Value Awareness & Multiagent Systems

Nardine Osman
Artificial Intelligence Research Institute (IIIA-CSIC)

Conexión AIHUB Escuela 2022
Palma de Mallorca
July 4, 2022

Collaborators

Nardine Osman

Carles Sierra

Juan Antonio Rodriguez

Maite López

Pablo Noriega

Nieves Montes
PhD student

Marc Serramià
Ex-PhD student

selected members of IIIA's "Ethics & AI" research theme

www.iiia.csic.es/en-us/research/themes/ethics-ai/

Outline

- Foundations
  - Understanding Norms
  - Understanding Values
  - Understanding the Norm–Value Relationship
- Vision & Motivation
- Value Awareness
  - A Selection of Models & Mechanisms

Outline

- Foundations
  - Understanding Norms
  - Understanding Values
  - Understanding the Norm–Value Relationship
- Vision & Motivation
- Value Awareness
  - A Selection of Models & Mechanisms

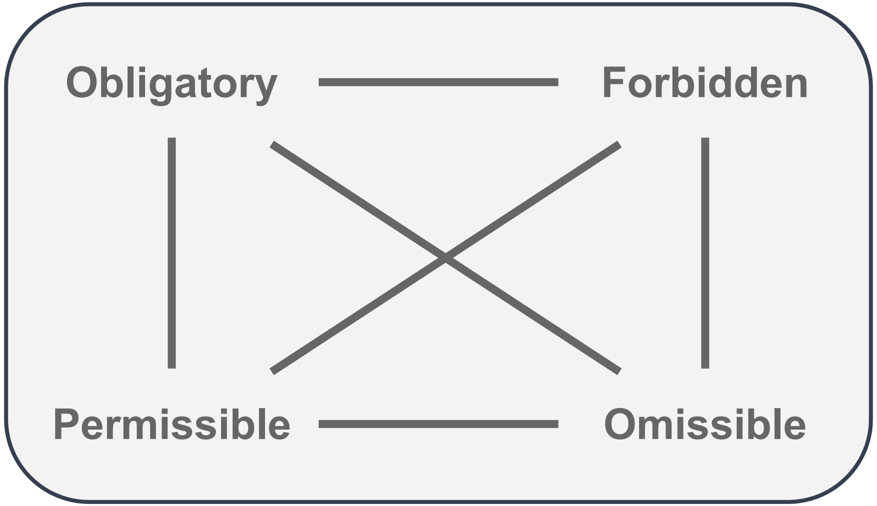
Norms in Multiagent Systems

Norms, an Overview

Norms are what **govern / regulate behaviour**

Norm Representation

Mostly based on **deontic concepts**



Norm Representation

Mostly based on **deontic concepts**

- **If-Then rules**
  - SIMPLE Language
- **Deontic Logic**
  - Conditional Deontic Logic with Deadlines
- **Event Calculus**
- **Expectations & Constraints**
  - Social Integrity Constraints
- **Commitments**
  - Object Constraint Language
- **Temporal Logic**
  - Hybrid Metric Interval Temporal Logic
  - Normative temporal logic (NTL)

Norm Representation

Mostly based on **deontic logic**

- **If-Then rules**
  - SIMPLE Language
- **Deontic Logic**
  - Conditional Deontic Logic with Deadlines
- **Event Calculus**
- **Expectations & Constraints**
  - Social Integrity Constraints
- **Commitments**
  - Object Constraint Language
- **Temporal Logic**
  - Hybrid Metric Interval Temporal Logic
  - Normative temporal logic (NTL)

If the auctioneer has announced the current price and no buyer has said 'mine!', then the auctioneer can say 'next!'.

de Jonge et al. (2016)

Norm Representation

Mostly based on **deontic logic**

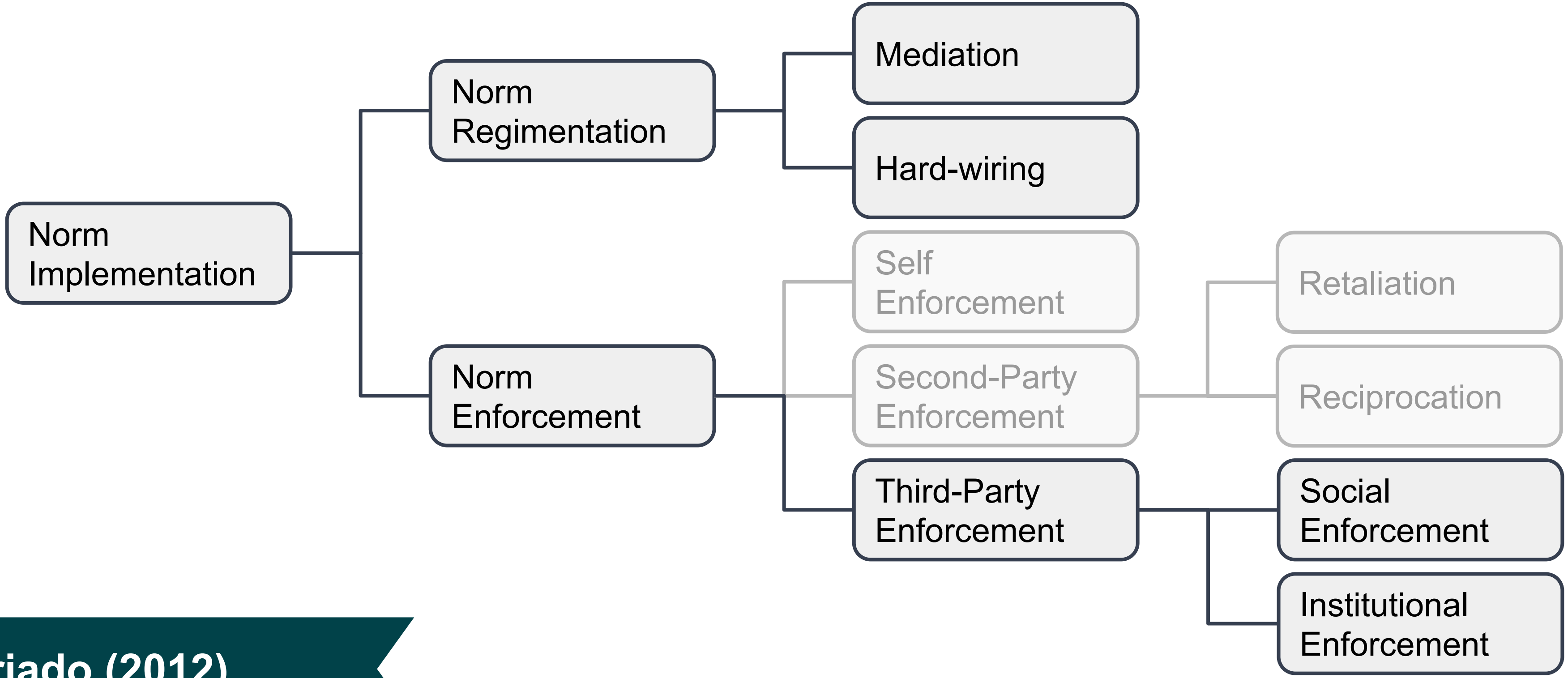
- **If-Then rules**
  - SIMPLE Language
- **Deontic Logic**
  - Conditional Deontic Logic with Deadlines
- **Event Calculus**
- **Expectations & Constraints**
  - Social Integrity Constraints
- **Commitments**
  - Object Constraint Language
- **Temporal Logic**
  - Hybrid Metric Interval Temporal Logic
  - Normative temporal logic (NTL)

```
PERMITTED(
        (user DO appoint(regular user))
         IF
         (access level(user, register,
                'full control')))

OBLIGED(
        (buyer DO bid(product,price))
         BEFORE
         (buyer DO exit(auction house)))
```

Vázquez-Salceda et al. (2004)

Norm Implementation



Criado (2012)

Norm Reasoning

- **Norm diagnosis.** Check and verify properties of norms.
- **Conflict resolution.** Check for inconsistencies.
- **Norm compliance.** Assess consequences of obeying norms.

Norm Creation

- **Top-Down Approaches**
  - **Offline design**
  - **Online norm synthesis**
    - driven by conflict detection

- **Bottom-Up Approaches**
  - **Norm Emergence:** usually focuses on internalisation of norms
  - **Norm Agreement**

Norm emergence triggers top-down norm creation

Take Home Message #1

Norms Guide Behaviour

Norms, usually specified as **deontic concepts**,
are used to mediate behaviour.

Norm **compliance** is ensured/motivated with
regimentation/enforcement techniques.

Hot topics in **AI & Ethics** are
value-driven norm assessment & creation/selection.

Understanding Values

Understanding Values

Values in the Social Sciences

Values in the Social Sciences

Why the interest in values?

“ Theorists have long considered values central to **understanding social behaviour** (e.g. Allport et al, 1960; Kluckhohn, 1951; Rokeach, 1973; Williams, 1968). This is because they view values as deeply rooted, abstract motivations that **guide, justify, and explain attitudes, norms, opinions, and actions** (Feather, 1985; Halman and de Moor, 1994; Rokeach, 1973; Schwartz, 1992). ”

Schwartz (2007)

Values in the Social Sciences

What are values?

Lewin (1952, p. 41). “Values influence behavior but have not the character of a goal (i.e., of a force field)... the individual does not try to ‘reach’ the value of fairness, but fairness is ‘guiding’ his behavior... values are not force fields but they “induce” force fields.”

Guth & Tagiuri (1965, p.124-125). “A value can be viewed as a conception, explicit or implicit, of what an individual or a group regards as desirable, and in terms of which he or they select, from among alternative available modes, the means and ends of action”.

Hutcheon (1972, p. 184). “... values are not the same as ideals, norms, desired objects, or espoused beliefs about the 'good', but are, instead, operating criteria for action...”.

Rokeach (1973, p. 5). “A value is an enduring belief that a specific mode of conduct or end-state of existence is personally or socially preferable to an opposite or converse mode of conduct or end-state of existence”.

Schwartz (1994, p.20). A value is “a belief pertaining to desirable end states or modes of conduct that transcends specific situations; guides selection or evaluation of behavior, people, and events; and is ordered by the importance relative to other values to form a system of value priorities”.

Feather (1996, p. 222). “I regard values as beliefs about desirable or undesirable ways of behaving or about the desirability or otherwise of general goals.”

Braithwaite & Blamey (1998, p.364). “Values...are principles for action encompassing abstract goals in life and modes of conduct that an individual or a collective considers preferable across contexts and situations”.

Friedman et al. (2006, p. 349). “A value refers to what a person or group of people

Van de Poel & Royakkers (2011, p. 72). Values are “lasting convictions or matter
not just for themselves to be able to lead a good life or realize a good society”

Cheng & Fleischmann (2010), Rohan (2000)

Values in the Social Sciences

What are values?

Lewin (1952, p. 41). “Values influence behavior but have not the character of a goal (i.e., of a force field)... the individual does not try to ‘reach’ the value of fairness, but fairness is ‘guiding’ his behavior... values are not force fields but they ‘induce’ force fields.”

Guth & Tagiuri (1965, p.124-125). “A value can be viewed as a conception, explicit or implicit, of what an individual or a group regards as desirable, and in terms of which he or they select, from among alternative available modes, the means and ends of action”.

Hutcheon (1972, p. 184). “... values are not the same as ideals, norms, desired objects, or espoused beliefs about the 'good', but are, instead, operating criteria for action...”.

Rokeach (1973, p. 5). “A value is an enduring belief that a specific mode of conduct or end-state of existence is personally or socially preferable to an opposite or converse mode of conduct or end-state of existence”.

Schwartz (1994, p.20). A value is “a belief pertaining to desirable end states or modes of conduct that transcends specific situations; guides selection or evaluation of behavior, people, and events; and is ordered by the importance relative to other values to form a system of value priorities”.

Feather (1996, p. 222). “I regard values as beliefs about desirable or undesirable ways of behaving or about the desirability or otherwise of general goals.”

Braithwaite & Blamey (1998, p.364). “Values...are principles for action encompassing abstract goals in life and modes of conduct that an individual or a collective considers preferable across contexts and situations”.

Friedman et al. (2006, p. 349). “A value refers to what a person or group of people consider important in life”.

van de Poel & Royakkers (2011, p. 72). Values are “lasting convictions or matters that people feel should be strived for in general and not just for themselves to be able to lead a good life or realize a good society”

What are values?

We adopt Schwartz's view of values.

"(1) **Values are beliefs** linked inextricably to affect.
(2) **Values refer to desirable goals** that motivate action.
(3) **Values transcend specific actions and situations**.
(4) **Values serve as standards or criteria**.
(5) **Values are ordered by importance** relative to one another.
(6) **The relative importance of multiple values guides action**."

Schwartz's Theory of Basic Human Values
Schwartz (2012)

Values in the Social Sciences

Value Categories

Schwartz's theory of basic human values

He conducted value surveys in 20 countries, resulting in a culturally universal conceptual framework for values, which is composed of 56 different values falling into 10 general values, which may be organised into 4 groups.

