What matters and how it matters: A choice-theoretic representation of moral theories

Franz Dietrich
Paris School of Economics and CNRS

Christian List
LSE

Cumberland Lodge
November 2017

Verification and Validation of Autonomous Systems: Ethical, Social and Trustworthy behaviour

The paper is forthcoming in *Philosophical Review*

Plan

Part 1: Motivation

Part 2: Options, contexts, rightness

Part 3: Ranking-based and reason-based explanations of rightness

Part 4: A taxonomy of moral theories

Part 5: Learning to be moral (for homans or robots)

Part 1:

Motivation

Capturing moral behaviour: rightness functions

- A 'rightness function' R maps each 'context' to the set of actions right/permissible in that context.
- 'Contexts' are situations where an agent (human or robot) must take an action.

Each moral theory has its own rightness function

- Example: For the utilitarian righness function, $R(K)$ consists of all feasible options in context K producing maximal total happiness.

A rightness function as the deontic content of a moral theory

Rightness functions vs. choice functions

Choice theorists use 'choice functions' instead of 'rightness functions'.

Formally the same object, except that

- $R(K)$ can be empty ('moral dilemmas')
- Our 'contexts' K can be richer than in choice theory.

Interpretively distinct:

- Choice functions are about actual (observed, real) choice
- rightness functions are about moral choice, i.e., how we or the the robot *should* act.

Could we simply tell the robot the rightness function which he then executes, full stop?

Explaining rightness

Choice theorists seek to *explain* or *represent* or *rationalize* a given rightness/choice function R:

- *Classical ranking-based explanation*: a fixed relation $\succsim$ that ranks all options, where the choice in any context K is given by $R(K) = \{\succsim\text{-best feasible options in } K\}$.
- *Many non-classical* explanations or models exist in choice theory.
- My favourite: *reason-based* explanation (Dietrich-List, 2016, 2017).

Part 2:

Options, contexts, rightness

Options

- X : an arbitrary set of *options*

Contexts

- $\mathcal{K}$: an arbitrary set of *contexts*
- What is a context?
  - In rational-choice-theory, it's merely a description of which options are 'feasible',
    $=>$ so it's a non-empty set of ('feasible') options $K \subseteq X$.
  - More generally, it could contain any additional information, such as room temperature, cultural environment, 'framing' information, or even information about the agent (e.g., his cultural identity).
    $=>$ e.g.: $K = (Y, t)$ where Y is the set of feasible options and t is room temperature, or cultural environment, or agent's cultural identity.

Contexts (cont.)

- Our contexts are fully general: they can be classical or richer.
- Formally: a context $K \in \mathcal{K}$ is something which induces a non-empty set of feasible options, denoted $[K] \subseteq X$
  - classical context: $[K] = K$
  - a non-classical example: $K = (Y, t)$ with feasible set $Y = [K]$ and additional information t.

Right choice

- Let $R : \mathcal{K} \longrightarrow 2^X$ be the *rightness function*, which for each context $K \in \mathcal{K}$ specifies the set $R(K) \subseteq [K]$ of permissible/right (feasible) options.
- This function captures all 'oughts', i.e., the deontic content of a moral theory.

Example: utilitarianism

- For the utilitarian righness function, we have

$$R(K) \;=\; \{x \in [K] : x \text{ gives at least as much overall happiness as each other } y \text{ in } [K]\}.$$

Part 3:

Ranking-based and reason-based explanation of moral choice

Ranking-based explanation

- A *ranking-based* or *classical* explanation or representation of the rightness function is a ('betterness') relation $\succsim$ on the set of options X such that in each context K the right options are $R(K) = \{x \in K : x \succsim y \text{ for all } y \in [K]\}$.

Two problems with ranking-based explanation

- **Problem 1:** inconsistent with many moral theories/rightness functions (e.g., context-dependent theories)
  Problem 2: uninformative, i.e., silent on the 'why'
  –> represenats essentially just the theory's deontic content, not the full theory (which also includes justifications and reasons)

Now: reason-based explanation

Properties: informally

- Which options are right depends on *properties*, i.e., properties of the options and/or the context, or formally, of option-context pairs.
- For instance, choosing x in context K might be right because of the (option) property of x that a life is saved, and the (context) property of K that there is no feasible option in which more than one life is saved.

Three types of properties: informally

- Pure **option properties** (act properties) pertain to the option only, e.g., the property that the act involves killing, or is expensive.
- Pure **context properties** pertain to the context only, e.g., the property that the cultural environment is traditional Indian, or that over 10 acts are feasible.
- **Relational** properties pertain to the relation between option and context, e.g., the property that the act has negative external effects in the context, or that the option is the most expensive one on offer.

Properties: formally

- An **option-context pair** is a pair (x, K) of an option x in X and a context K in $\mathcal{K}$.
  $->$ it represents the choice of x in context K.
- A **property** is something which is satisfied by certain option-context pairs (which *have* the property).
  - If you like, *identify* a property P with the set of option-context pairs having it: $P \subseteq X \times \mathcal{K}$.

