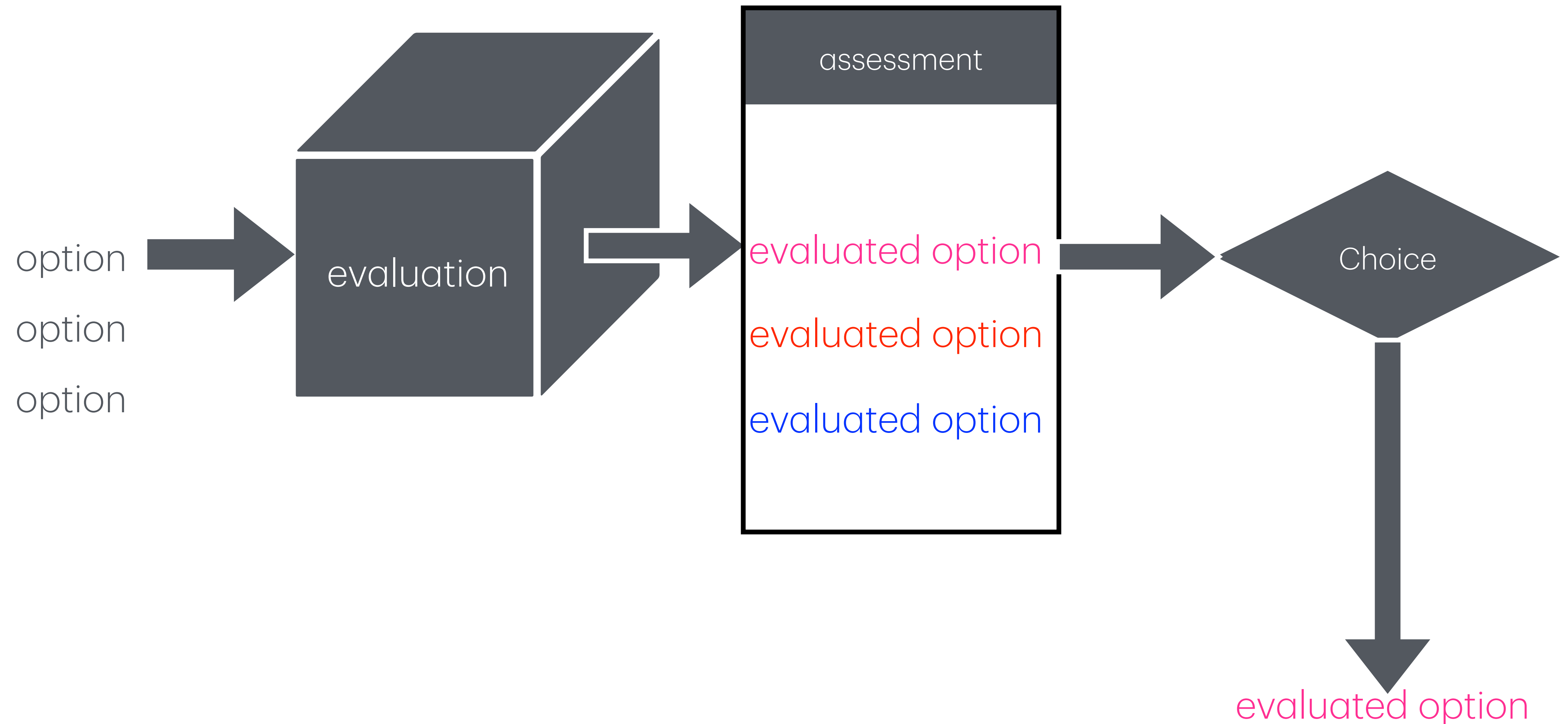
Bello, P., & Malle, B. F. (2023). Computational approaches to morality. In R. Sun (Ed.), *Cambridge Handbook of Computational Cognitive Sciences* (pp. 1037-1063). Cambridge University Press.