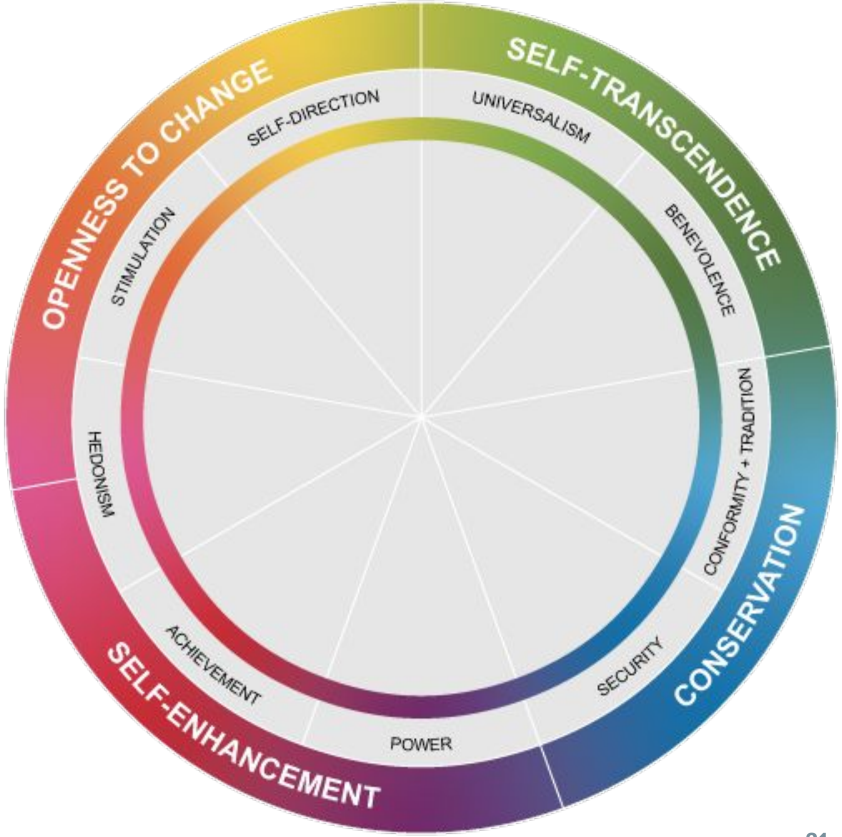


Image source: https://www.open.edu/openlearncreate/mod/oucontent/view.php?id=143807&section=1.2

Value Categories in technology design

Friedman focused on values implicated in technology design, based on conceptual, empirical, and technical investigations.

Key values identified:

(1) human welfare
(2) ownership and property
(3) privacy
(4) freedom from bias
(5) universal usability
(6) trust
(7) autonomy
(8) informed consent
(9) accountability
(10) courtesy
(11) identity
(12) calmness
(13) environmental sustainability

Value Categories in technology design

> [According to Davis & Nathan (2015),] values should be more open-ended and should bottom-up elicit values from stakeholders.
>
> van de Poel (2021)

> values themselves may be subject to change ***during*** the lifetime of a product
>
> van de Poel (2021)

Understanding Values

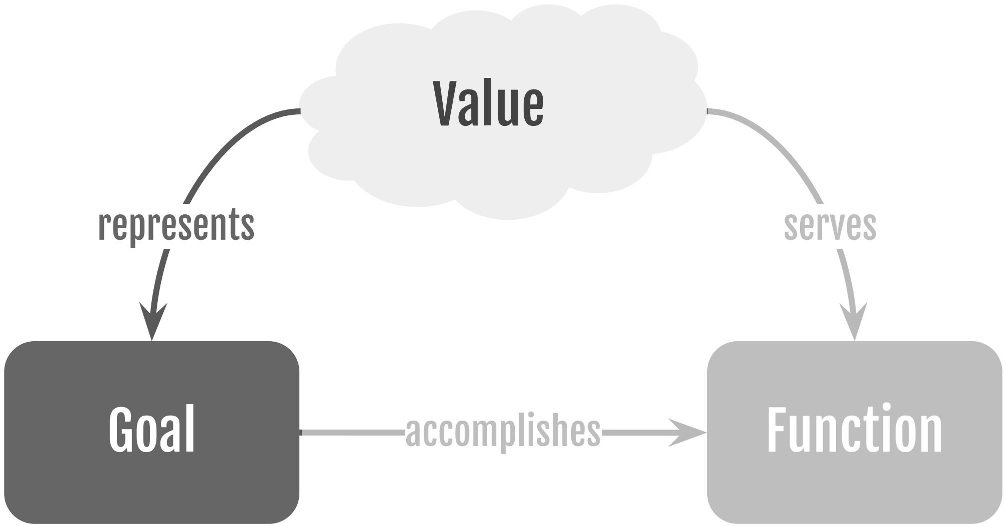
Values as Formal Objects

Values, goals, and requirements

"… values … are grounded in one or more of three universal **requirements** of human existence … These requirements are needs of individuals as biological organisms, requisites of coordinated social interaction, and survival and welfare needs of groups. Individuals cannot cope successfully with these requirements of human existence on their own. Rather, people must **articulate appropriate goals** to cope with them, communicate with others about them, and gain cooperation in their pursuit. **Values** are the socially desirable **concepts** used to **represent these goals** mentally and the vocabulary used to **express them in social interaction**."

Schwartz's Theory of Basic Human Values
Schwartz (2012)

Values, goals, and requirements



- Biological needs
- Social coordination
- Group survival & welfare

Values as Goals

Proposal.

Abstract values are grounded into permanent goals, that agents actively pursue.

Not all Goals are Equal

So from a computational perspective, what is the difference between goals that ground values and traditional AI goals?

- Permanency
- Degree of satisfaction

Examples of Value-Grounding Goals

Gender Equality:

- women get equal pay to men with the same job
- equal access to education
- equal maternity and paternity rights
- …

Democracy:

- elected representatives determine government policy
- transparent financing of political parties
- no restrictions on internet access
- …

Preferences over Values

Value-grounding goals must be prioritised from the most esteemed to the least important.

"(1) **Values are beliefs** linked inextricably to affect.
(2) **Values refer to desirable goals** that motivate action.
(3) **Values transcend specific actions and situations**.
(4) **Values serve as standards or criteria**.
(5) **Values are ordered by importance** relative to one another.
(6) **The relative importance of multiple values guides action**."

Schwartz's Theory of Basic Human Values
Schwartz (2012)

Preferences help address conflicts between values!

Take Home Message #2

Values as Formal Objects

We propose values to be grounded into **permanent goals**.

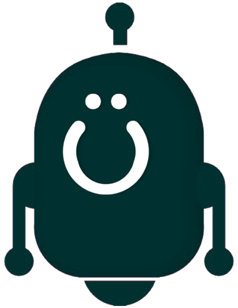
Preferences over values need to be specified, to help with conflicts.

Dynamic nature of values must be acknowledged and accounted for.

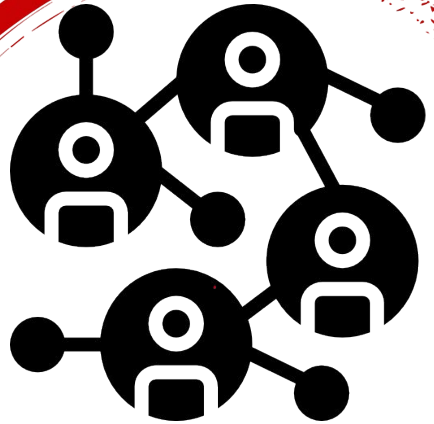
The Norm–Value Relationship

Implications of Value-Grounding Goals

So what are the computational implications of having values grounded into goals?



Supports decision making over agent actions



Support norm selection/generation

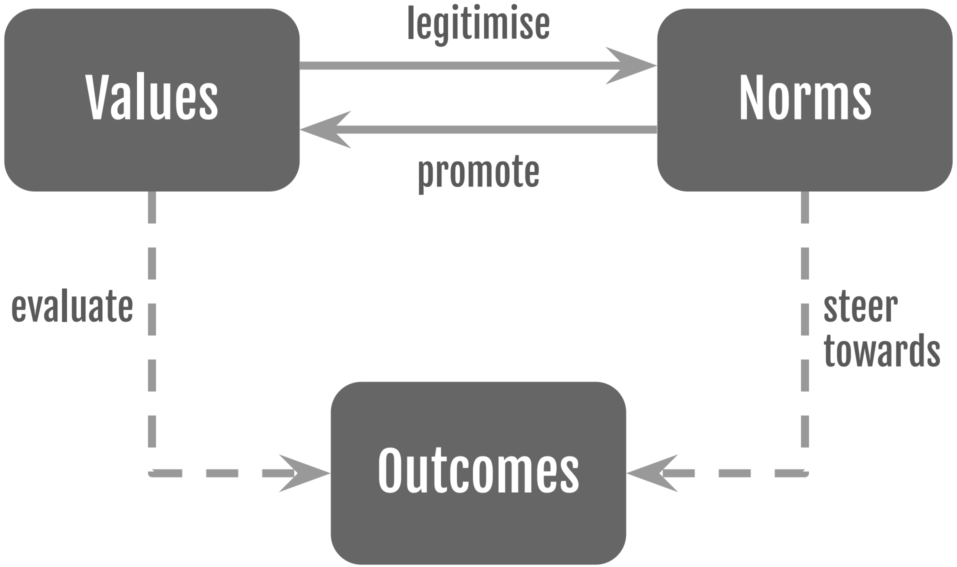
The Norm–Value Relationship

Norms govern behaviour.

Values are grounded through goals.

When a norm facilitates the achievement of
goals that ground the meaning of a value,
we say **the norm is aligned with respect to that value**.

The Norm–Value Relationship



Take Home Message #3

The Norm–Value Relationship

We say **a norm is aligned with respect to a value** if it facilitates the achievement of goals that ground the meaning of that value.

Outline

- Foundations
  - Understanding Norms
  - Understanding Values
  - Understanding the Norm–Value Relationship
- Vision & Motivation
- Value Awareness
  - A Selection of Models & Mechanisms

Vision & Motivation

Vision

Value-aware AI

Noun [U]

/'væl.ju: ə'weə^{r} eɪ aɪ/

an AI system that <u>understands</u> and <u>abides</u> by a value system, <u>explains</u> its own behaviour and that of others in terms of that value system.

Understands a value system

➥ value representation & reasoning

Abides by a value system

➥ value-alignment mechanisms

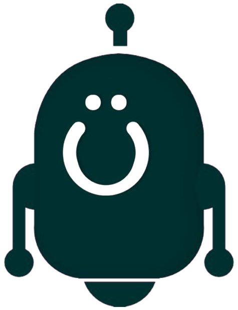
Explains behaviour in terms of a value system

➥ value-based explainability

Vision

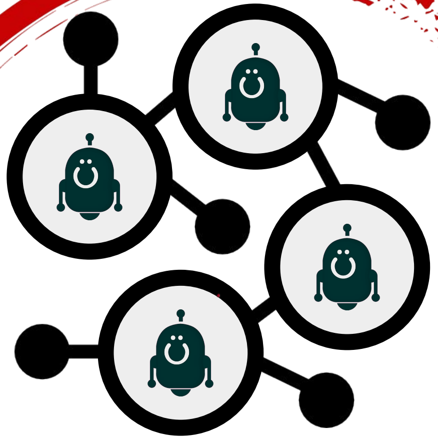
Value-aware AI systems

an individual agent



behaviour shaped by decision making
[value-driven decision making]

a system of interacting agents



behaviour shaped by norms
[value-aligned norms]

Motivation

- Ethical AI [value-aware and value-aligned AI : values are engineered]
- Trustworthy AI

Outline

- Foundations
  - Understanding Norms
  - Understanding Values
  - Understanding the Norm–Value Relationship
- Vision & Motivation
- **Value Awareness**
  - A Selection of Models & Mechanisms

Value Awareness: Models & Mechanisms

Value Alignment

Sierra et al. (2019)

Values as Preferences

$S \prec S'$

Male's salary > Female's salary

Male's salary = Female's salary

Values are understood as preferences over behaviour, or preferences over the states of the world.

Values as Preferences

Values are understood as preferences over behaviour, or preferences over the states of the world.

We define a **value-based preference** over states of the world: $\mathrm{Prf}_{v}^{\alpha}(s,s')$

Defining Value-Based Preferences

A value-based preference depends on the **satisfaction of state properties** relevant to the value: Φ_v

Examples of state properties:

- No gender pay gap
- Equal rights to education
- Equal rights in marriage, divorce, & property/land ownership and inheritance
- ...

