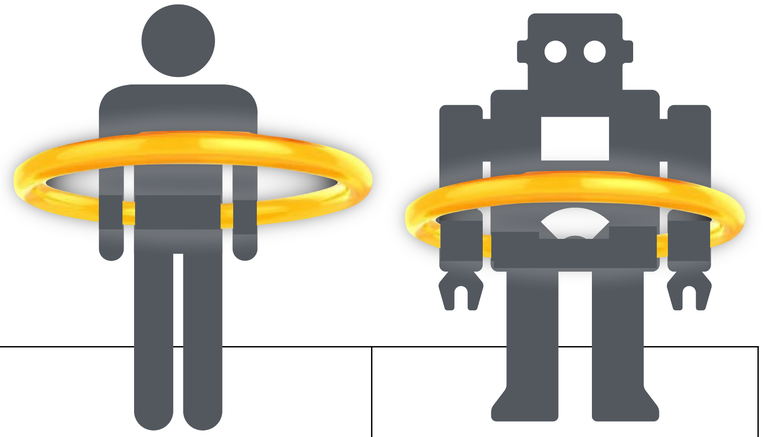
The difference between implicit and explicit

(as computer scientists)

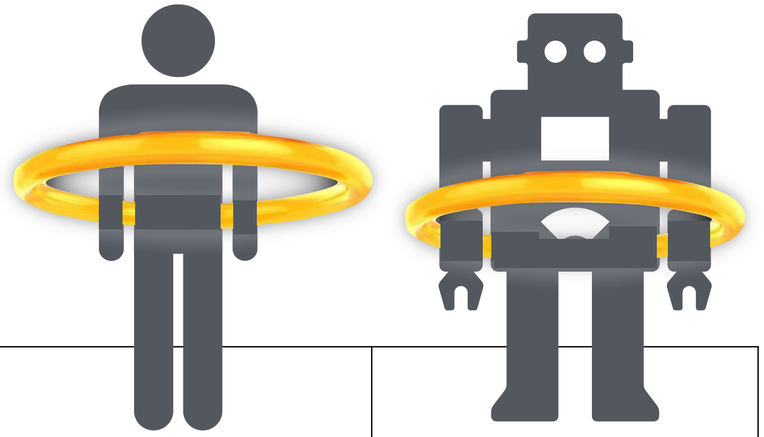


explicitly ethical	discerns right/wrong after guidinence	can morally evaluat options & situatio
implicitly ethical	e.g., in prison	actions constrained to good
source of morally relevant information	grew up among humans	?
verification of moral behaviour	exists among humans	?

A moral theory from moral philosophy

The difference between implicit and explicit

(as computer scientists)



explicitly ethical	discerns right/wrong after guidinence	can morally evaluat options & situatio
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?

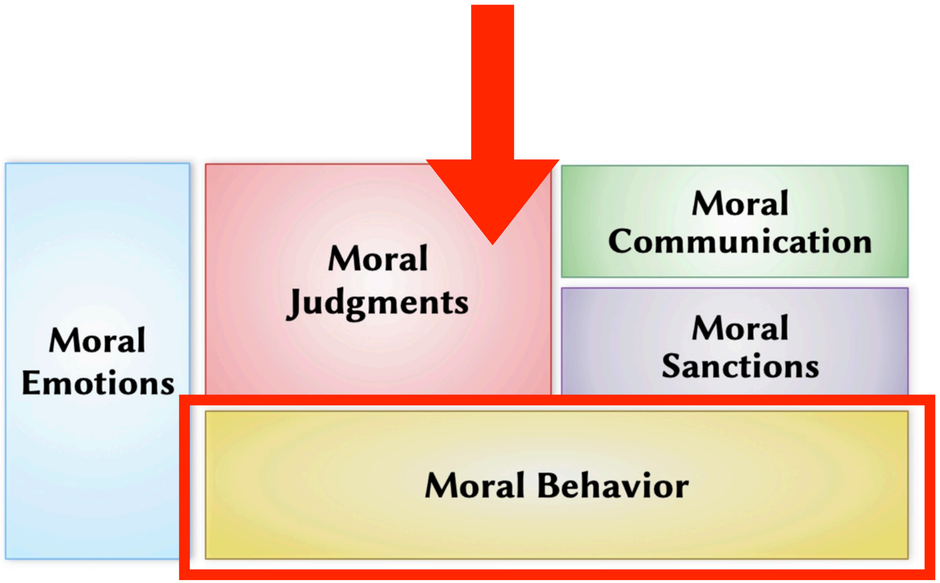


Figure 31.1 Five major moral phenomena: moral behavior (including moral decision making), moral judgments, moral emotions, moral sanctions, and moral communication.

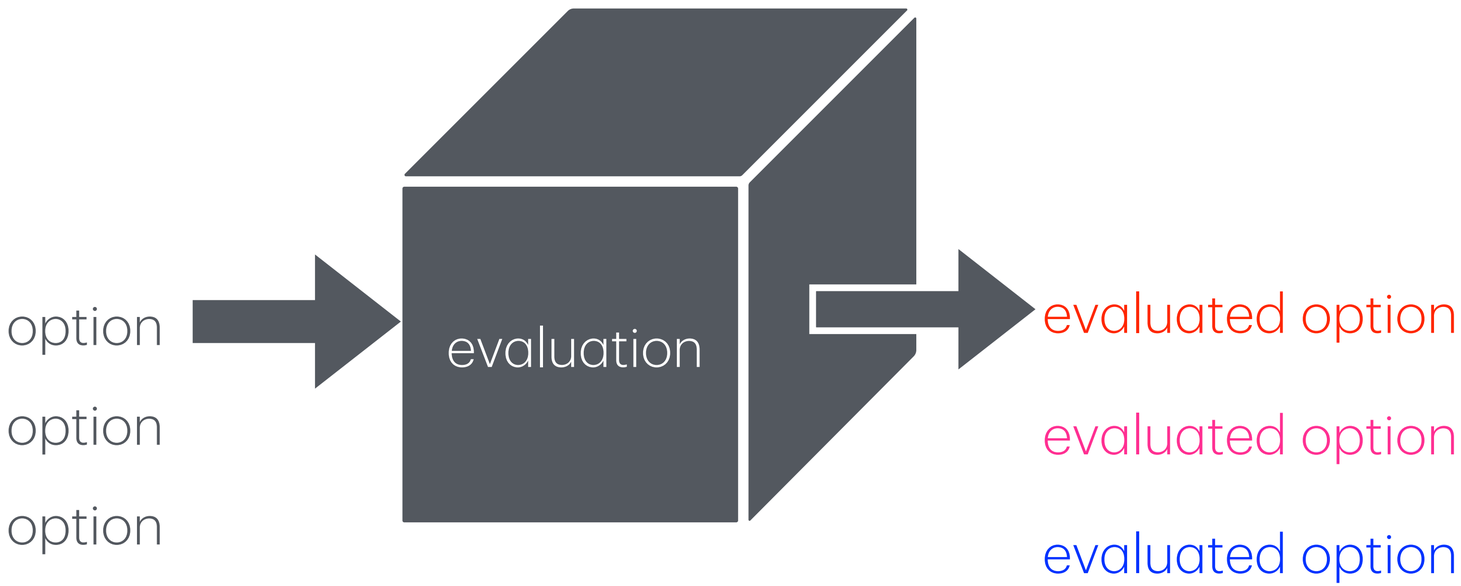
Bello, P., & Malle, B. F. (2023). Computational approaches to morality. In R. Sun (Ed.), *Cambridge Handbook of Computational Cognitive Sciences* (pp. 1037-1063). Cambridge University Press.