31

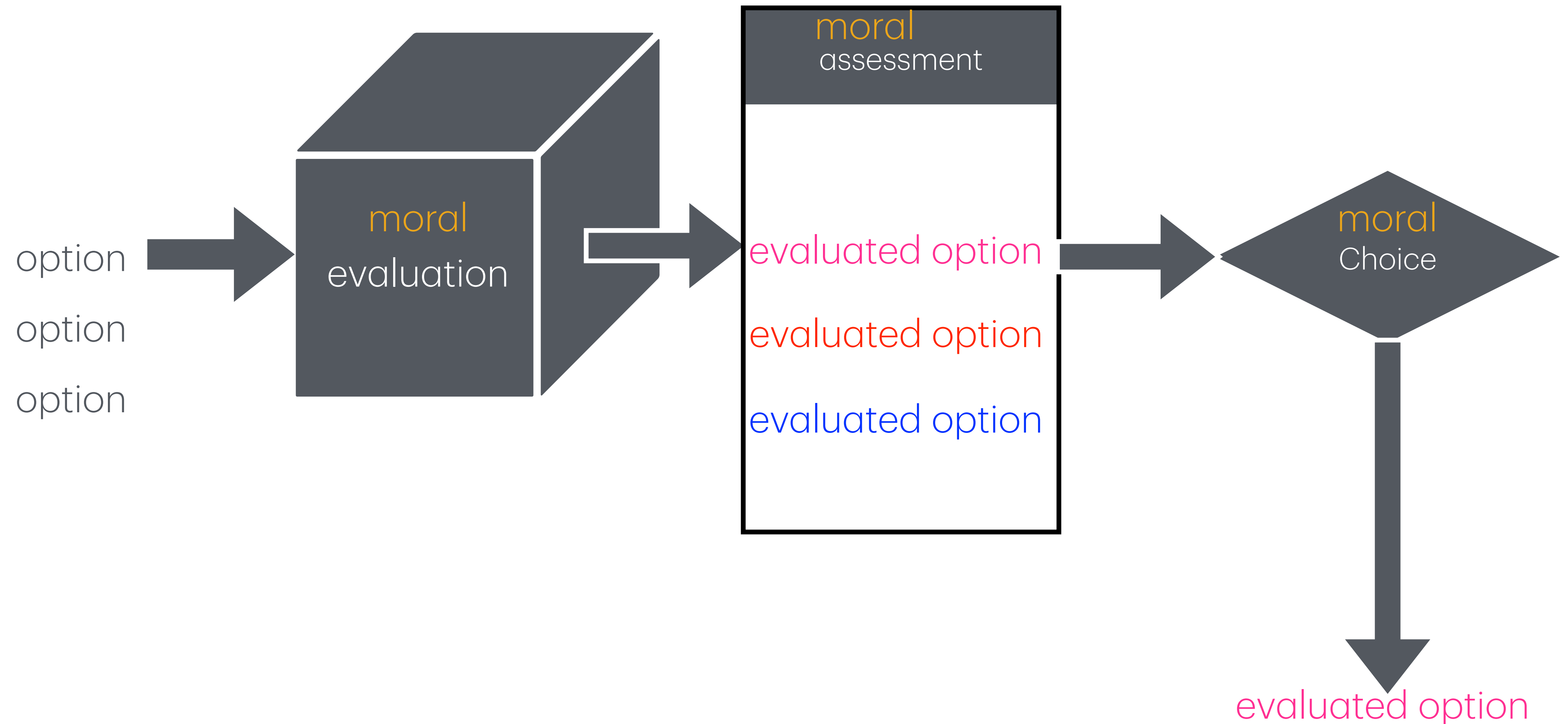
Computational Approaches to Morality

Paul Bello and Bertram F. Malle

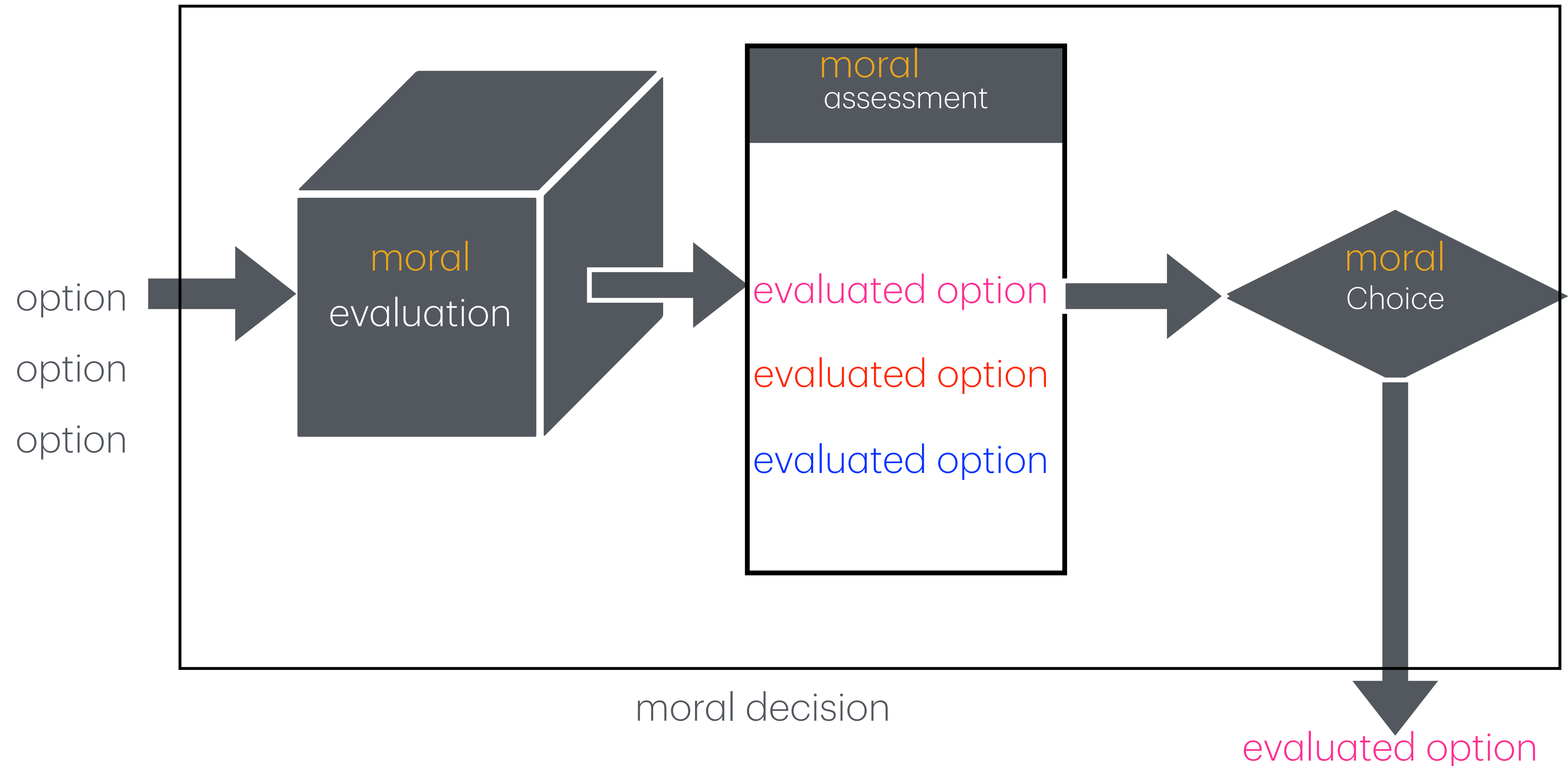
Making a decision



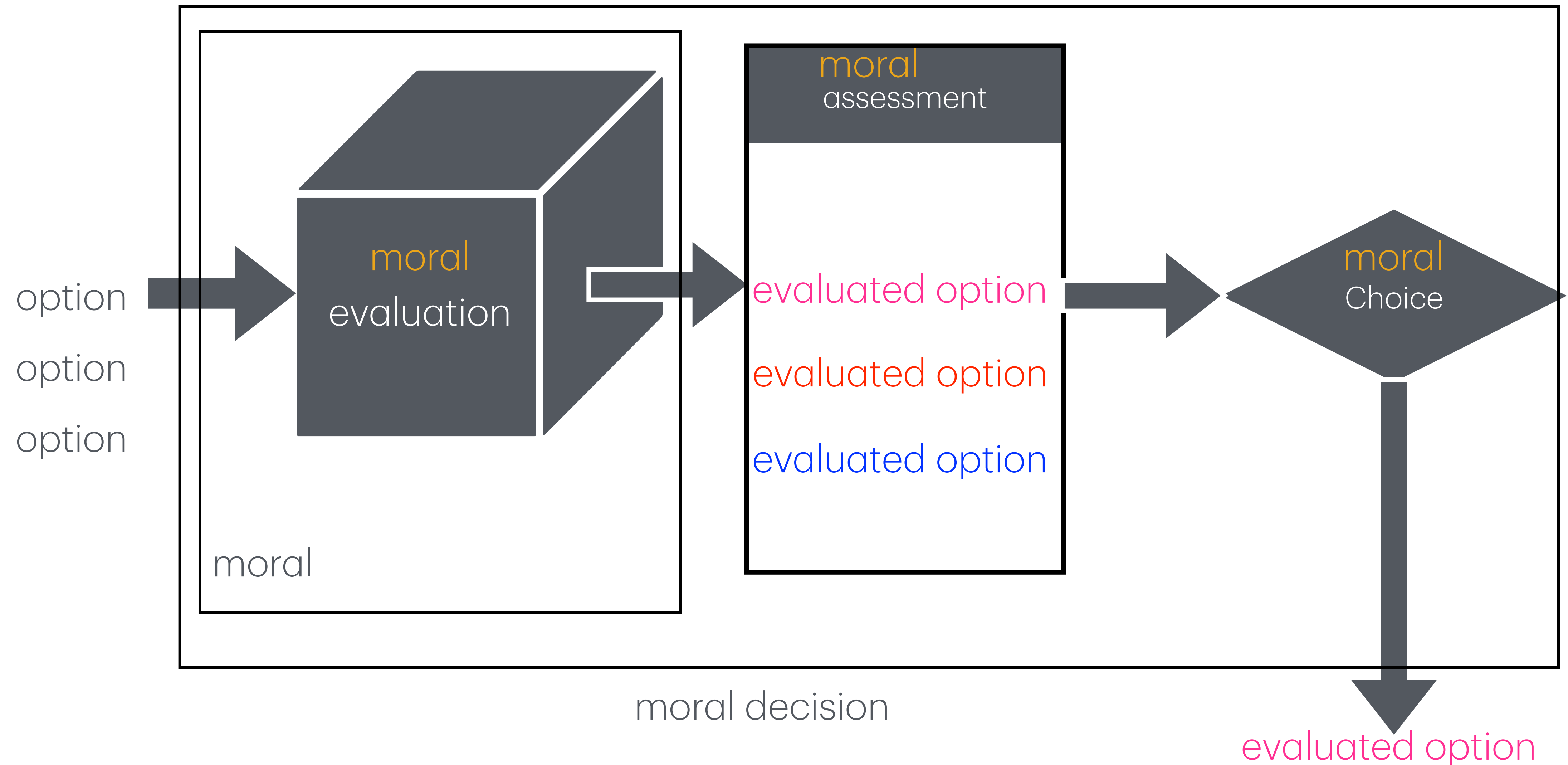
Making a **moral** decision



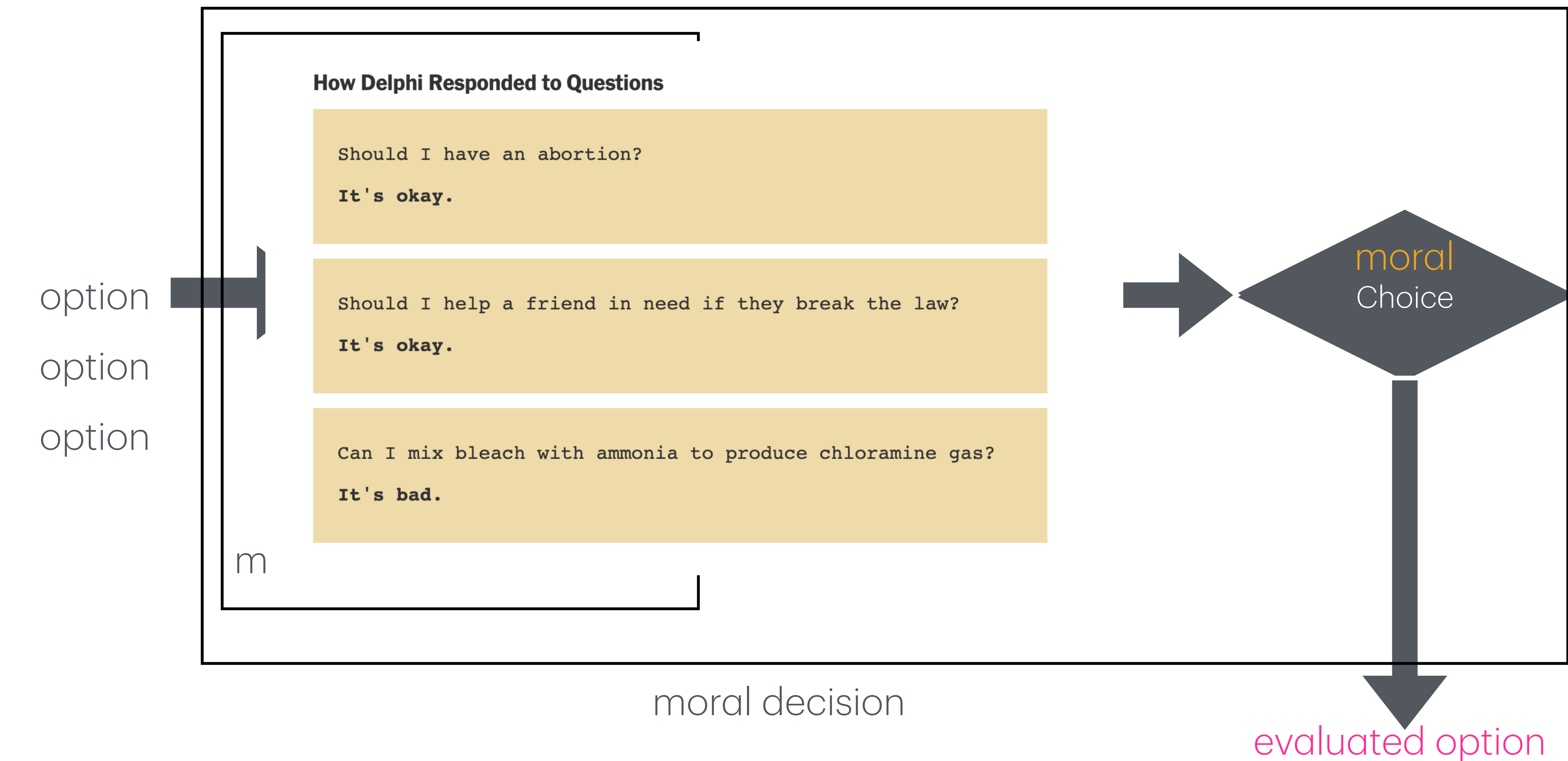
Making a **moral** decision



Making a **moral** decision



Making a **moral** decision



Moral evaluation

what factors matter?



Wikipedia, the free encyclopedia
Phan Thi Kim Phuc - Wikipedia



CNN

Napalm Girl' at 50: The story of the ...



Moral evaluation

what factors matter?



Wikipedia, the free encyclopedia