Defining Value-Based Preferences

S $\prec$ S'

Male's salary > Female's salary

Male's salary = Female's salary

A value-based preference depends on the **satisfaction of state properties** relevant to the value: Φ_v

$$\mathrm{Prf}_v(s,s') = f\left(\,P(s \vDash \Phi_v),\ P(s' \vDash \Phi_v)\,\right)$$

a computational approach

Sets of Values & Groups of People

What about preferences over **sets of values**
& for **groups of people**?

$$
\begin{tikzcd}
\mathrm{Prf}_{v}^{\alpha}(s,s') \arrow[r] \arrow[d] & \mathrm{Prf}_{v}^{G}(s,s') \arrow[d] \\
\mathrm{Prf}_{V}^{\alpha}(s,s') \arrow[r] & \mathrm{Prf}_{V}^{G}(s,s')
\end{tikzcd}
$$

$$\mathrm{Prf}_{V}^{\alpha}(s,s') = \frac{\sum_{v\in V} \mathrm{Prf}_{v}^{\alpha}(s,s')}{|V|}$$

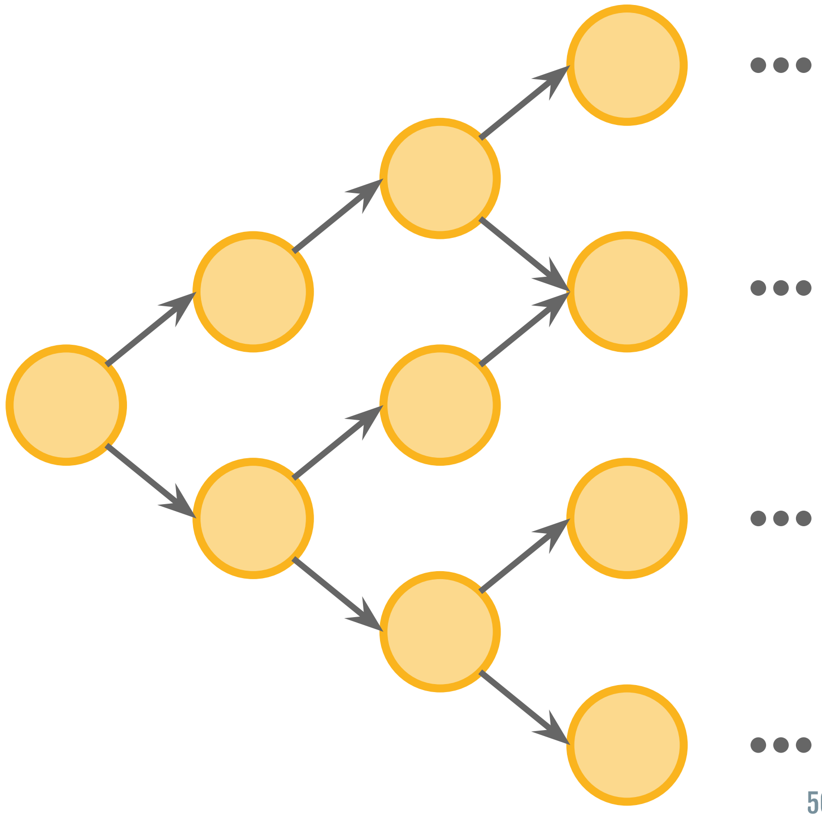
$$\mathrm{Prf}_{v}^{G}(s,s') = \frac{\sum_{\alpha\in G} \mathrm{Prf}_{v}^{\alpha}(s,s')}{|G|}$$

$$\mathrm{Prf}_{V}^{G}(s,s') = \frac{\sum_{v\in V} \mathrm{Prf}_{v}^{G}(s,s')}{|V|}$$

$$= \frac{\sum_{\alpha\in G} \mathrm{Prf}_{V}^{\alpha}(s,s')}{|G|}$$

Norms

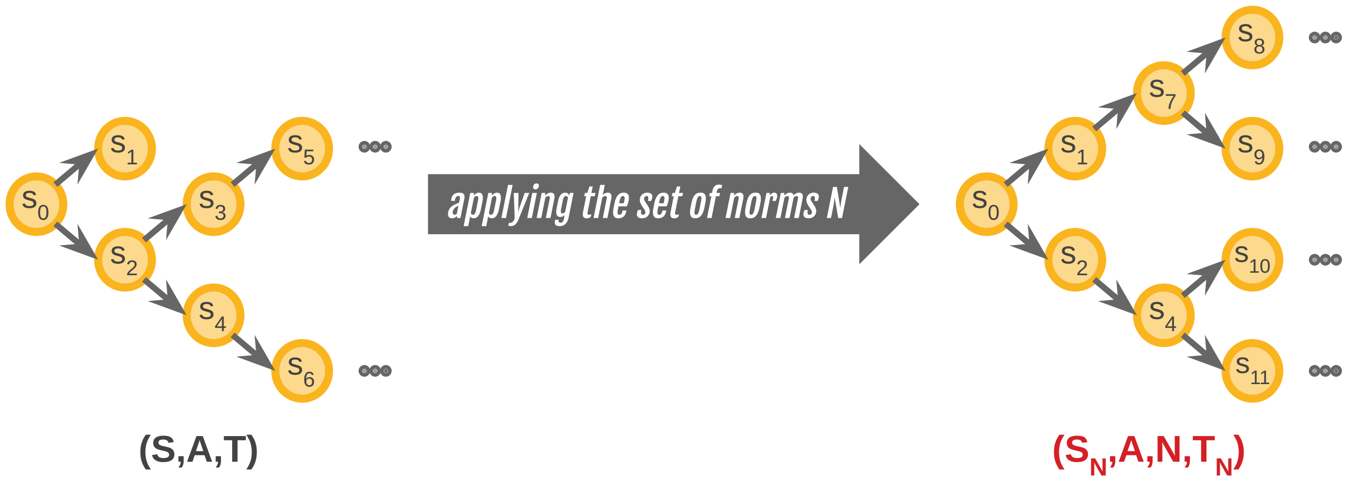
Norms define the possible worlds.



Norms

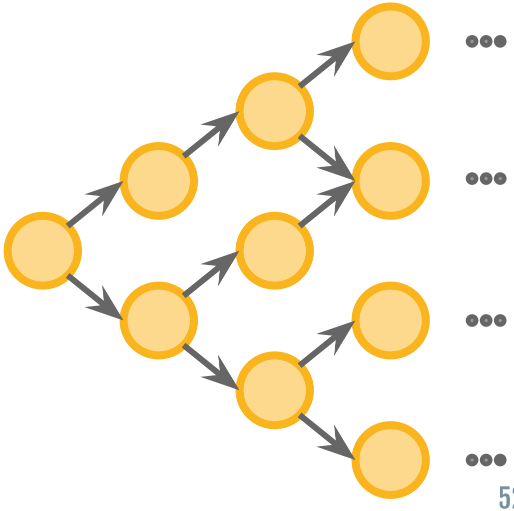
Norms define the possible worlds.

Changing the norms result in changing the possible worlds.



Value-Alignment

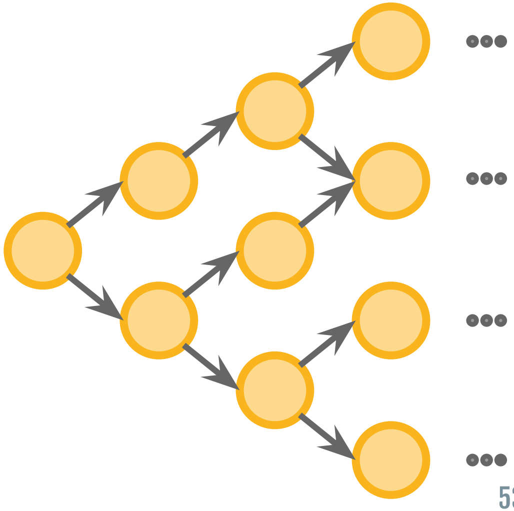
The degree of alignment of a norm n with a value v for agent α is the **accumulation of preferences** along the transitions.



Value-Alignment

The degree of alignment of a norm n with a value v for agent α is the **accumulation of preferences** along the transitions.

And we consider **all possible paths**.

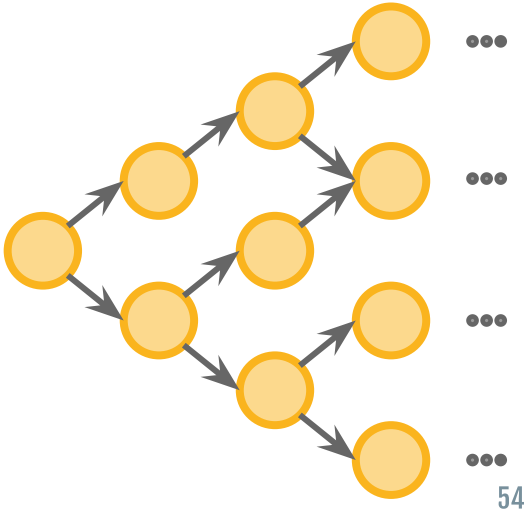


Value-Alignment

The degree of alignment of a norm n with a value v for agent α is the **accumulation of preferences** along the transitions.

And we consider **all possible paths**,
giving **equal weight** to all paths and all transitions.

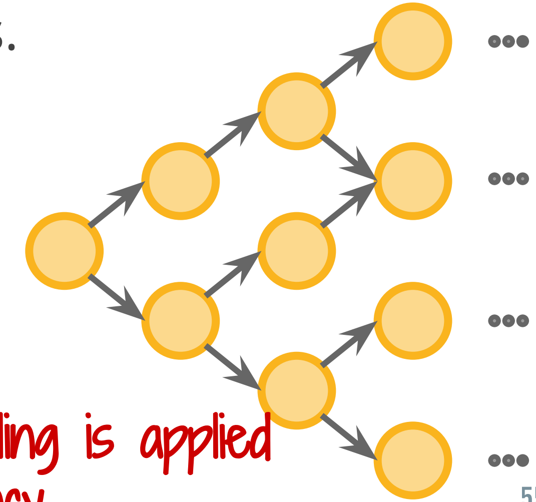
Big assumption!!!



Value-Alignment

The degree of alignment of a norm n with a value v for agent α is the **accumulation of preferences** along the transitions.

And we consider **all possible paths**,
giving **equal weight** to all paths and all transitions.



$$\mathsf{Algn}^{\alpha}_{n,v}(\mathcal{S},\mathcal{A},T) = \frac{\sum\limits_{p \in paths} \sum\limits_{d \in [1, length(p)]} \mathsf{Prf}^{\alpha}_{v}(p_I[d], p_F[d])}{\sum\limits_{p \in paths} length(p)}$$

Monte Carlo sampling is applied to address efficiency

Relative Alignment

The relative alignment of norm n_1 w.r.t. norm n_2
is defined as the **difference** in their **alignments**!

$$\mathsf{RAlgn}^{\alpha}_{n_1/n_2,v}(\mathcal{S},\mathcal{A},T) = \mathsf{Algn}^{\alpha}_{n_1,V}(\mathcal{S},\mathcal{A},T) - \mathsf{Algn}^{\alpha}_{n_2,V}(\mathcal{S},\mathcal{A},T)$$

Sets of Values & Groups of People

With the right **aggregation functions**, and just like preferences, we can talk about alignment / relative alignment over **sets of values** & for **groups of people**?

$$\begin{tikzcd} \mathrm{Prf}_{v}^{\alpha}(s,s') \arrow[r] \arrow[d] & \mathrm{Prf}_{v}^{G}(s,s') \arrow[d] \\ \mathrm{Prf}_{V}^{\alpha}(s,s') \arrow[r] & \mathrm{Prf}_{V}^{G}(s,s') \end{tikzcd}$$

$$\begin{tikzcd} \mathrm{Algn}_{v}^{\alpha}(S,A,T) \arrow[r] \arrow[d] & \mathrm{Algn}_{v}^{G}(S,A,T) \arrow[d] \\ \mathrm{Algn}_{V}^{\alpha}(S,A,T) \arrow[r] & \mathrm{Algn}_{V}^{G}(S,A,T) \end{tikzcd}$$

$$\begin{tikzcd} \mathrm{RAlgn}_{n/n',v}^{\alpha}(S,A,T) \arrow[r] \arrow[d] & \mathrm{RAlgn}_{n/n',v}^{G}(S,A,T) \arrow[d] \\ \mathrm{RAlgn}_{n/n',V}^{\alpha}(S,A,T) \arrow[r] & \mathrm{RAlgn}_{n/n',V}^{G}(S,A,T) \end{tikzcd}$$

Example: Prisoner's Dilemma

Agents' actions (cooperate (c) & defect (d)) results in certain gains.

Let the relevant state parameters describe accumulated gains: (x,y)

	β co-operates	β defects
α co-operates	6,6	0,9
α defects	9,0	3,3

Example: Prisoner's Dilemma

Value-based preferences.

❶ States with higher **equality** in accumulated gain are preferred.

❷ States with higher **equality** in accumulated gain are preferred, **only if my personal gain is not lower**.

❸ States with higher **personal gain** are preferred, **only if equality is not lower**.

❹ States with higher **personal gain** are preferred.

Example: Prisoner's Dilemma

Value-based preferences.

❶ $$\mathsf{Prf}(s,s') = \frac{|x-y|}{\max\{x,y\}} - \frac{|x'-y'|}{\max\{x',y'\}}$$

❷ $$\mathsf{Prf}(s,s') = \left(1 - \frac{|y'-x'|}{\max\{x',y'\}}\right) \cdot \frac{x'-x}{\max\{x',x\}}$$

❸ $$\mathsf{Prf}(s,s') = \frac{x'-x}{2(\max\{x',x\})} - \frac{y'-y}{2(\max\{y',y\})}$$

❹ $$\mathsf{Prf}(s,s') = \frac{x'-x}{\max\{x',x\}}$$

Example: Prisoner's Dilemma

Value-based preferences.