31

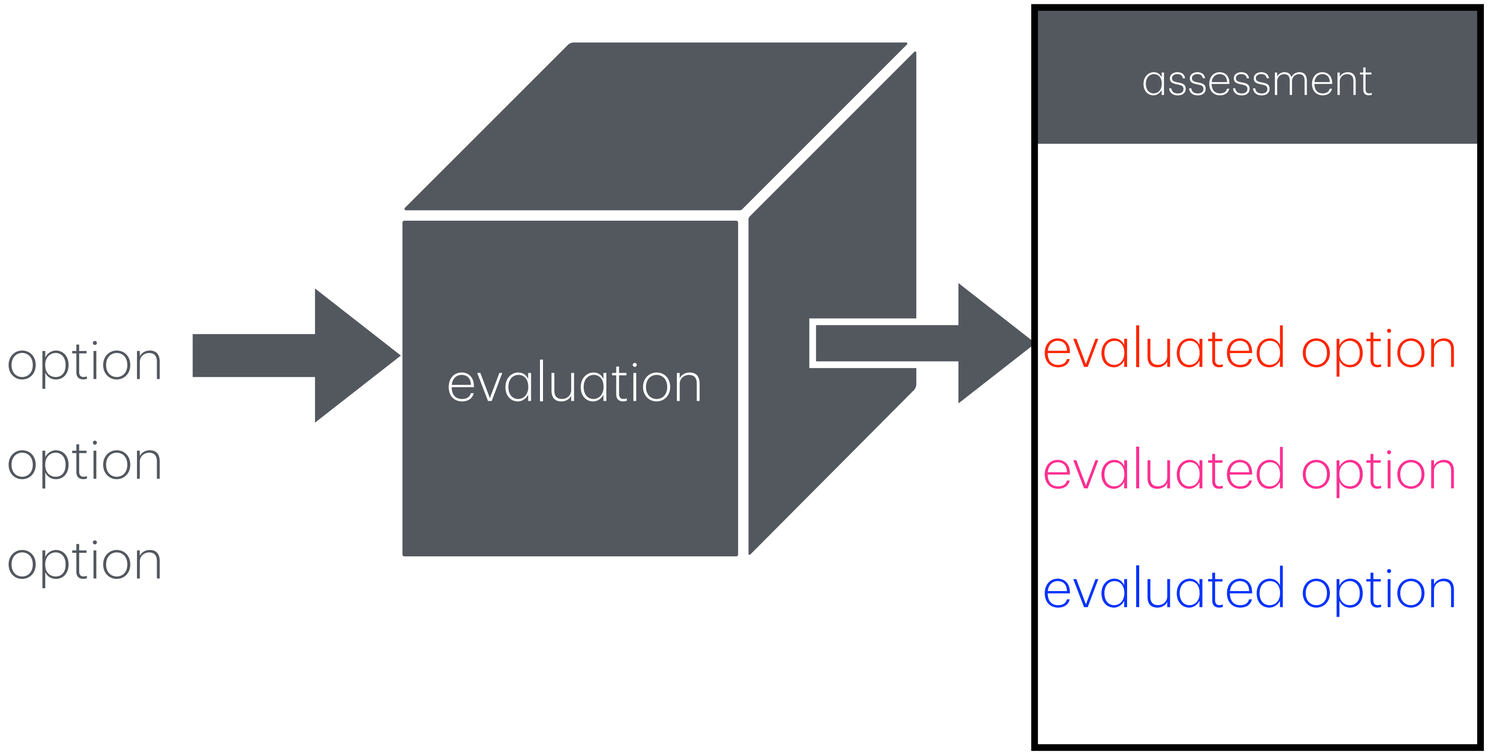
Computational Approaches to Morality

Paul Bello and Bertram F. Malle