Option properties defined

- A property P is a (pure) **option property** if

  (x, K) satisfies $P \Leftrightarrow (x, K')$ satisfies P (for all $x \in X$, $K, K' \in \mathcal{K}$
- An option property satisfied by (x, K) is simply called a property *'of x'*.

Context properties defined

- A property P is a (pure) **context property** if

  (x, K) satisfies $P \Leftrightarrow (x', K)$ satisfies P (for all $x, x' \in X$, $K \in \mathcal{K}$)
- A context property satisfied by (x, K) is simply called a property '*of* K'.

Relational properties defined

- A relational property is a property which is neither a (pure) option property nor a (pure) context property.

Notation

- $\mathcal{P}$: fixed set of *all* properties considered
- $\mathcal{P}(x, K)$: set of properties of (x, K) of any kind.
- $\mathcal{P}(x)$: set of option properties of x
- $\mathcal{P}(K)$: set of context properties of K

Righness can be explicated in terms of properties!

Two examples...

Utilitarianism

- For each $t \geq 0$ consider the (option!) property H_t of producing total happiness t.
- The utilitarian rightness function R picks the feasible option(s) whose happiness property H_t has highest t.

Another (stylised) example

- Consider the choice of a sweet from a basket of sweets served to the agent.
- So X consists of various sweets: dark Belgium chocolate, white Swiss chocolate, Austrian Mozart balls, American Mars, Snickers, ...
- A context is a non-empty set $K \subseteq X$ of sweets on offer; so $\mathcal{K} = 2^X \setminus \{\varnothing\}$.

Stylised example: the properties

We consider the following properties:

- $healthy$: the (option) property that the sweet is healthy
- $vulgar$: the (context) property that the basket contains a mars or snickers
- $polite$: the (relational) property that the sweet is not the only healthy one on offer (so can be chosen politely)

Stylised example: right choice

According to the moral theory:

- In non-vulgar contexts, one should choose a sweet that is polite (if available) and healthy (if available), where politeness has priority over health if the two can't be both achieved.
- In vulgar contexts, politeness no longer matters, so that one should simply choose a healthy sweet (if available).
- This defines a rightness function R (the formal details are obvious).

In both examples, righness is driven by properties.

But how exactly?

Reasons structures

A **reasons structure** is a pair $(N, \geq)$ containing:

- a function N, the **(normative) relevance function**, which assigns to each context $K \in \mathcal{K}$ a set of properties $N(K) \subseteq \mathcal{P}$, the **(normatively) relevant properties** in context K, such that normative relevance is determined by the context properties, i.e.,

$$\mathcal{P}(K) = \mathcal{P}(K') \Rightarrow N(K) = N(K') \text{ (for all } K, K' \in \mathcal{K}).$$

- a binary relation over property bundles $\geq$ ($\subseteq 2^{\mathcal{P}} \times 2^{\mathcal{P}}$), the **(normative) weighing relation**.

Reasons structures as formalized moral theories

Example: the utilitarian reasons structure

For classical utilitarianism,

- $N(K) = \{H_t : t \geq 0\}$, the set of happiness properties.
- $\{H_t\} \geq \{H_{t'}\} \Leftrightarrow t \geq t'$ ('more happiness is better')

Example: the reasons structure in the 'sweet example'

This example suggests the following reasons structure:

- $N(K) = \begin{cases} \{polite, healthy\} & \text{if } vulgar \notin \mathcal{P}(K) \\ \{healthy\} & \text{if } vulgar \in \mathcal{P}(\end{cases}$
- $\{polite, healthy\} > \{polite\} > \{healthy\} > \varnothing.$

Derivative notions

A reasons structure $\mathcal{R} \equiv (N, \geq)$ induces

(1) a **moral description** of options

(2) a notion of **rightness**.

Details on next slides!

(1) Moral description of options

- Option x **as described morally** in context K is the set

$$N(x, K) := \mathcal{P}(x, K) \cap N(K)$$

of normatively relevant properties of x in context K.

(2) Right choice

- The reasons structure implies the R rightness function which prescribes choosing an option with best normatively relevant properties:

$$R(K) := \{x \in K : N(x, K) \geq N(y, K) \text{ for all } y \in K\}.$$

- This righness function R is said to be **explained** or **represented** by the reasons structure $\mathcal{R}$.

The utilitarian example again

The above 'utilitarian reasons structure' indeed explains the utilitarian rightness function.

To see why, note that any option x is morally described in any context K by its unique happiness property H_t:

$$N(x, k) = \{H_t\}.$$

The 'sweet example' again

- Again, the above reasons structure indeed explains the intended rightness function.
- To see why, note that:
  - in non-vulgar contexts, any sweet is normatively described as one of four possible property combinations:

$$\{polite, healthy\},\ \{polite\},\ \{healthy\},\ \varnothing,$$

  - in vulgar contexts, the normative description of a sweet is either $\{healthy\}$ or $\varnothing$.

Two ways the context may matter

A reasons structure $(N, \geq)$ (or the moral theory it represents) displays

- **context-relevance or non-consequentialism** if at least one $N(K)$ contains not just option properties (i.e., context or relational properties),
- **context-variance or relativism** if $N(K)$ is not the same for all contexts $K \in \mathcal{K}$.