❶ States with higher **equality** in accumulated gain are preferred.

❷ States with higher **equality** in accumulated gain are preferred, **only if my personal gain is not lower**.

❸ States with higher **personal gain** are preferred, **only if equality is not lower**.

❹ States with higher **personal gain** are preferred.

Norms.

n_0 **No taxing:**
No taxes are to be payed.

n_1 **Incremental taxing:**
No taxes to be paid when the gain is 0 or 3, 3 to be paid as taxes when the gain is 6, & 5 to be paid as taxes when the gain is 9.

n_2 **Fixed taxing:**
1/3 of the gain is to be paid as taxes.

Example: Prisoner's Dilemma

Value-based preferences.

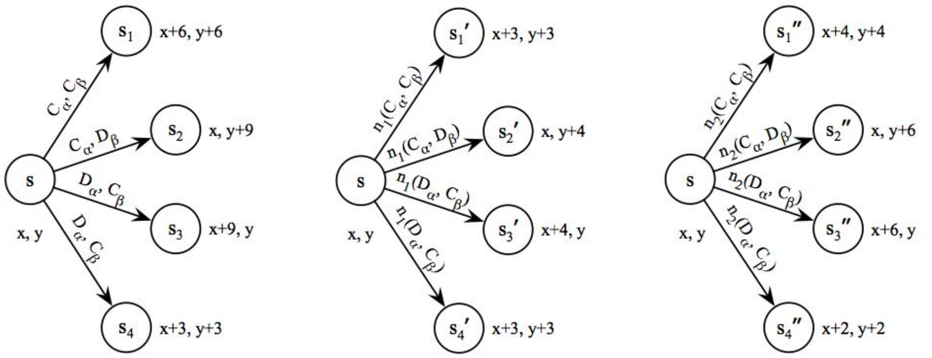
❶ States with higher **equality** in accumulated gain are preferred.

❷ States with higher **equality** in accumulated gain are preferred, **only if my personal gain is not lower**.

❸ States with higher **personal gain** are preferred, **only if equality is not lower**.

❹ States with higher **personal gain** are preferred.

Norms.



(a) n_0 applied (b) n_1 applied (c) n_2 applied

Example: Prisoner's Dilemma

Which norms are better aligned with an agent's interpretation of 'equality'?

3 norms: n_0, n_1, n_2

4 interpretations of 'equality': ❶, ❷, ❸, ❹

	α's actions	β's actions	Relative Alignments
❶	{c}	{c,d}	$n_1 \succ n_0 \sim n_2$
❷	{c}	{c,d}	$n_0 \sim n_1 \sim n_2$
❸	{c}	{c,d}	$n_0 \sim n_1 \sim n_2$
❹	{c}	{c,d}	$n_0 \succ n_2 \succ n_1$
❶	{d}	{c,d}	$n_1 \succ n_0 \sim n_2$
❷	{d}	{c,d}	$n_0 \sim n_1 \sim n_2$
❸	{d}	{c,d}	$n_0 \sim n_1 \sim n_2$
❹	{d}	{c,d}	$n_0 \sim n_1 \succ n_2$
❶	{c,d}	{c}	$n_1 \succ n_0 \sim n_2$
❷	{c,d}	{c}	$n_0 \sim n_1 \sim n_2$
❸	{c,d}	{c}	$n_0 \sim n_1 \sim n_2$
❹	{c,d}	{c}	$n_0 \sim n_1 \sim n_2$
❶	{c,d}	{d}	$n_1 \succ n_0 \sim n_2$
❷	{c,d}	{d}	$n_1 \succ n_0 \sim n_2$
❸	{c,d}	{d}	$n_1 \succ n_0 \sim n_2$
❹	{c,d}	{d}	$n_0 \sim n_1 \succ n_2$
❶❷❸	{c,d}	{c,d}	$n \sim n \sim n$

Example: Prisoner's Dilemma

Which norms are better aligned with an agent's interpretation of 'equality'?

3 norms: n_0, n_1, n_2

4 interpretations of 'equality': ❶, ❷, ❸, ❹

The norm better aligned with a strong support of equality (❶) is norm n_1.

	α's actions	β's actions	Relative Alignments
❶	{c}	{c,d}	$n_1 \succ n_0 \sim n_2$
❷	{c}	{c,d}	$n_0 \sim n_1 \sim n_2$
❸	{c}	{c,d}	$n_0 \sim n_1 \sim n_2$
❹	{c}	{c,d}	$n_0 \succ n_2 \succ n_1$
❶	{d}	{c,d}	$n_1 \succ n_0 \sim n_2$
❷	{d}	{c,d}	$n_0 \sim n_1 \sim n_2$
❸	{d}	{c,d}	$n_0 \sim n_1 \sim n_2$
❹	{d}	{c,d}	$n_0 \sim n_1 \succ n_2$
❶	{c,d}	{c}	$n_1 \succ n_0 \sim n_2$
❷	{c,d}	{c}	$n_0 \sim n_1 \sim n_2$
❸	{c,d}	{c}	$n_0 \sim n_1 \sim n_2$
❹	{c,d}	{c}	$n_0 \sim n_1 \sim n_2$
❶	{c,d}	{d}	$n_1 \succ n_0 \sim n_2$
❷	{c,d}	{d}	$n_1 \succ n_0 \sim n_2$
❸	{c,d}	{d}	$n_1 \succ n_0 \sim n_2$
❹	{c,d}	{d}	$n_0 \sim n_1 \succ n_2$
❶❷❸	{c,d}	{c,d}	$n \sim n \sim n$

Example: Prisoner's Dilemma

Which norms are better aligned with an agent's interpretation of 'equality'?

3 norms: n_0, n_1, n_2

4 interpretations of 'equality': ❶, ❷, ❸, ❹

All norms (n_0, n_1, n_2) are equally aligned for moderate supporters of equality (❷,❸).

	α's actions	β's actions	Relative Alignments
❶	{c}	{c,d}	$n_1 > n_0 \sim n_2$
❷	{c}	{c,d}	$n_0 \sim n_1 \sim n_2$
❸	{c}	{c,d}	$n_0 \sim n_1 \sim n_2$
❹	{c}	{c,d}	$n_0 > n_2 > n_1$
❶	{d}	{c,d}	$n_1 > n_0 \sim n_2$
❷	{d}	{c,d}	$n_0 \sim n_1 \sim n_2$
❸	{d}	{c,d}	$n_0 \sim n_1 \sim n_2$
❹	{d}	{c,d}	$n_0 \sim n_1 > n_2$
❶	{c,d}	{c}	$n_1 > n_0 \sim n_2$
❷	{c,d}	{c}	$n_0 \sim n_1 \sim n_2$
❸	{c,d}	{c}	$n_0 \sim n_1 \sim n_2$
❹	{c,d}	{c}	$n_0 \sim n_1 \sim n_2$
❶	{c,d}	{d}	$n_1 > n_0 \sim n_2$
❷	{c,d}	{d}	$n_1 > n_0 \sim n_2$
❸	{c,d}	{d}	$n_1 > n_0 \sim n_2$
❹	{c,d}	{d}	$n_0 \sim n_1 > n_2$
❶❷❸	{c,d}	{c,d}	$n \sim n \sim n$

Example: Prisoner's Dilemma

Which norms are better aligned with an agent's interpretation of 'equality'?

3 norms: n_0, n_1, n_2
4 interpretations of 'equality': ❶, ❷, ❸, ❹

When there is a random strategy for both agents, leading to an egalitarian society, all norms (n_0, n_1, n_2) are equally aligned for all the various supporters of equality (❶ , ❷ , ❸ , ❹).

	α's actions	β's actions	Relative Alignments
❶	{c}	{c,d}	$n_1 > n_0 \sim n_2$
❷	{c}	{c,d}	$n_0 \sim n_1 \sim n_2$
❸	{c}	{c,d}	$n_0 \sim n_1 \sim n_2$
❹	{c}	{c,d}	$n_0 > n_2 > n_1$
❶	{d}	{c,d}	$n_1 > n_0 \sim n_2$
❷	{d}	{c,d}	$n_0 \sim n_1 \sim n_2$
❸	{d}	{c,d}	$n_0 \sim n_1 \sim n_2$
❹	{d}	{c,d}	$n_0 \sim n_1 > n_2$
❶	{c,d}	{c}	$n_1 > n_0 \sim n_2$
❷	{c,d}	{c}	$n_0 \sim n_1 \sim n_2$
❸	{c,d}	{c}	$n_0 \sim n_1 \sim n_2$
❹	{c,d}	{c}	$n_0 \sim n_1 \sim n_2$
❶	{c,d}	{d}	$n_1 > n_0 \sim n_2$
❷	{c,d}	{d}	$n_1 > n_0 \sim n_2$
❸	{c,d}	{d}	$n_1 > n_0 \sim n_2$
❹	{c,d}	{d}	$n_0 \sim n_1 > n_2$
❶❷❸❹	{c,d}	{c,d}	$n \sim n \sim n$

Value Alignment

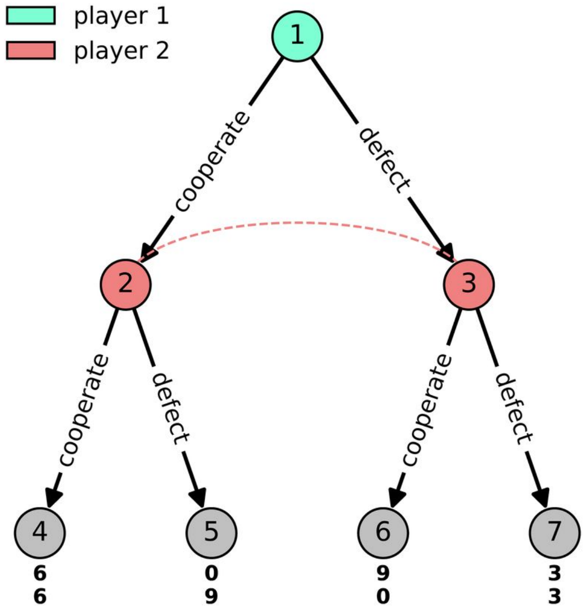
Montes & Sierra (2021 a)

Modelling Interactions

A Game Theoretic Approach.

Agents' interactions modelled as normal-form games.

Given a set of norms N governing a multiagent system, agents adopt a particular strategy to play the game.



Concepts from Game Theory

In a **nash equilibrium**,
no player has anything to gain
by changing only their own strategy.

In **pareto optimality**,
no player can improve their reward
without damaging someone else's.

Concepts from Game Theory

In a nash equilibrium,
no player has anything to gain
by changing only their own strategy.

In a **nash alignment equilibria**,
no player can improve its alignment
by changing only their own strategy.

In pareto optimality,
no player can improve their reward
without damaging someone else's.

Pareto optimal alignment
corresponds to situations where
no agent can improve its alignment
without damaging someone else's.

Value Aligned Agent Strategies

We can calculate which agent strategies lead to nash alignment equilibrium & pareto optimal alignment, for a given value.