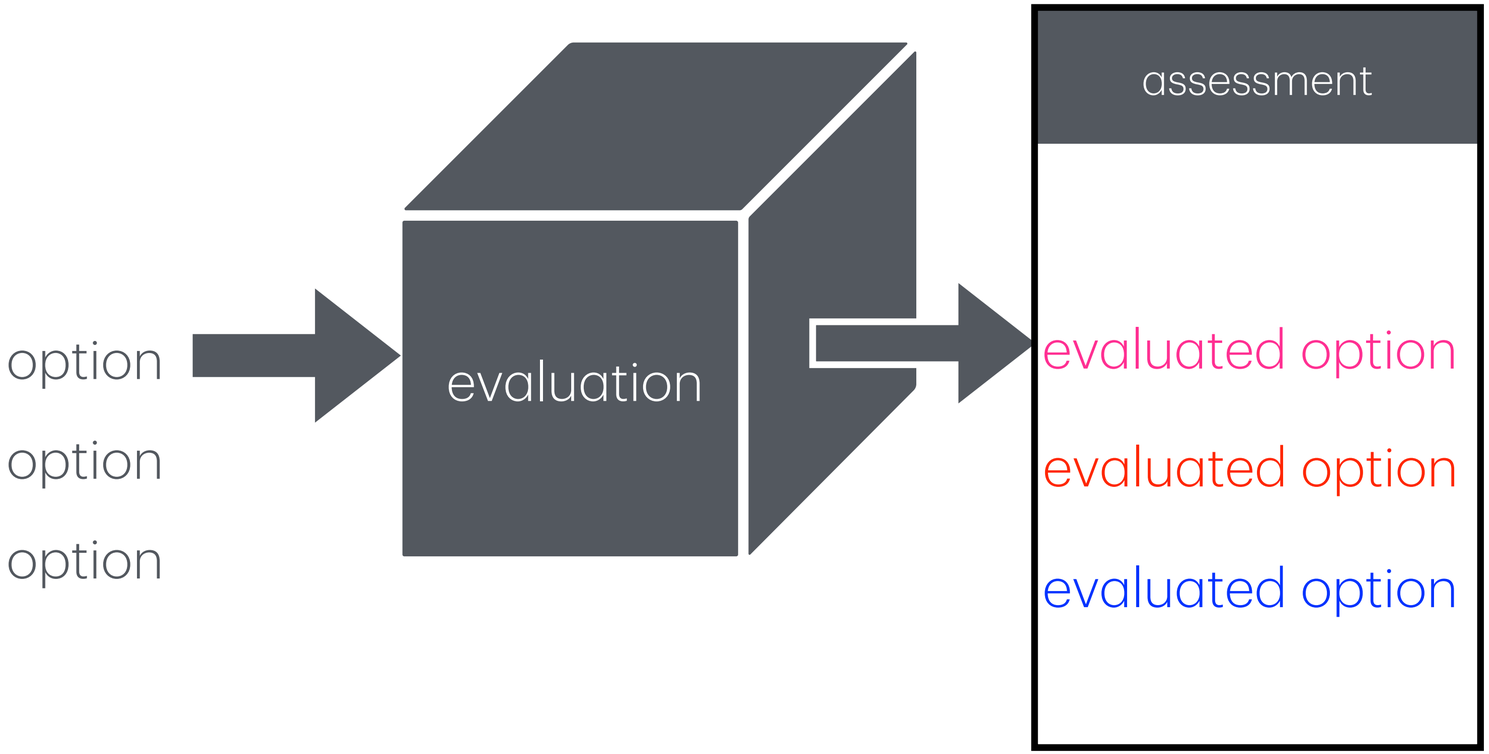
Making a decision



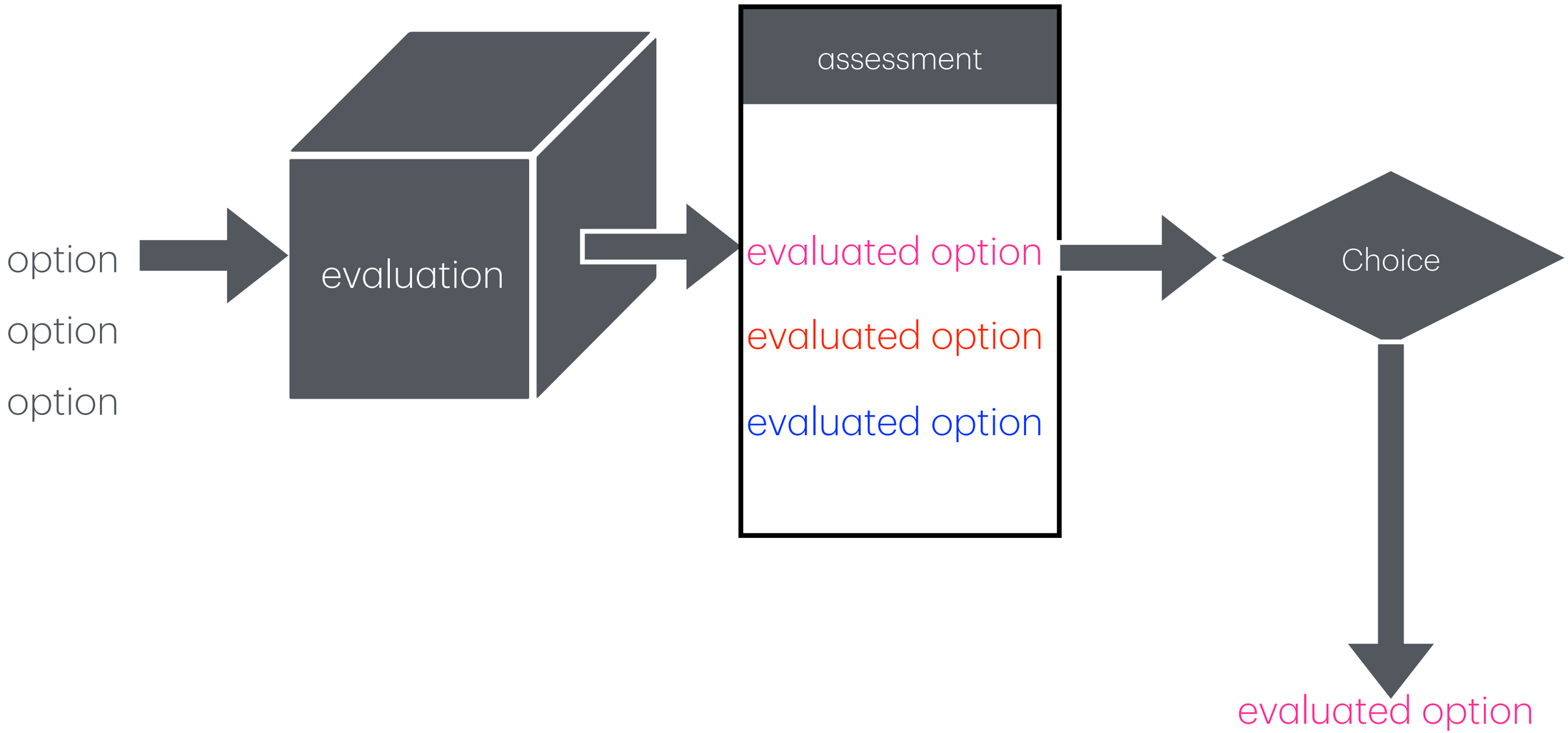
Making a decision



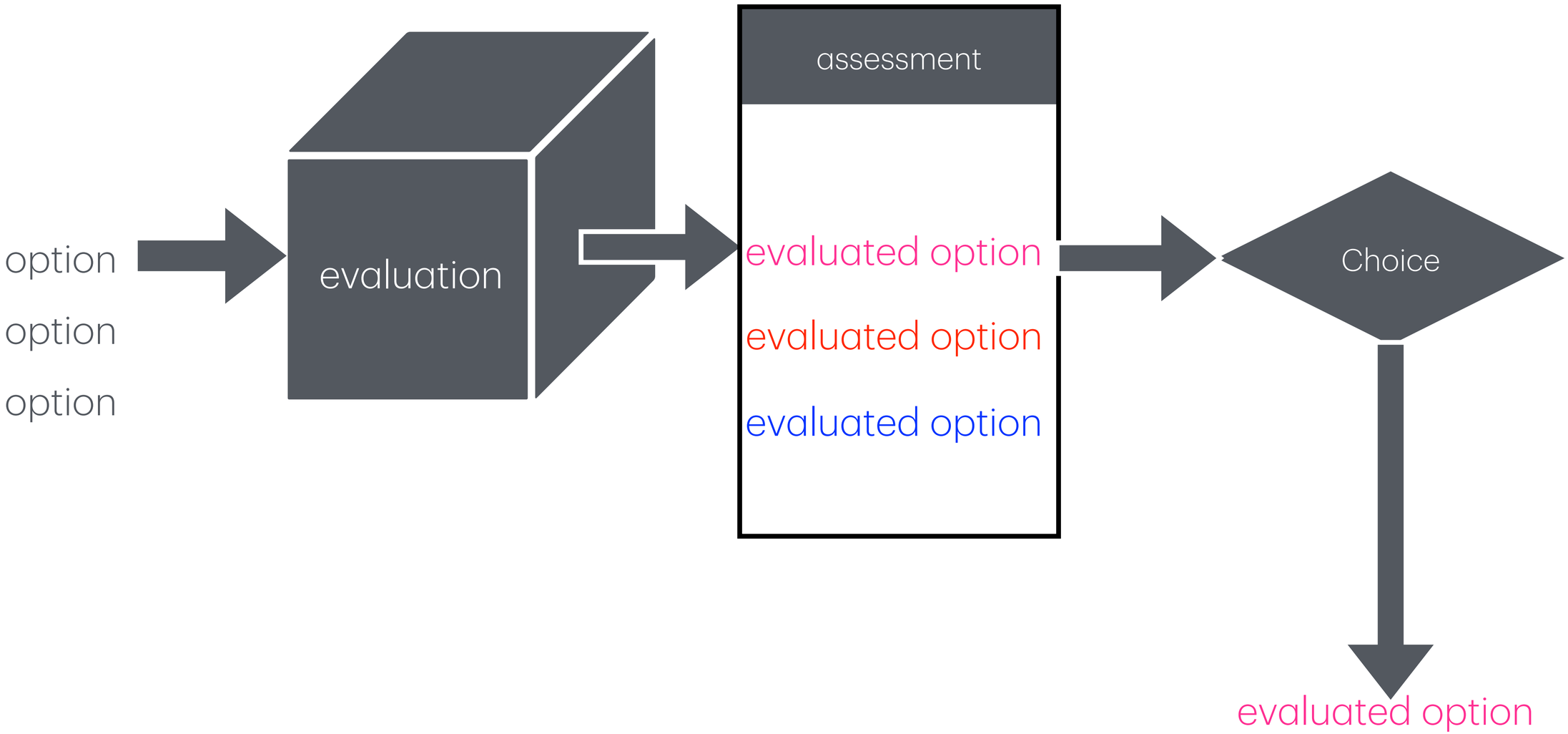