Phan Thi Kim Phuc - Wikipedia



CNN
Napalm Girl' at 50: The story of the ...

- Considering all factors that matter for the problem in the context

Moral evaluation

what factors matter?



naked child

historical document

censoring

- Considering all factors that matter for the problem in the context

Moral evaluation

what factors matter?



naked child

historical document

censoring

- Considering all factors that matter for the problem in the context
- Evaluate the options with respect to the morally relevant factors

Moral evaluation

what factors matter?



Wikipedia, the free encyclopedia
Phan Thi Kim Phuc - Wikipedia

naked child

historical document



CNN
Napalm Girl' at 50: The story of the ...

historical
document

censoring

- Considering all factors that matter for the problem in the context
- Evaluate the options with respect to the morally relevant factors



censoring

Moral evaluation

what factors matter?



naked child

historical document



historical
document

censoring



censoring

- Considering all factors that matter for the problem in the context
- Evaluate the options with respect to the morally relevant factors
- Associate each option with one or more moral qualities

Moral evaluation

what factors matter?

harm
fidelity



naked child
historical document

harm
fidelity



historical
document
censoring

fidelity



censoring

- Considering all factors that matter for the problem in the context
- Evaluate the options with respect to the morally relevant factors
- Associate each option with one or more moral qualities

Moral assessment

Classifying options into normative categories according to their deontic moral status

- This step enables the decision
- Assigns qualities or quantities that allow for the options to be compared.