Example: Prisoner's Dilemma

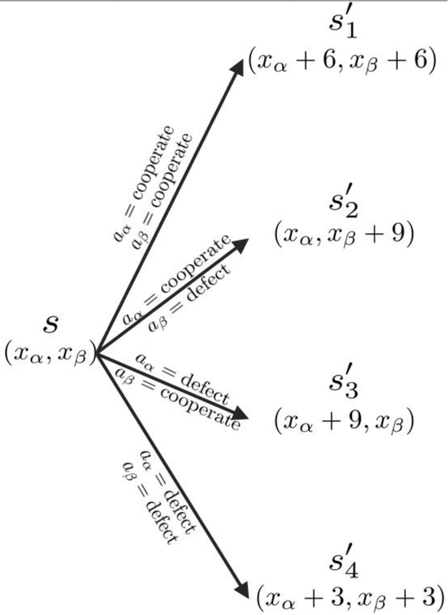
Norms. Sierra et al. (2019)'s game

	β co-operates	β defects
α co-operates	6,6	0,9
α defects	9,0	3,3

Example: Prisoner's Dilemma

Norms. Sierra et al. (2019)'s game

	β co-operates	β defects
α co-operates	6,6	0,9
α defects	9,0	3,3



Example: Prisoner's Dilemma

Norms. Sierra et al. (2019)'s game

	β co-operates	β defects
α co-operates	6,6	0,9
α defects	9,0	3,3

Values. Equality

$$\mathsf{Prf}_{\text{equality}}(s, s') = 1 - 4 \cdot GI(s') = 1 - 2 \cdot \frac{|x'_\alpha - x'_\beta|}{x'_\alpha + x'_\beta}$$

Example: Prisoner's Dilemma

Norms. Sierra et al. (2019)'s game

	β co-operates	β defects
α co-operates	6,6	0,9
α defects	9,0	3,3

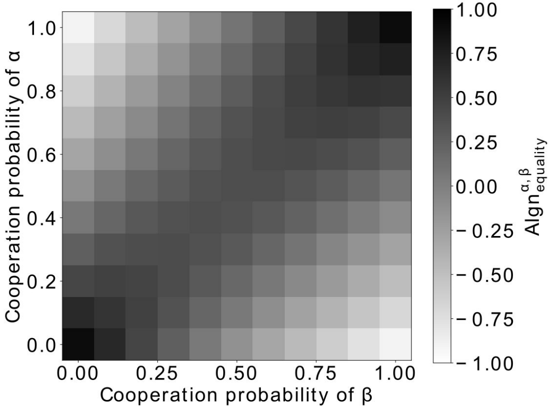
Values. Equality

$$\mathsf{Prf}_{\text{equality}}(s, s') = 1 - 4 \cdot GI(s') = 1 - 2 \cdot \frac{|x'_\alpha - x'_\beta|}{x'_\alpha + x'_\beta}$$

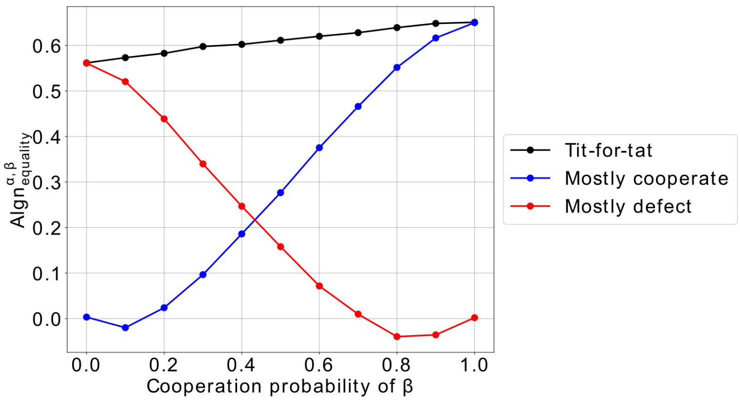
Strategies.

- Random-action Profiles
- Heterogeneous Profiles: a) tit for tat,
  b) mostly cooperate,
  c) mostly defect

Results



Alignment under
random actions profiles



Alignment under
heterogeneous profiles

- Under random profiles, alignment is highest when both players have similar cooperation probabilities.
- Under heterogeneous profiles, tit-for-tat results in stable alignment.

Value-Aligned Norms

We have assessed value-aligned strategies.

Can we assess value-aligned norms?

Example: Prisoner's Dilemma

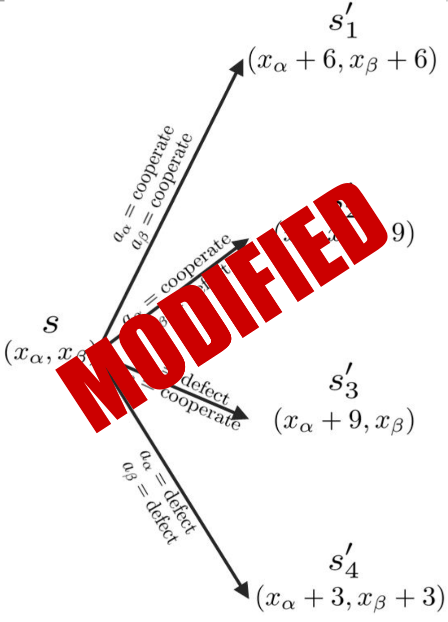
Norms. Sierra et al. (2019)'s game

	β co-operates	β defects
α co-operates	6,6	0,9
α defects	9,0	3,3

Example: Prisoner's Dilemma

Norms. Sierra et al. (2019)'s game

	β co-operates	β defects
α co-operates	6,6	0,9
α defects	9,0	3,3



New Norms.

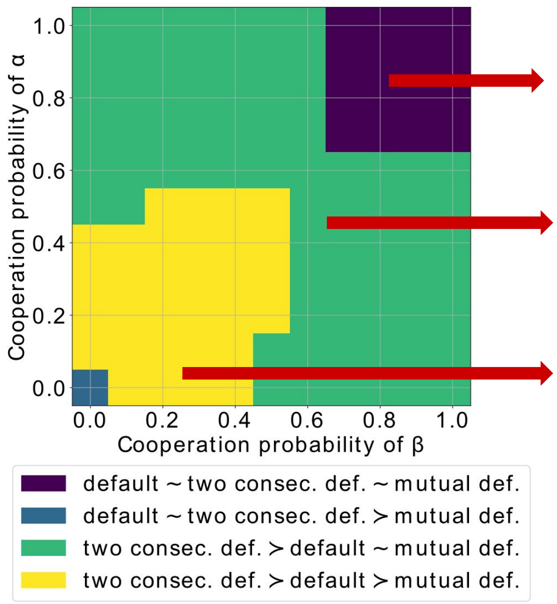
❶ **Ban on 2 consecutive defections:**

If a player defect twice in a row, then they are obliged to cooperate next.

❷ **Ban on mutual defection:**

Both players defect, then the outcome is as if one had cheated on the other (random toss).

Results



Relative Alignment of the norms

Cooperative society →
no gain by introducing the new norms

Exploitative society →
banning consecutive defections improves alignment

Defective society →
either of the new norms improves alignment

Montes (2020)

Value Alignment

Montes & Sierra (2021 b)

Value Aligned Norm Synthesis

Previous work:

Norms + Values → Alignment

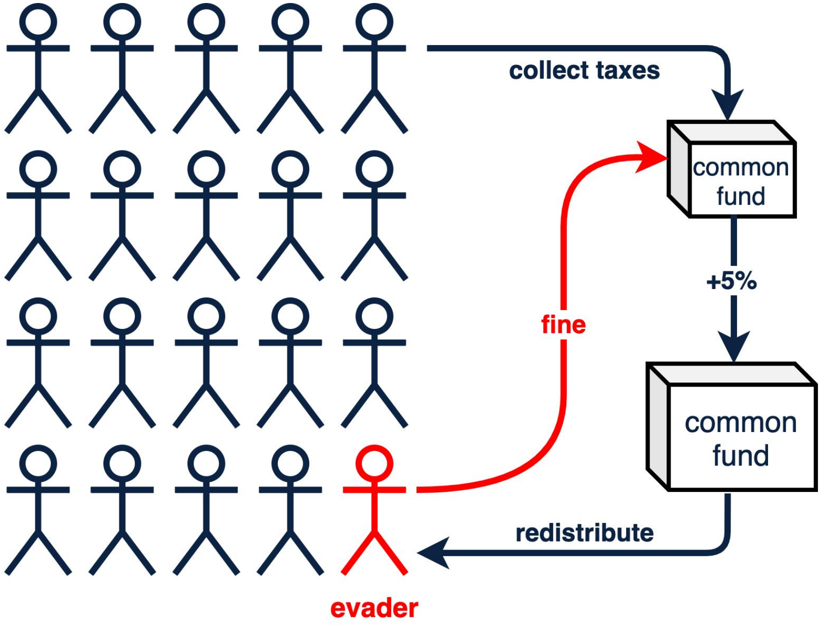
Current work:

Values → Norms

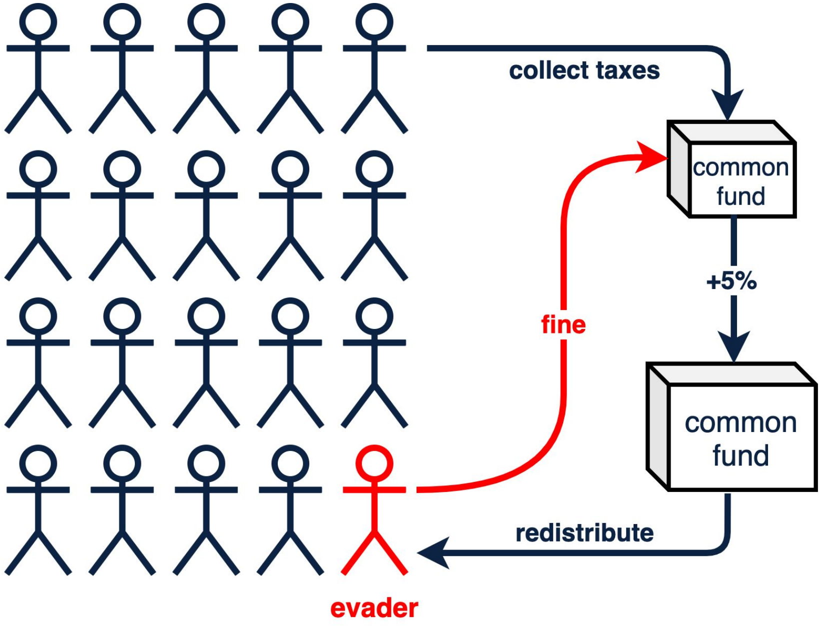
Find the norms that maximise alignment!

Example

A simple tax model.



A simple tax model.



n_1 **collecting rates**, specifies the percentage of taxes to be paid per wealth segment

n_2 **redistribution rates**, specifies the redistribution of revenue per wealth segment

n_3 **evader detection probability**

n_4 **fine rate**

Values

Equality.

$$\mathsf{Prf}_{\text{equality}}(s, s') = 1 - 4 \cdot GI(s') = 1 - 2 \cdot \frac{|x'_\alpha - x'_\beta|}{x'_\alpha + x'_\beta}$$

Value-Aligned Norm Synthesis

Find the norm parameters that maximise alignment.

An Optimisation Problem.

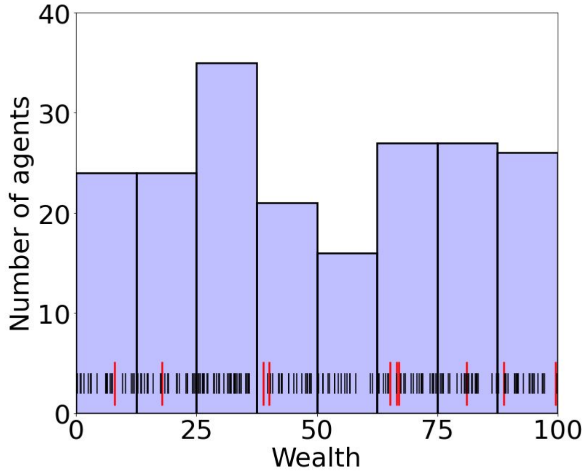
$$N^* = argmax_{N' \subseteq N} \mathsf{Algn}_{N',V}$$

A **genetic algorithm** searches the parameters of the norms in order to maximise their alignment w.r.t. the aspired values.

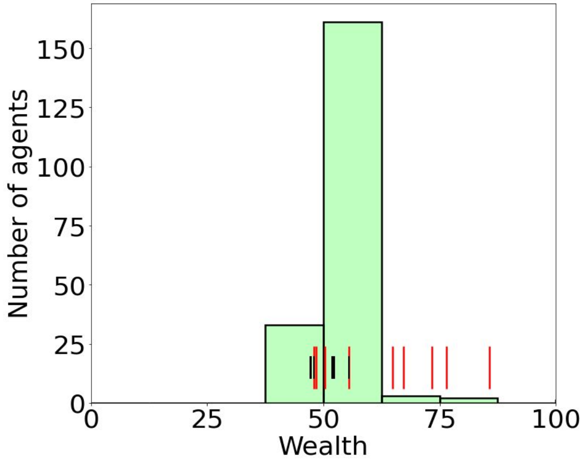
Results

Value and target function	**Optimal normative parameters P^*_N**	**Optimal alignment $Algn^{G\,*}_{N,v}$**
Equality, eq. (3)	*collect* = [20%, 29%, 26%, 35%, 27%]	0.95
	redistribute = [20%, 22%, 19%, 26%, 13%]	
	catch = 44%	
	fine = 61%	

Results



Wealth distribution at the beginning



Wealth distribution at the end

- Wealth evenly distributed at the end.

Individual Norm's Impact on Alignment

What is the contribution of each individual norm to the overall alignment?

The **shapley value** concept of game theory is useful.

When a coalition of players cooperate,
and a certain gain is realised,
the shapley value helps compute
how important is one player to that cooperation.

Individual Norm's Impact on Alignment

What is the contribution of each individual norm to the overall alignment?

The **shapley value** concept of game theory is useful.

$$\phi_i(v) = \sum_{N' \subseteq N \setminus \{n_i\}} \frac{|N'|!\,(|N| - |N'| - 1)!}{|N|!} \cdot \left(\mathsf{Algn}_{N' \cup \{n_i\}, v} - \mathsf{Algn}_{N', v} \right)$$

Results

Value	Norm	Shapley value
Equality	n_1	0.50
	n_2	0.03
	n_3	0.08
	n_4	0.01

- Collection of taxes (n_1) is enough to shrink wealth distribution!

Value Compatibility

How compatible are values v_1 and v_2 under norms N?

Given a fixed set of norms N
that maximises alignment for value v_1,
what is the alignment w.r.t. a new value v_2?