Making a decision



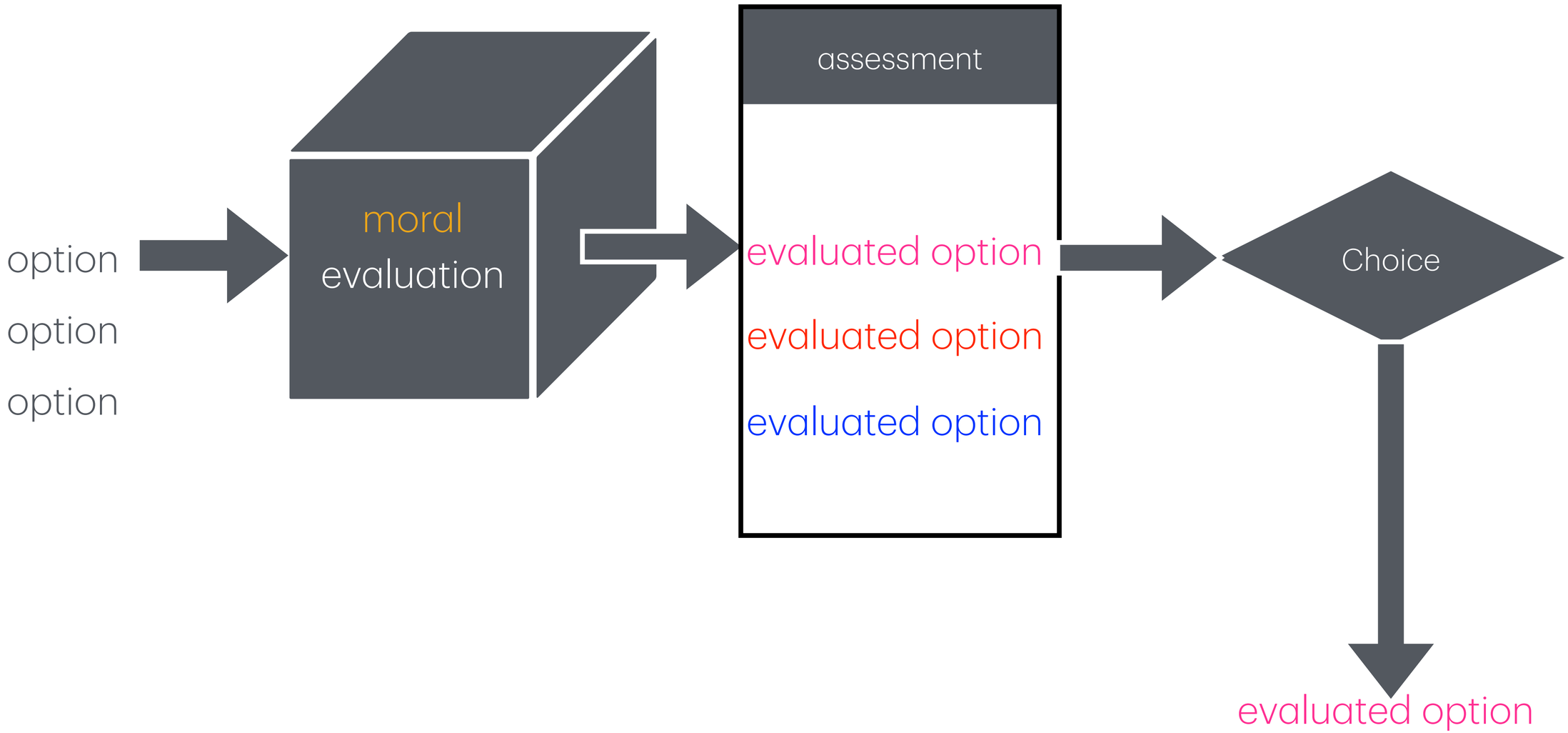
Making a decision



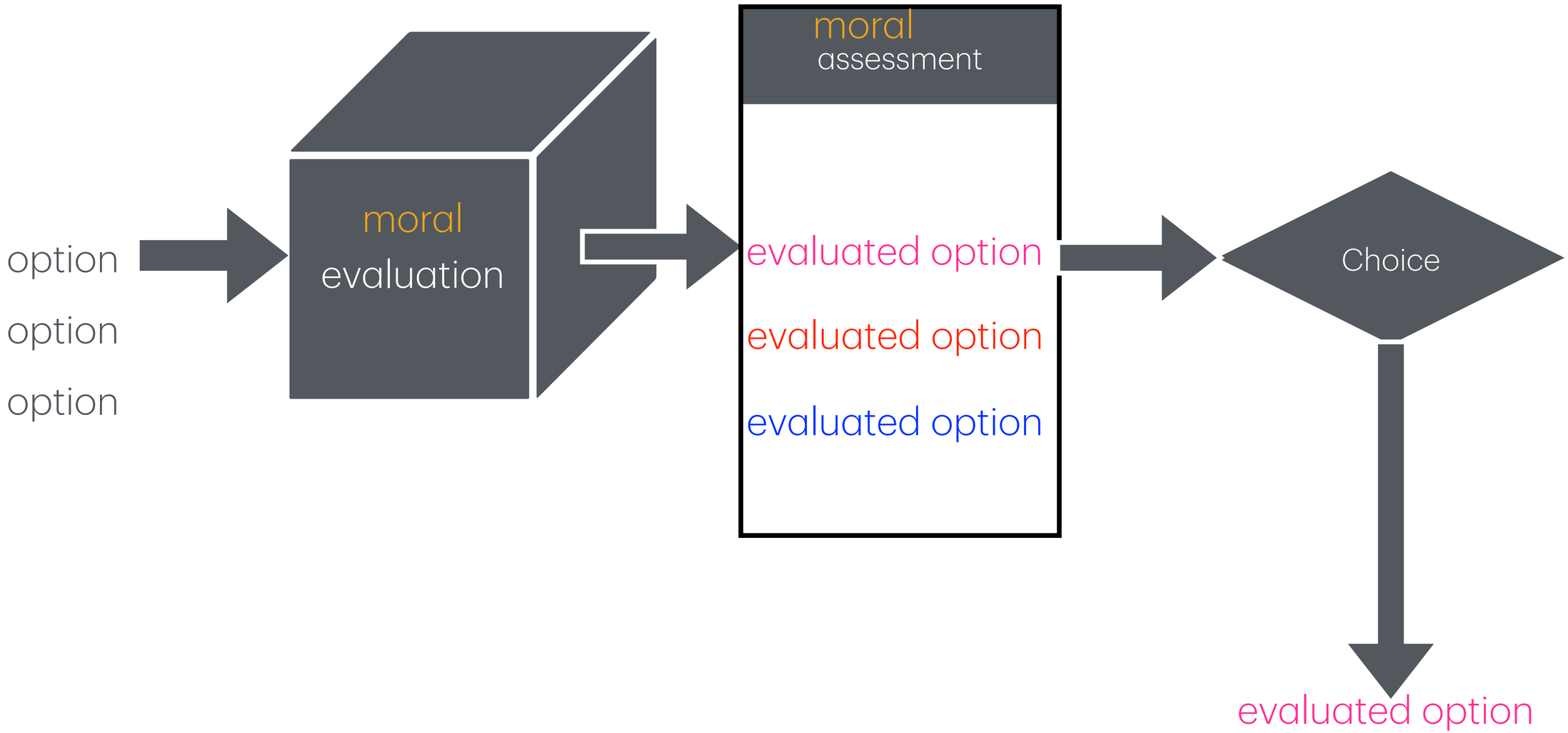