-1



0



-3

Moral assessment

Classifying options into normative categories according to their deontic moral status

- This step enables the decision
- Assigns qualities or quantities that allow for the options to be compared.



-1
wrong



0
maybe less wrong



-3
wrong

Moral choice

- The act of selecting a decision
- The act of executing a decision in a morally acceptable way
- Ideally we always choose from the morally acceptable set
- Ideally, the choice can be explained, justified, defended

Different normative domains

what is moral?

Good/Bad	Right/Wrong
gradable allow neutral states non-privative opposites not duals not alternative dependent	not gradable no neutral states privative opposites duals alternative dependent

Right and wrong are the paradigm examples from the class of what the literature calls ‘deontic categories’. This class also includes *required*, *obligatory*, *forbidden*, *prohibited*, *permissible*, *optional*, and their many cognates. Deontic categories are also often picked out using such terms as ‘ought to’, ‘must’, ‘may’, and others.² The deontic categories form a class because they resemble each other in several ways, and because they are related to each other in ways that they are not related to non-deontic normative categories, cf. (Berker 2022).

- Deontic categories (right, wrong etc)
- Evaluative categories (good, bad, etc)
- Fittingness categories (appropriate, justified etc.)
- Reason-related categories

Which hard thing to do?

Option



7

Option



3

Option



0

Which hard thing to do?

Option



7

harmful to 1

Option



3

good for John

Option



0

harms privacy increases safety

Hard choices

moral conflicts... often considered to be the same as dilemmas

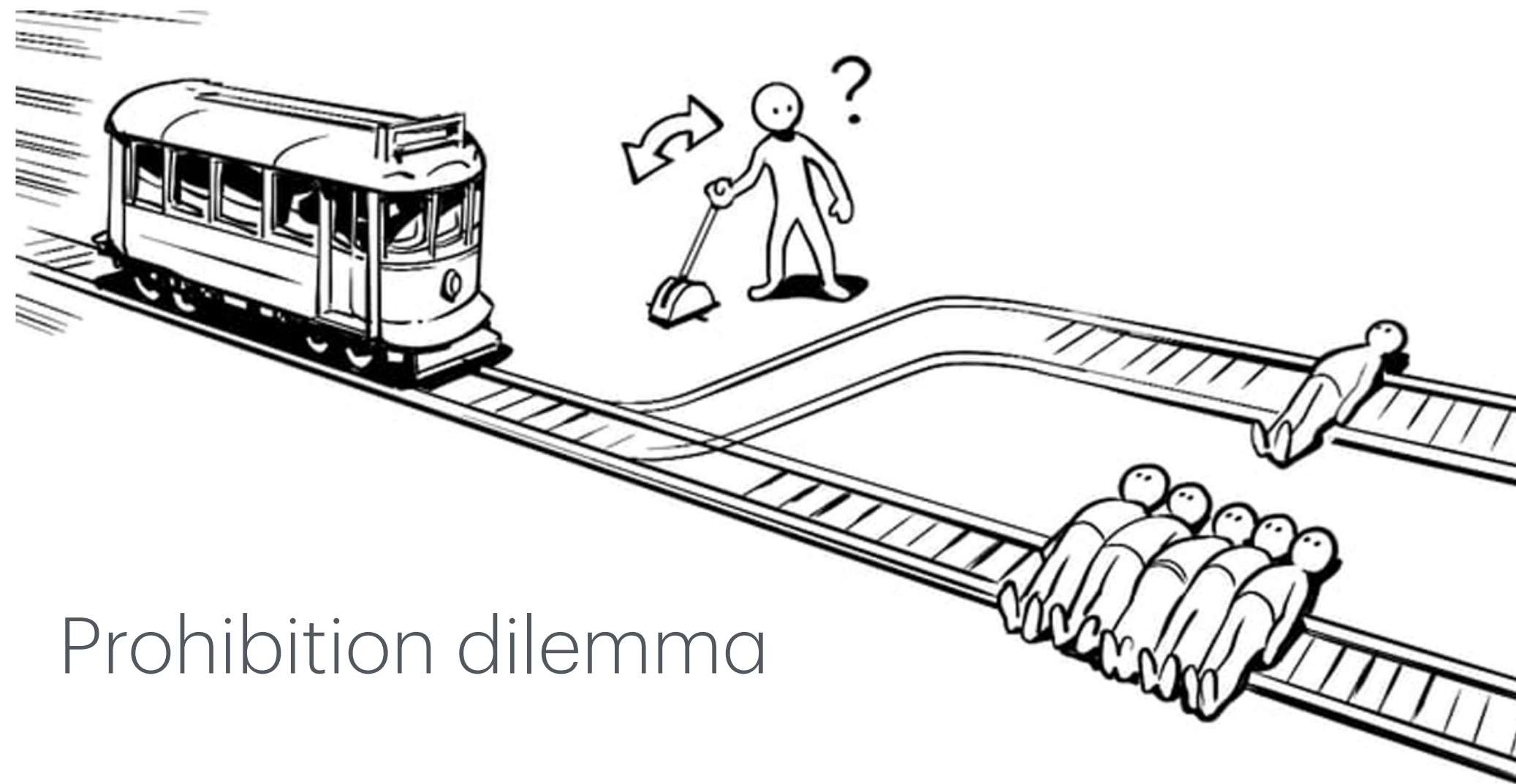
- Choice 1: purchase a plane ticket for Gran Canaria.
Human was very busy.
- Choice 2: donate to doctors without borders. Human expressed desire to be more environmentally aware