Results

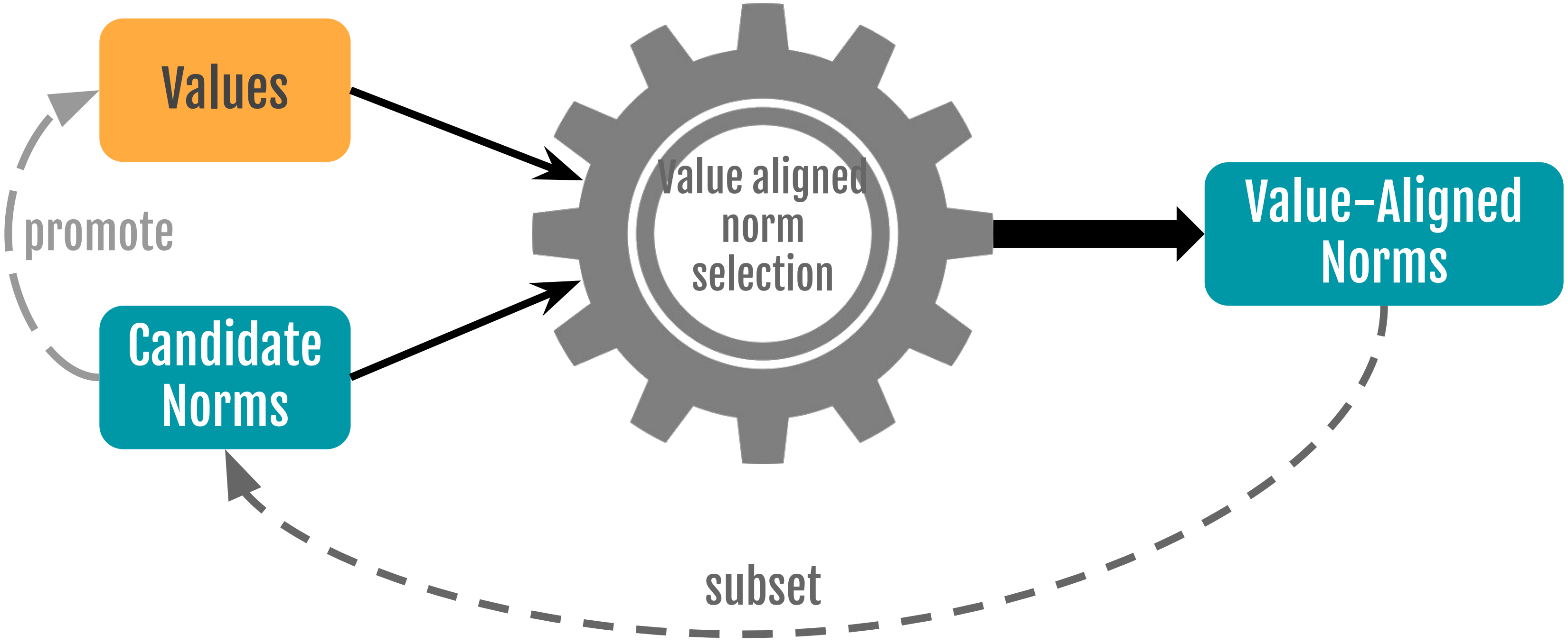
		v_j	
		Equality	Fairness
v_i	Equality	-	-0.28
	Fairness	0.60	-

- Strong pursue of equality neglects fairness.
- Seeking fairness respects equality to a large degree.

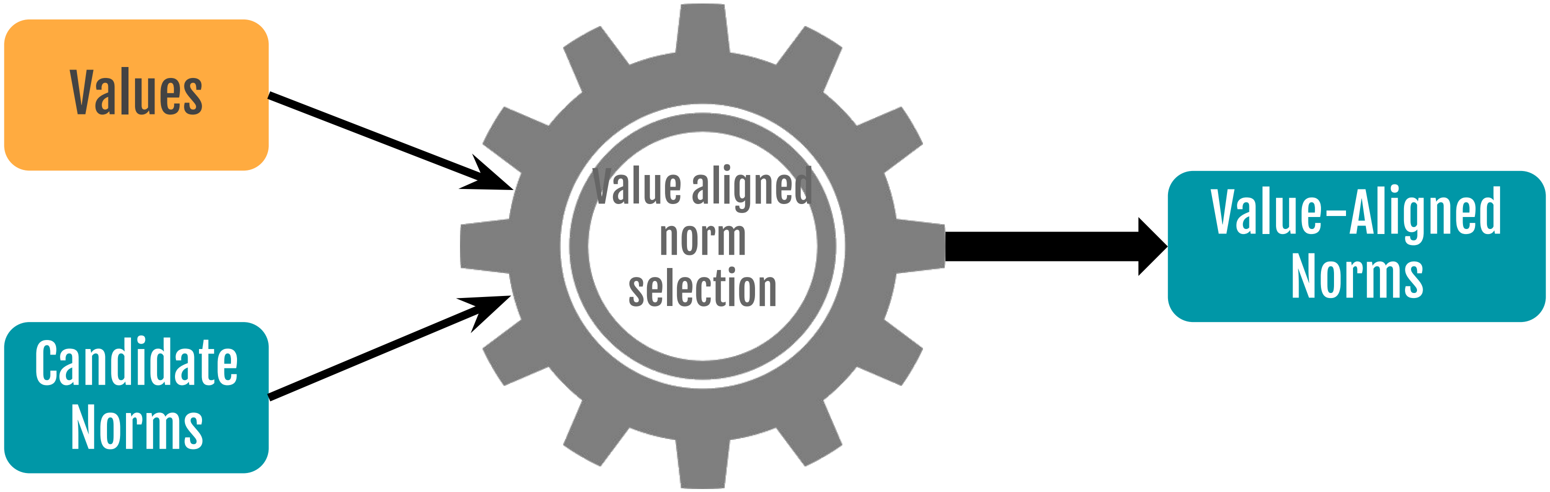
Value Alignment

Serramia et al. (2018)

Value-Aligned Norm Selection



Value-Aligned Norm Selection



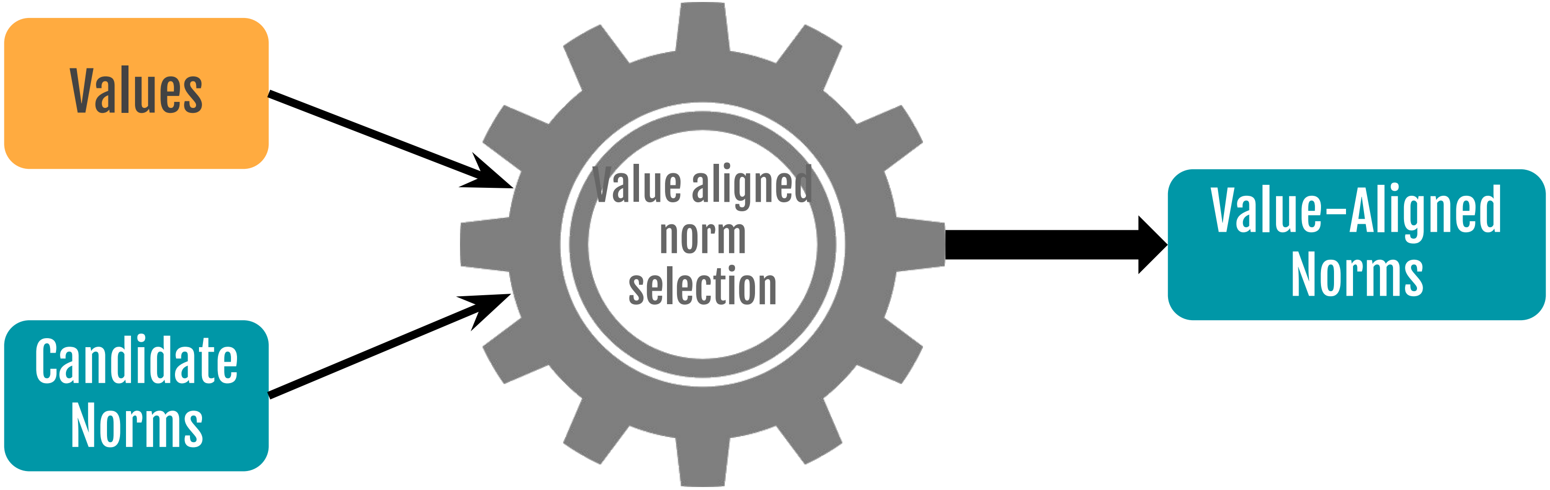
Value-Aligned Norm Selection

$v_1 \succcurlyeq v_2 \succcurlyeq v_3$

The **value system** is a structure containing:

- a set of moral values
- preferences over these values (a ranking $\succcurlyeq$)

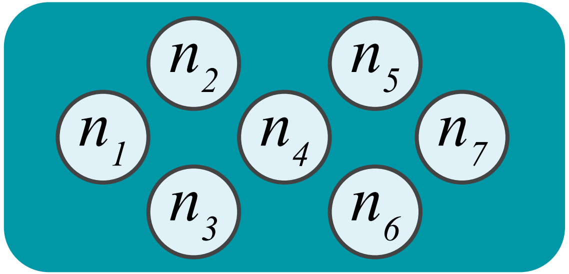
Value-Aligned Norm Selection



Value-Aligned Norm Selection

A **norm net** contains:

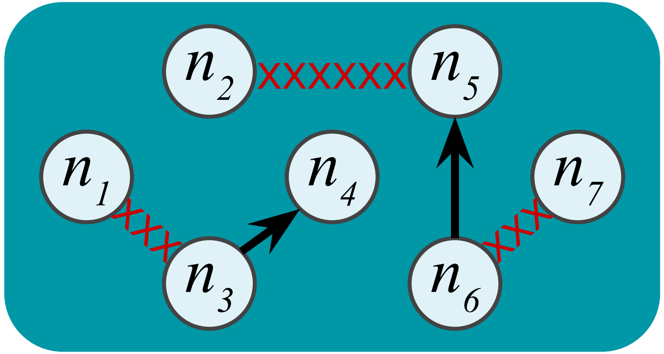
- a set of norms
- norm relations



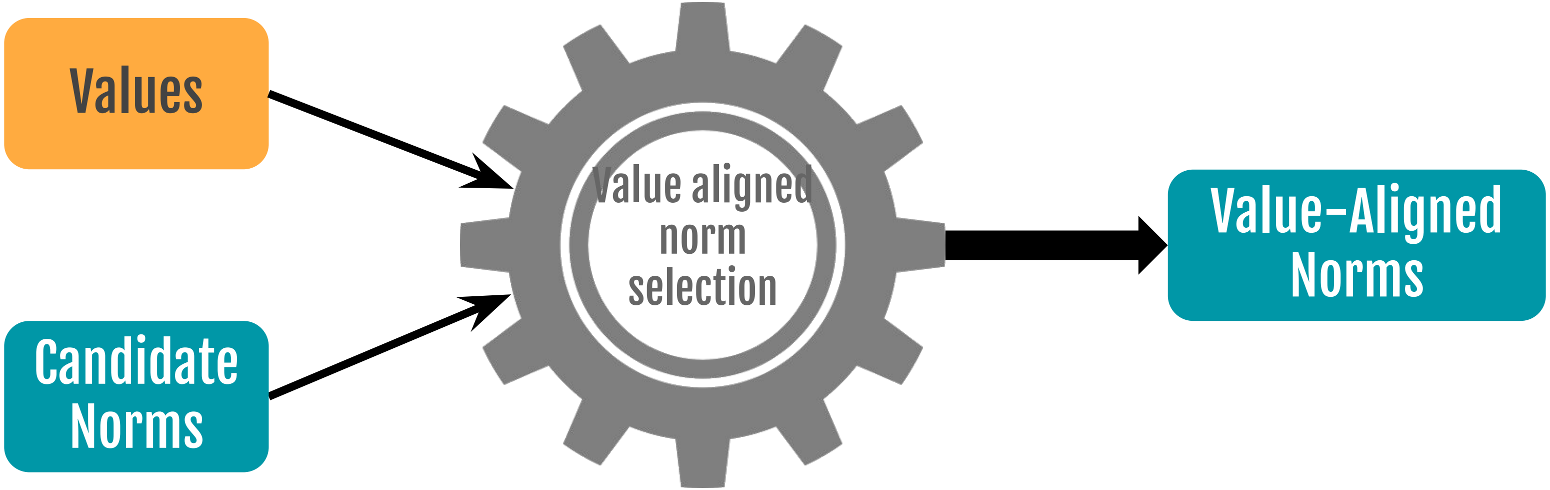
Value-Aligned Norm Selection

A **norm net** contains:

- a set of norms
- norm relations:
  - **exclusivity**
  - **generalisation**



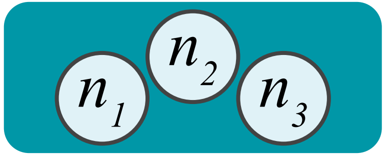
Value-Aligned Norm Selection



Value-Aligned Norm Selection

Objective of value-aligned norm selection is to find a **norm system** that:

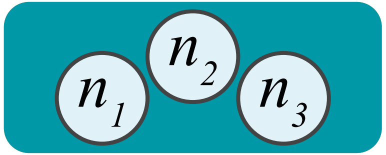
- best aligns the values system
- is sound