Making a moral decision



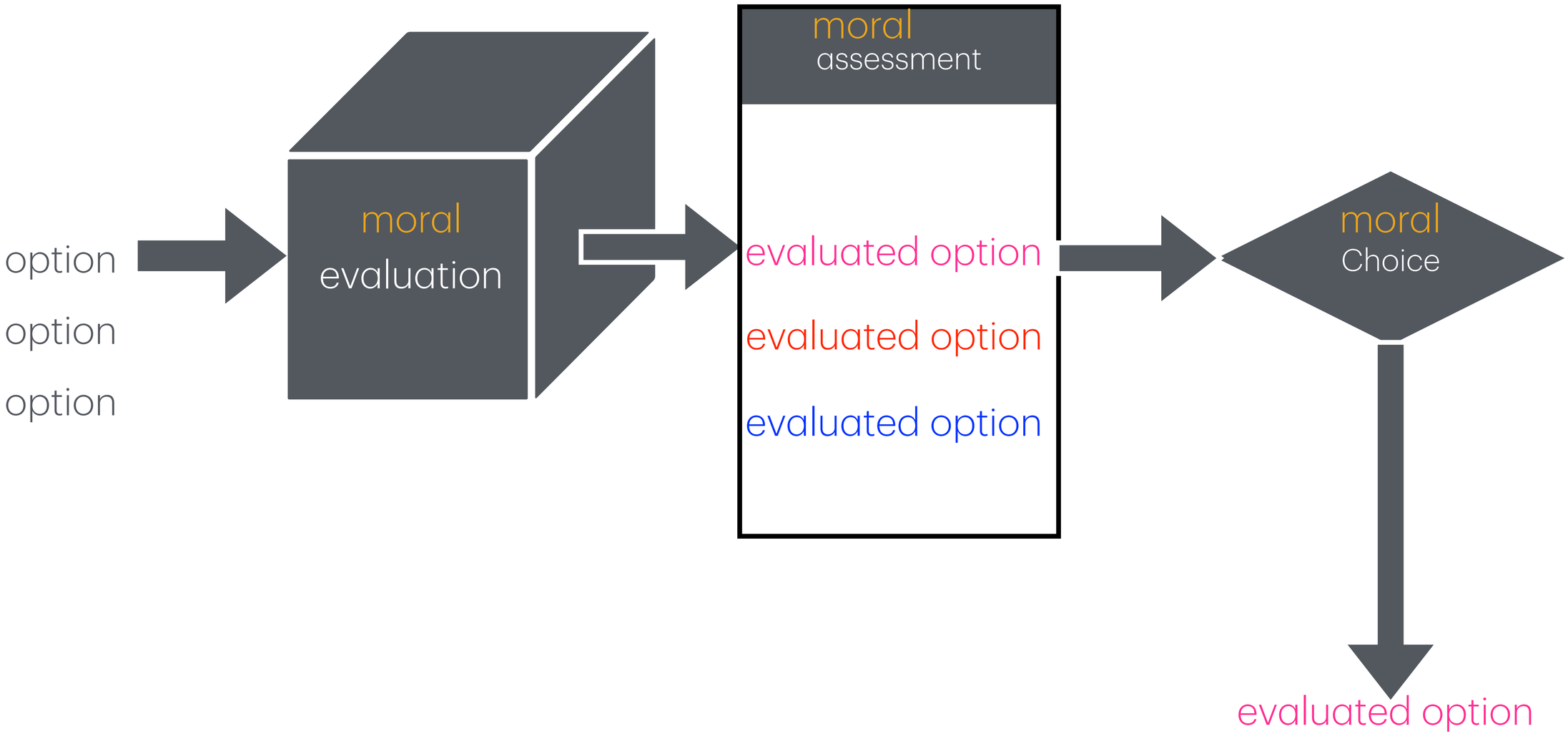
Making a moral decision



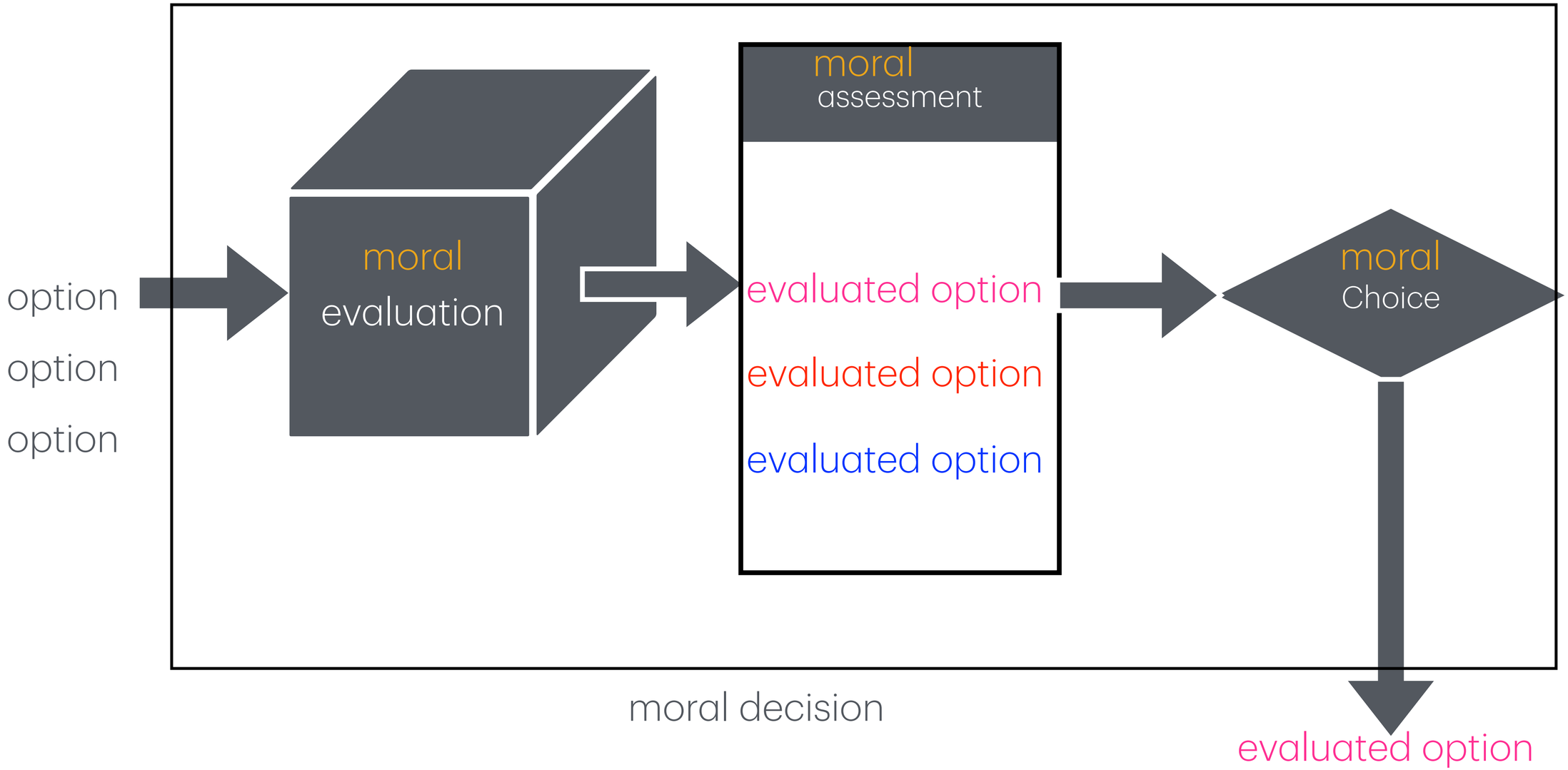