Hard choices

moral dilemmas

- *Obligation dilemma*. All the feasible actions are mandatory. The agent cannot do more than one action, so she has to make a choice based upon some sort of preferential reasoning;
- *Prohibition dilemma*. All the feasible actions are forbidden. The agent has to do one action;



Prohibition dilemma

Approximate solutions of moral dilemmas in multiple agent system

Regular Paper | Published: 16 October 2008

Volume 18, pages 157–181, (2009) [Cite this article](#)






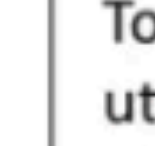





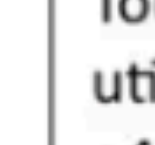
✓ Access provided by Università degli Studi di Bologna – Area Biblioteche e Servizi allo Studio

[Download PDF](#) ↓

[Matteo Cristani](#) ✉ & [Elisa Burato](#)

Obligation dilemma

Life-Saving Drug

Give drug to David	 +100	 -100	 -100	 -100	 -100	 -100	Total utility: -400
Give drug to the five	 -100	 +100	 +100	 +100	 +100	 +100	Total utility: +400

Moral uncertainty (descriptive)



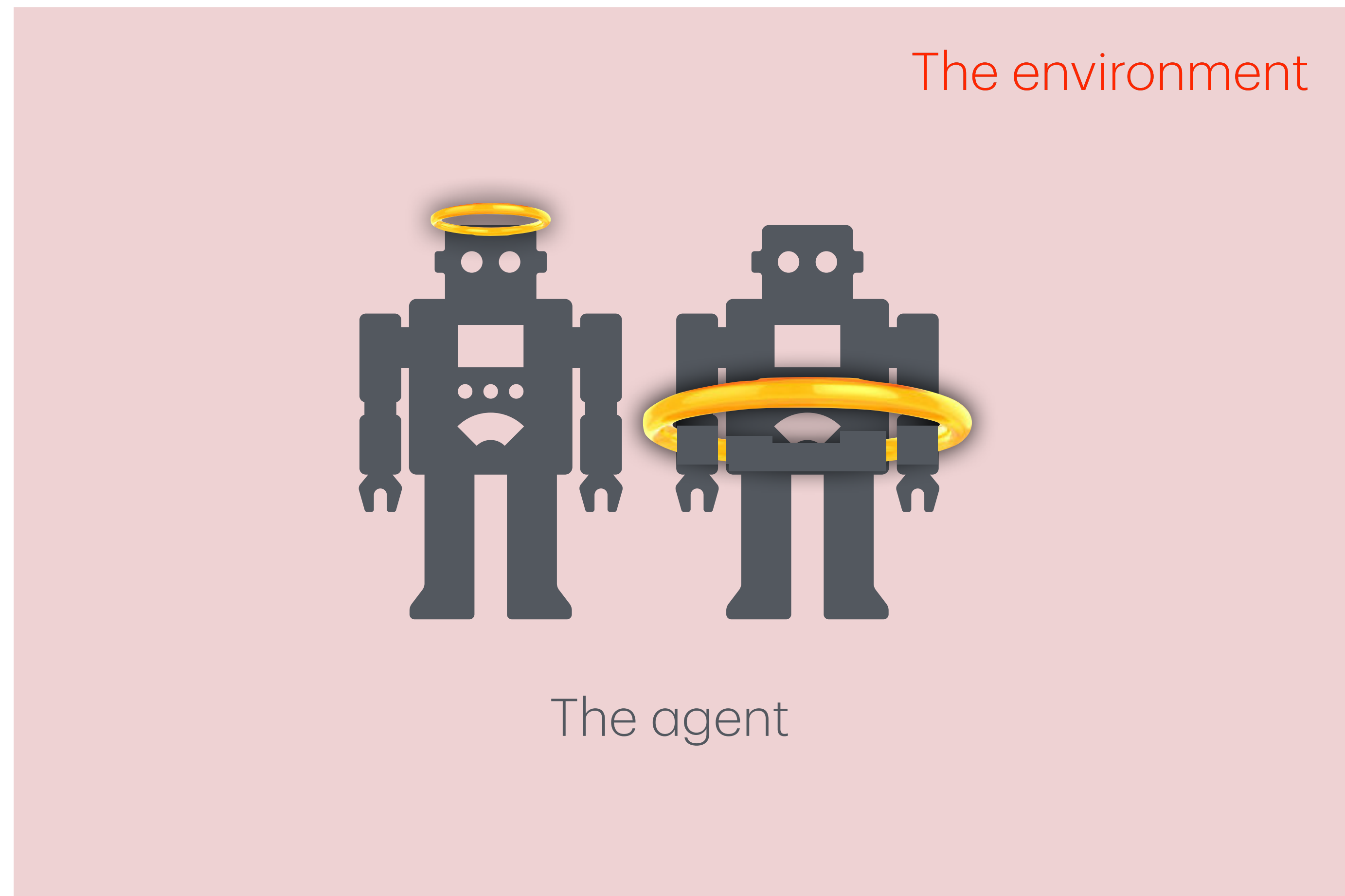