Value-Aligned Norm Selection

Objective of value-aligned norm selection is to find a norm system that:

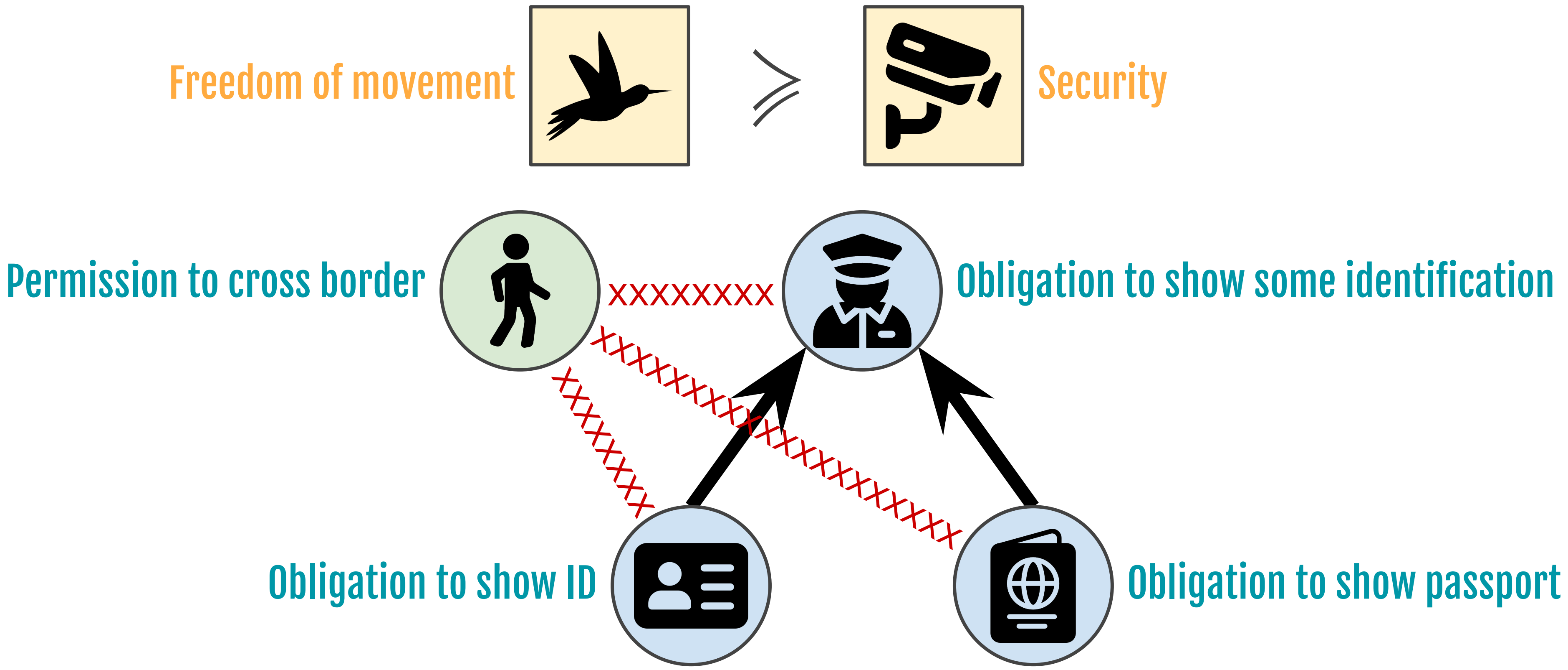
- best aligns the values system
- is sound



A norm system is sound if it is:

- **Conflict-free:** It does not contain exclusive norms.
- **Non-redundant:** It does not contain specific norms and those generalising them.

Example


Freedom of movement
≻
Security
Permission to cross border
Obligation to show some identification
Obligation to show ID
Obligation to show passport

Proposal

Assess the value alignment of a norm as a utility, then the solution is the sound norm system that maximises its cumulative utility.

Norm Utility

Norm utilities, that describe the value alignment of norms, depend on:

- **Value utilities**
- **Norm-value utilities**

Value Utility

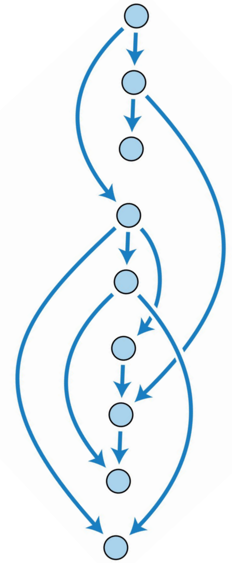
 $\succcurlyeq$ 

$$v \geq v' \Rightarrow u(v) \geq u(v')$$

Value Utility

 $\succ$ 

$$u(v) = 1 + \sum_{v \succ v'} u(v')$$



Value Utility

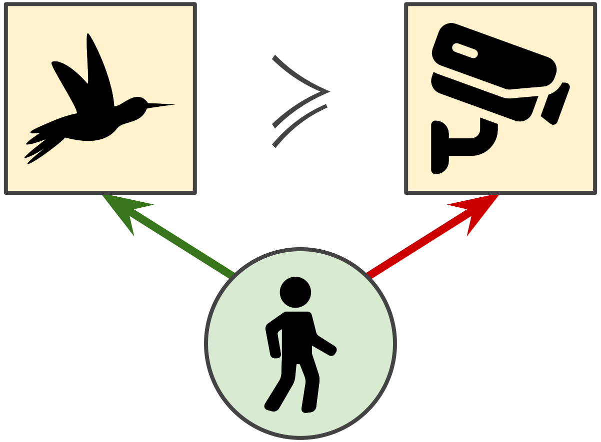
 $\succ$ 

$$u(v) = 1 + \sum_{v \succ v'} u(v')$$

Example.

$$u(v_{sec}) = 1$$
$$u(v_{free}) = 2$$

Norm–Value Utility



$$u(n,v) \in [-1, 1]$$

Norm–Value Utility

Example.

				
	1	-0.2	-0.2	-0.2
	-1	0.7	0.7	0.7

Norm Utility

Recall. This is the utility describing the value alignment of a norm
with respect to all values, and hence,
taking into consideration the utility of each value.

$$u(n) = \sum_{v} u(n,v) \cdot u(v)$$

Norm Utility

Example.

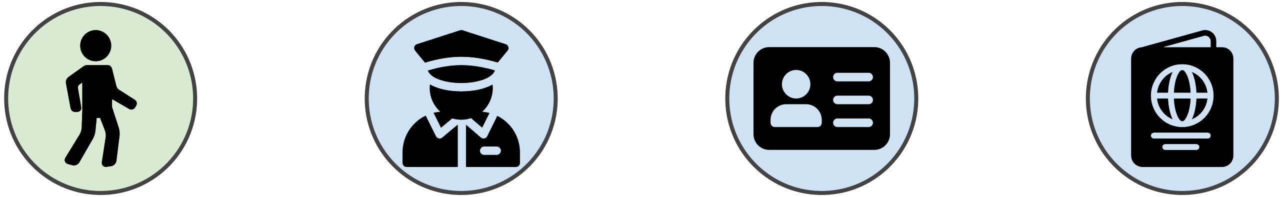
				
	1	-0.2	-0.2	-0.2
	-1	0.7	0.7	0.7

$$u(v_{sec}) = 1$$
$$u(v_{free}) = 2$$

Example.

				
	1	-0.2	-0.2	-0.2
	-1	0.7	0.7	0.7

$$u(v_{sec}) = 1$$
$$u(v_{free}) = 2$$



$u(n_1) = 1$ $u(n_2) = 0.3$ $u(n_3) = 0.3$ $u(n_4) = 0.3$

Norm System Utility

The **utility of a norm system** is the sum of the utilities of its norms.

$$u(\Omega) = \sum_{n \in \Omega} u(n)$$

Value-Aligned Norm Selection

Find the sound norm system that maximises this utility.

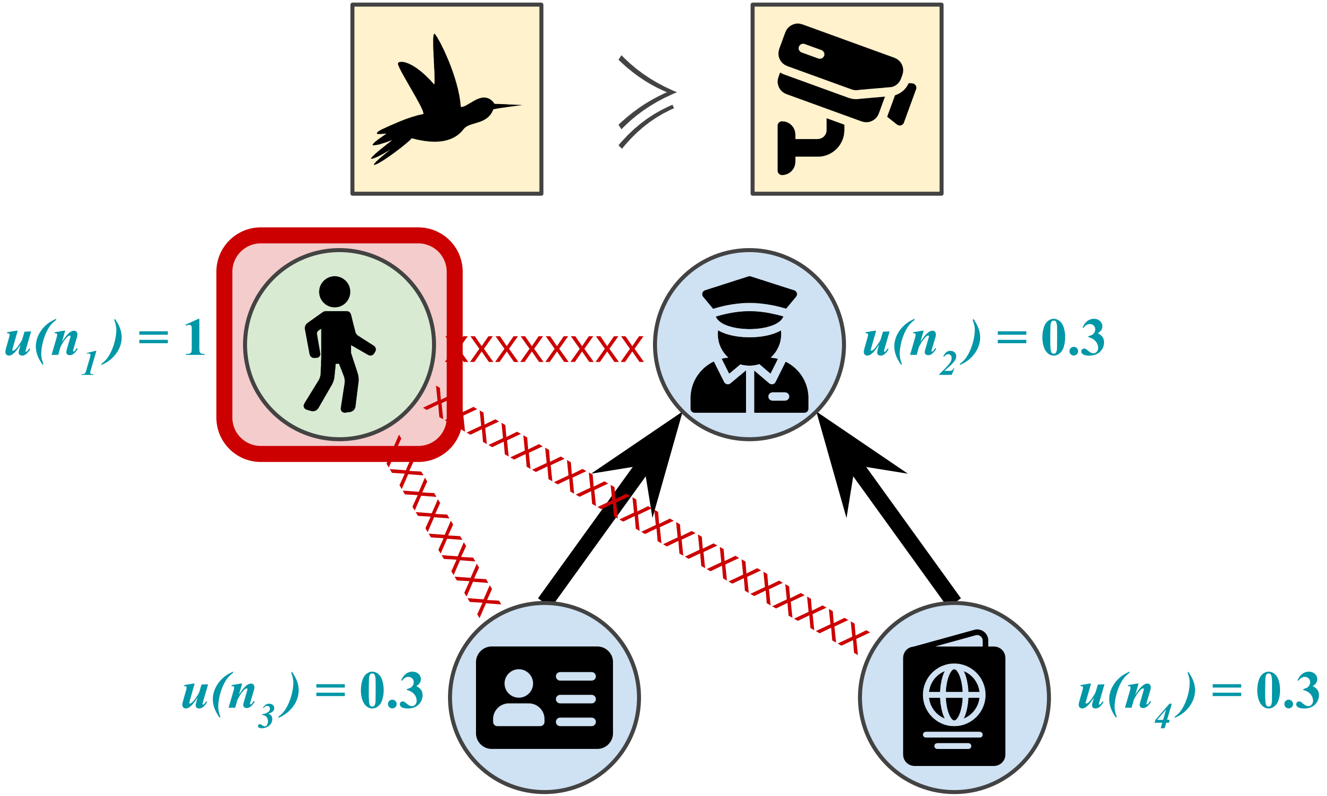
$$u(\Omega) = \sum_{n \in \Omega} u(n)$$

An Optimisation Problem.