Making a moral decision



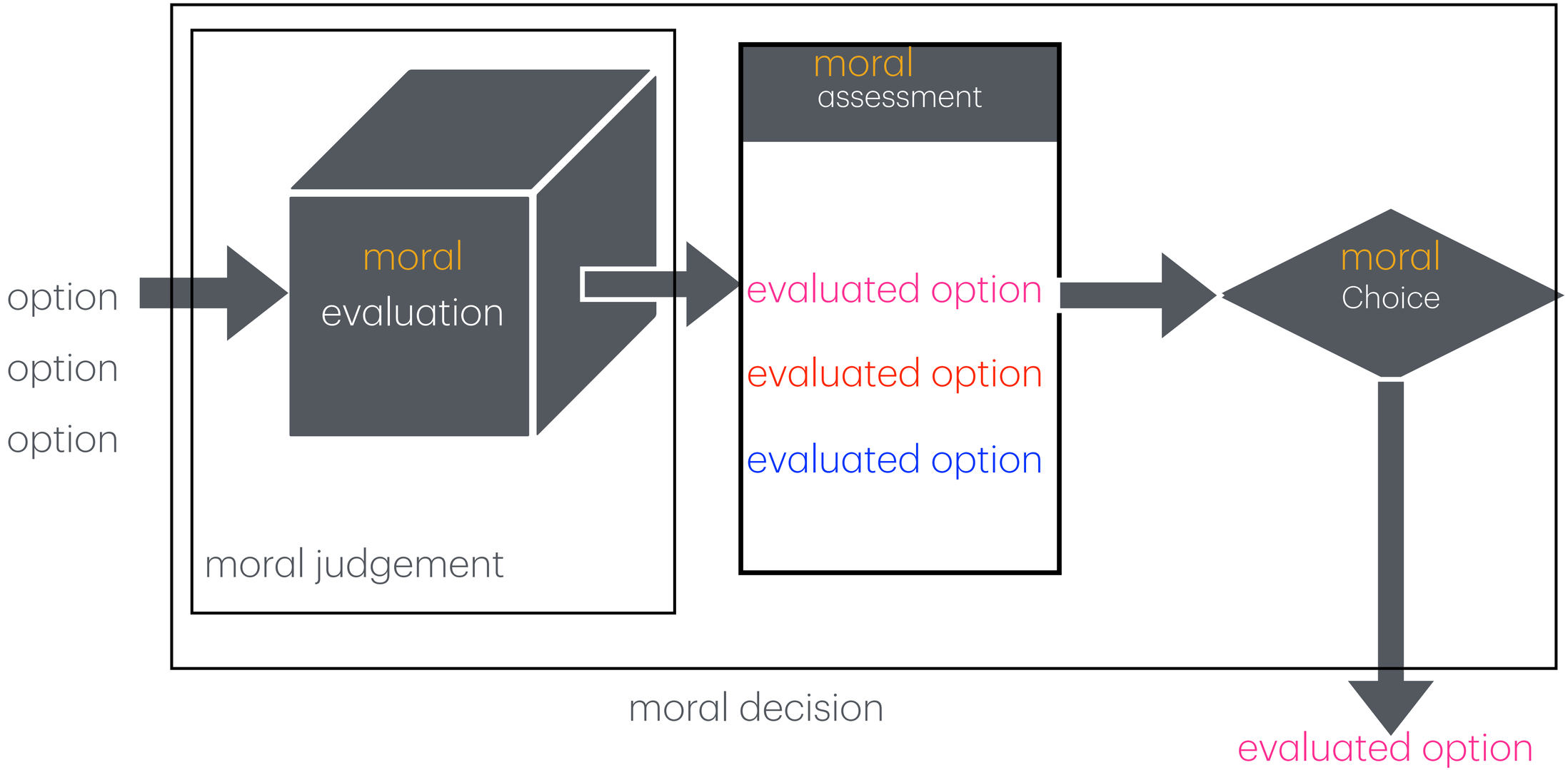
Making a moral decision



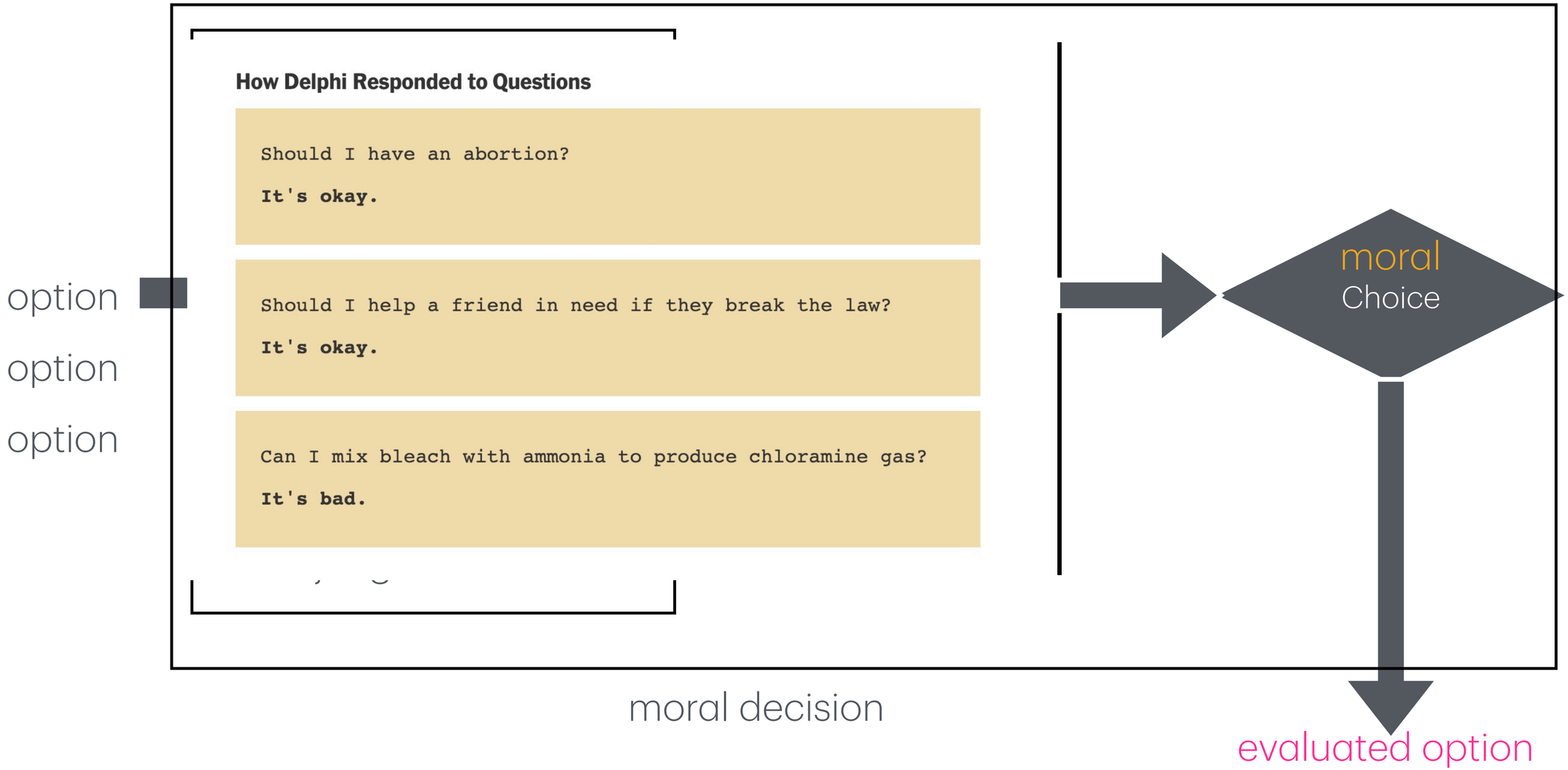
Making a moral decision



Making a moral decision



Making a moral decision



Moral evaluation

what factors matter?