Moral uncertainty (normative)

Choose what is most fair

Choose what is least harm fun



What can we engineer?



Engineering the environment

Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)

Multi-Objective Reinforcement Learning for Designing Ethical Environments

Manel Rodriguez-Soto¹, Maite Lopez-Sanchez², Juan A. Rodriguez-Aguilar¹

¹Artificial Intelligence Research Institute (IIIA-CSIC), Bellaterra, Spain

²Universitat de Barcelona (UB), Barcelona, Spain

{manel.rodriguez, jar}@iiia.csic.es, maite_lopez@ub.edu

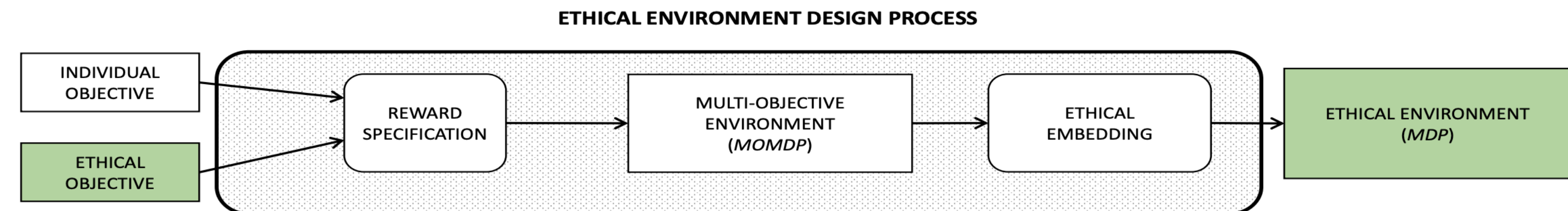


Figure 1: The process of designing an ethical environment is performed in two steps: a reward specification and an ethical embedding. Our algorithm computes the latter. Rectangles stand for objects whereas rounded rectangles correspond to processes.

Beyond agents, environments, judgments and decisions

- Moral decisions need to be explainable, justifiable and verifiable
- No systematic development, many ad-hoc approaches