The ‘sweet example’

Here the moral theory (reasons structure) is doubly context-dependent: it is

- non-consequentialist as the relational property *polite* is sometimes relevant,
- relativist as $N(K)$ varies with the context K.

Context-dependence summarized

the context...	**metaethical meaning**
matters	non-consequentialism
affects what matters	relativism

Reason-based explicability

- In 4 theorems, we characterize axiomatically the choice functions that are:
  (1) reason-based explicable
  (2) reason-based explicable in a universalist way
  (3) reason-based explicable in a consequentialist way
  (4) reason-based explicable in a universalist and consequentialist way.

Part 4:

A taxonomy of moral theories

Metaethical positions formalized

Prominent metaethical positions can be represented:

metaethical position	**formalised as...**
consequentialism or non-consequentialism?	context-irrelevance?
universalism or relativism?	context-invariance?
monism or pluralism?	one or many properties in $N(x, K)$?
teleology or non-teleology?	$\geq$ transitive or not?
atomism or holism?	$\geq$ separable or not?

Agent-relativity: a form of non-conseqntialism or relativism?

- We can model both forms of agent-relativity, by building the agent's identity into the context, and
  - *either* letting the agent's identity matter,
  - *or* letting it affect what matters.

Part 5:

Learning to be moral (for humans or robots)

How learn the rightness function?

Two approaches for how an agent (e.g. a robot) can come to know the rightness function R:

- The agent is told R or an underlying theory/explanation of R (e.g., a ranking $\succsim$ or a reasons structure $(N, \geq)$).
- The agent learns R bit by bit.

Let's focus on the second, learning-based approach.

A simple model of learning

Step 1: The agent obtains a restricted rightness function R^* on a subdomain of contexts $\mathcal{K}^* \subseteq \mathcal{K}$.

- Think of R^* as the restriction of the unknown R to $\mathcal{K}^*$.
- Possible sources of R^*:
  - explicitly programmed
  - learnt from positive and negative feedbacks after acting in contexts in $\mathcal{K}^*$.

A simple model of learning

Step 2: The agent/robot builds a theory of R^*, i.e.:

- either a ranking $\succsim^*$ of the options occurring in contexts in $\mathcal{K}^*$.
  - Problem: often no $\succsim^*$ exists for R^*
- or a reasons structure $(N^*, \succsim^*)$ with for the domain $\mathcal{K}^*$ rather than $\mathcal{K}$.
  - Problem: often more than one such reasons structure exists ('moral underdetermination problem').

A simple model of learning

Step 3: Form new moral judgments, i.e., use the theory to extend R^* to a larger domain of contexts $\mathcal{R}^{**} \supseteq \mathcal{R}$.

- Problem if the theory is a ranking $\succsim^*$: we won't be able to extend R^* much (no non-trivial new judgments!)
- But a reason-based theory lends itself to beautiful extensions :-).
  - These extensions depend on the theory $(N^*, \succsim^*)$ used to explain R^*, which reinforces the moral underdetermination problem.

Complementary slides

(Non-)consequentialism formalized

- **Consequentialism** claims that moral rightness of actions only depends on consequences of actions.
- **Non-consequentialism** denies this, claiming that the choice process/context may also matter.

So (if options are identified with full consequences[1]):

(1) Consequentialism implies context-irrelevance ('the context never matters').

(2) Non-consequentialism implies context-variance ('the context may matter').

[1]This is a common assumption on the specification of options. More generally, (1) only assumes that options are specified richly enough for consequences to depend only on options, and (2) only assumes that options haven't been specified so richly as to carry information beyond consequences.

Universalism and relativism formalized

- **Relativism** claims that the moral rightness of actions may depend on the context (culture, time, ...)
  - Certain forms of politeness may be required in India, but not in Europe.
- **Universalism** denies this.

So:

(1) Relativism implies context-variance ('the context determines what matters').

(2) Universalism implies context-invariance ('what matters is fixed').

Monism and pluralism formalized

- **Monism** claims that the moral value of actions is constituted by a single kind of thing.
  - Classical utilitarianism is monist: only 'happiness' counts.
- **Pluralism** claims that moral value is constitutively plural.
  - e.g., satisfaction, happiness, meaning in life, etc. all count – as ends, i.e., for instance not because they all contribute to something else (such as happiness).

So (as long as by 'relevant' property we mean 'ultimately relevant' rather than 'instrumentally relevant' property[2]):

[2]Relevant properties can be described at different levels, leading to different M functions and hence different moral descriptions of alternatives x_K. Consider for instance classical utilitarianism. At the ultimate level, only happiness counts, so that $M(K)$ contains happiness-level properties only. At an instrumental level, $M(K)$ contains all sorts of other properties – about daily life activities, health, etc. – as such properties are instrumentally relevant in the production of happiness.

a context K) by a *singleton* property set $N(x, K) = \{P\}$.

(2) Pluralism implies that $N(x, K)$ can be a many-element set $\{P, Q, ...\}$.