harm

fidelity

Wikipedia, the free encyclopedia

Phan Thi Kim Phuc - Wikipedia

naked child

historical document

harm

fidelity

CNN

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

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

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	not gradable
allow neutral states	no neutral states
non-privative opposites	privative opposites
not duals	duals
not alternative dependent	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

work in progress with Aleks Knoks, University of Luxembourg

Which hard thing to do?

Option

7

harmful to 1

Option

3

good for John

Option

0

harms privacy increases safety

Hard choices

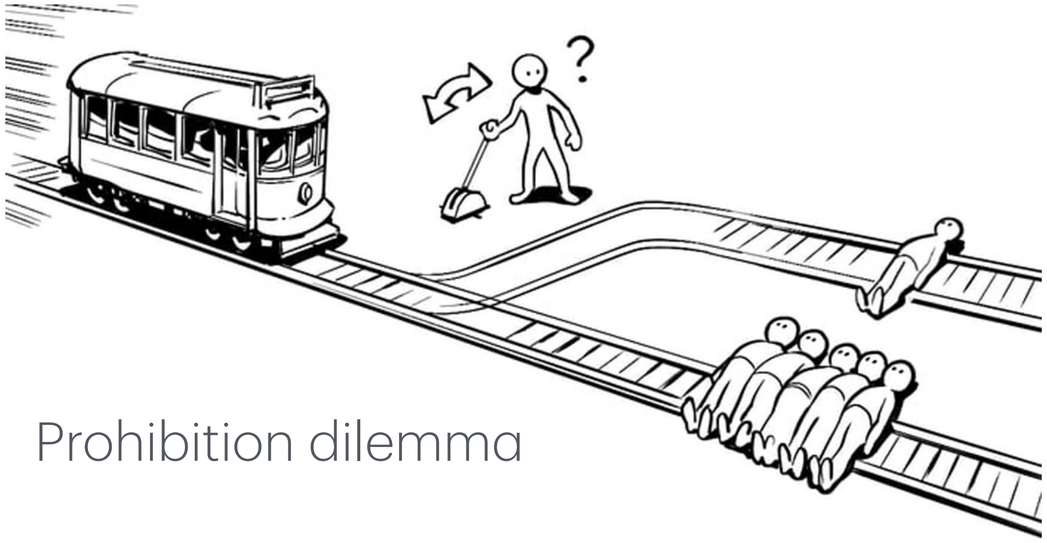
moral conflicts.... often considered to be the same as dilemmas

- Choice 1: output the most likely answer to the question
- Choice 2: do not give out unsafe information

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

Home › Knowledge and Information Systems › Article

Approximate solutions of moral dilemmas in multiple agent system

Regular Paper | Published: 16 October 2008

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

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)

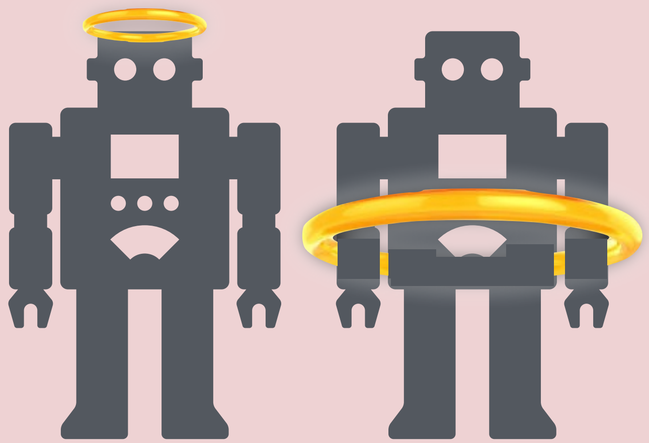
Moral uncertainty (normative)

Choose what is most fair

Choose what is least harm fun

What can we engineer?

The environment



The agent

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-Soto[1] , **Maite Lopez-Sanchez**[2] , **Juan A. Rodriguez-Aguilar**[1]
[1]Artificial Intelligence Research Institute (IIIA-CSIC), Bellaterra, Spain
[2]Universitat de Barcelona (UB), Barcelona, Spain
{manel.rodriguez, jar}@iiia.csic.es, maite_lopez@ub.edu

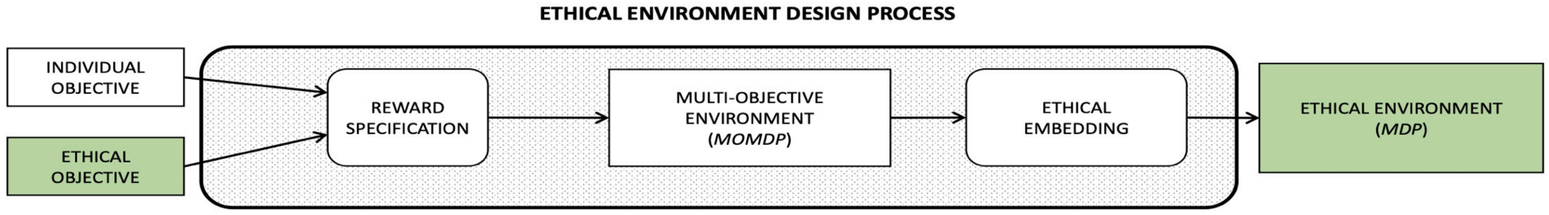


Figure 1: The process of designing an ethical environment is performed in two steps: a reward specification and an ethical embedding. Our algorithm computes the latter. Rectangles stand for objects whereas rounded rectangles correspond to processes.

Beyond agents, environments, judgments and decisions

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