Encoded as a linear program,
that also considers the constraints of norm relations.

$$\max_{x_1,\dots,x_k \in \{0,1\}} \; x_1\, u(n_1) + \ldots + x_k\, u(n_k)$$

Example



Value Alignment

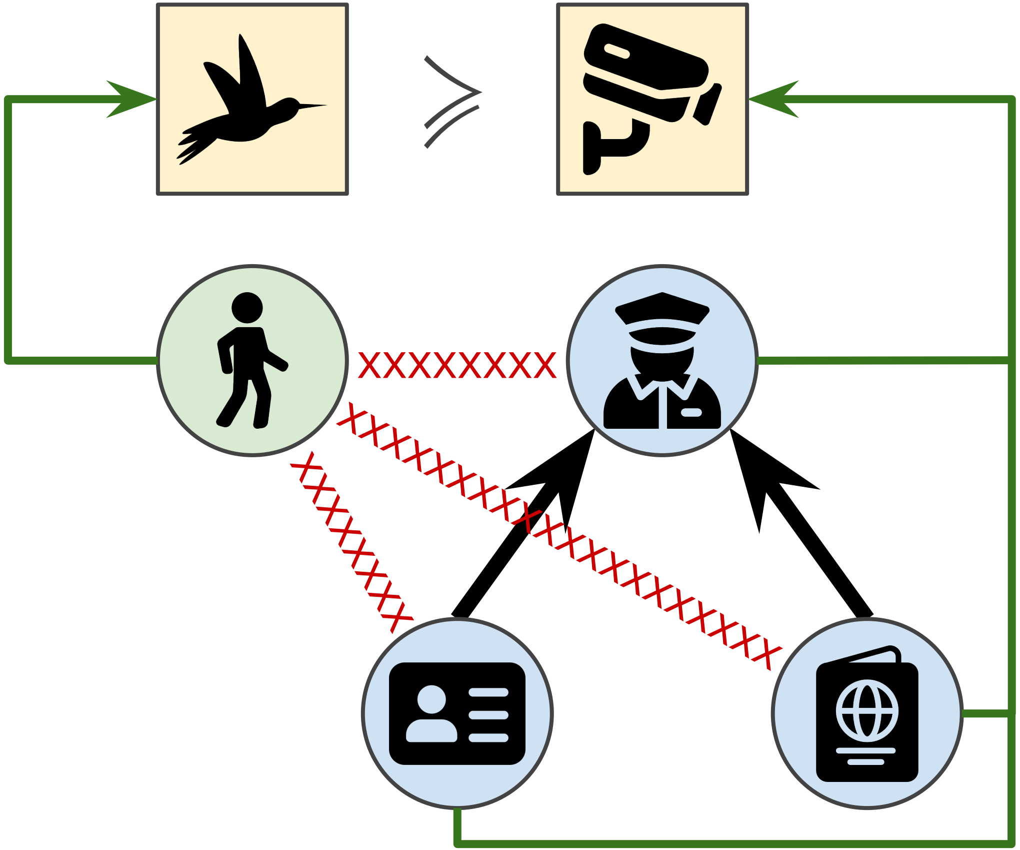
Serramia et al. (2020)

Value-Aligned Norm Selection

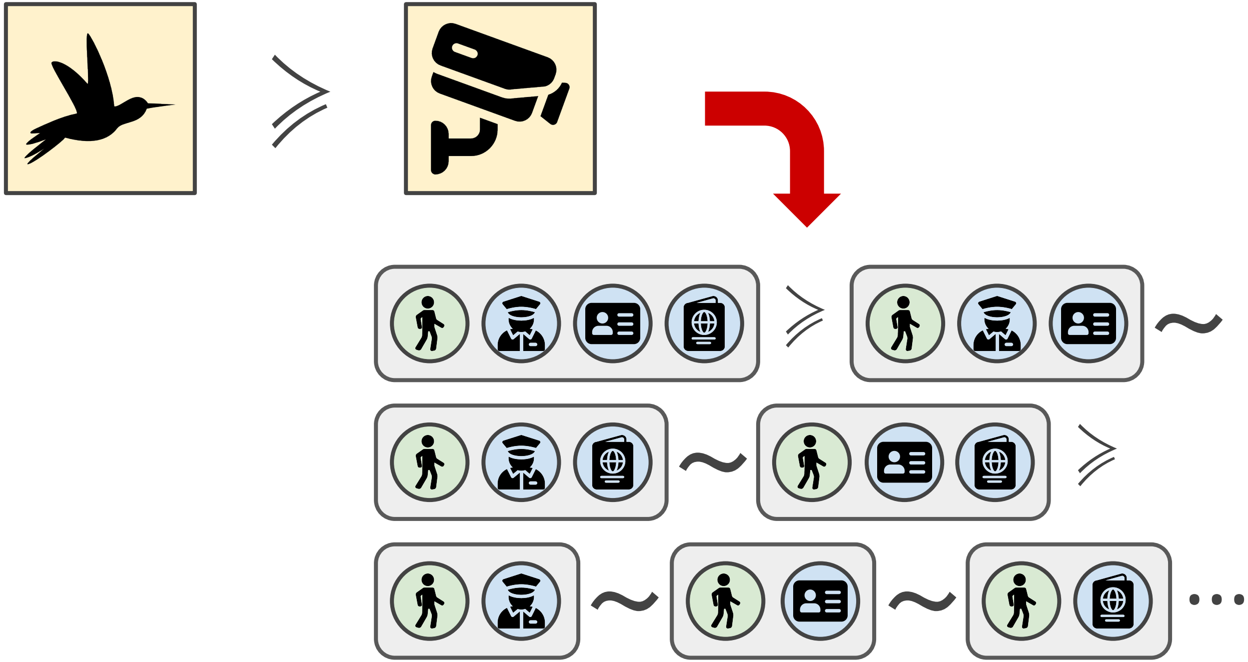
Quantitative Approach ➡ **Qualitative Approach**

Transform **value preferences** into **preferences over all norm systems**,
Then the solution is the **most preferred sound** norm system.

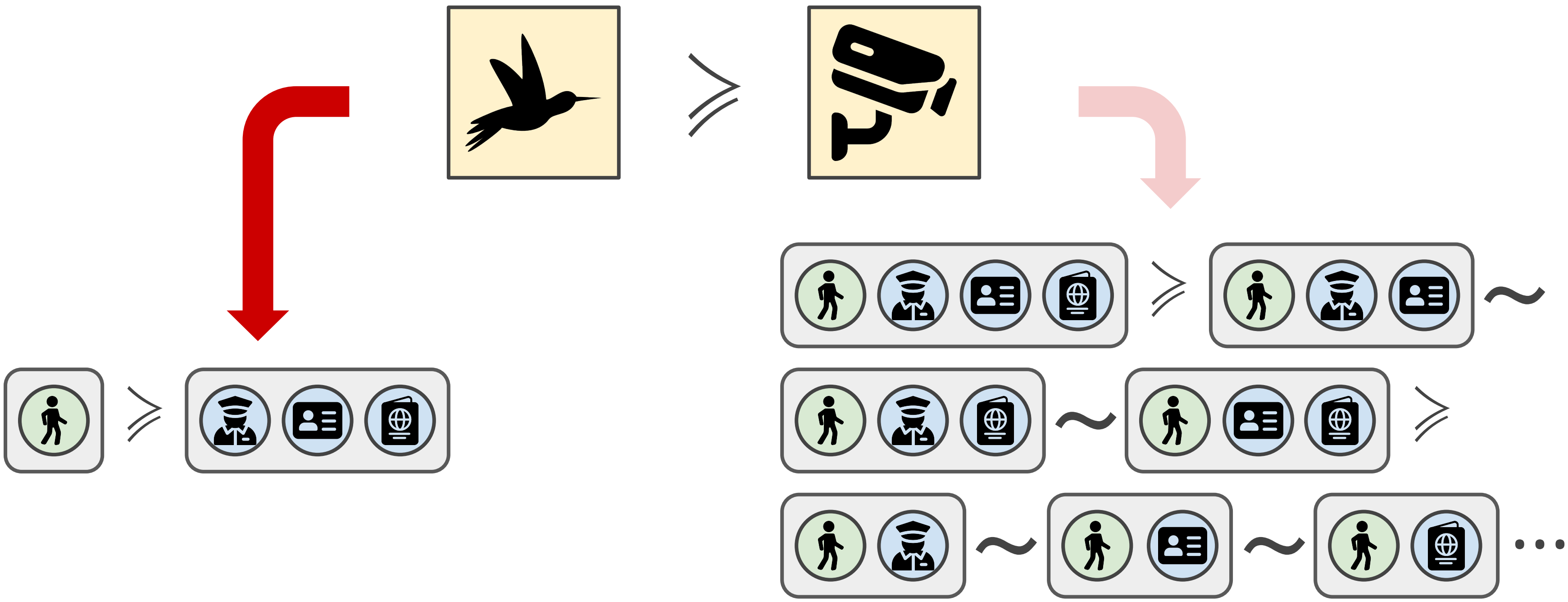
Value-Aligned Norm Selection



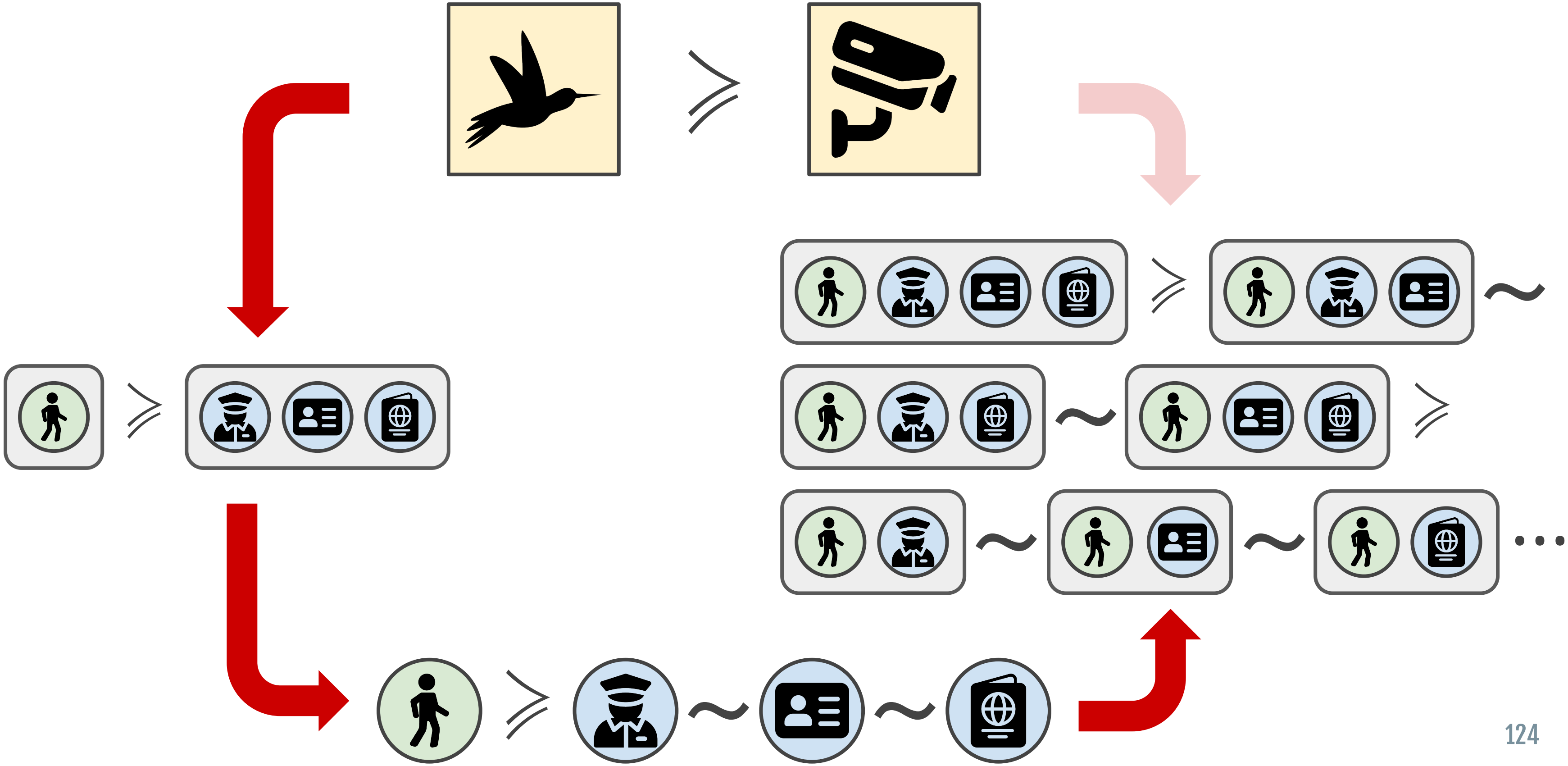
Value-Aligned Norm Selection



Value-Aligned Norm Selection

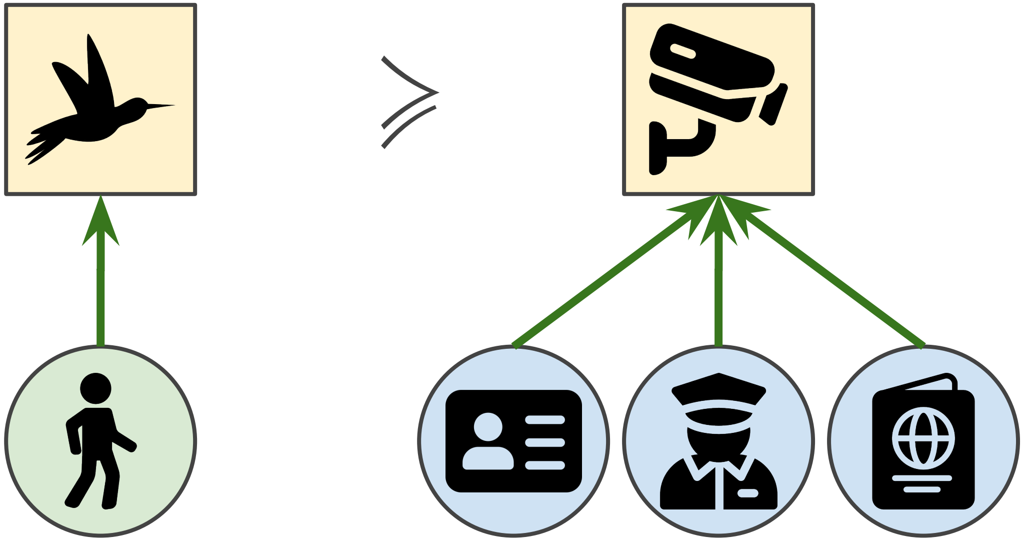


Value-Aligned Norm Selection



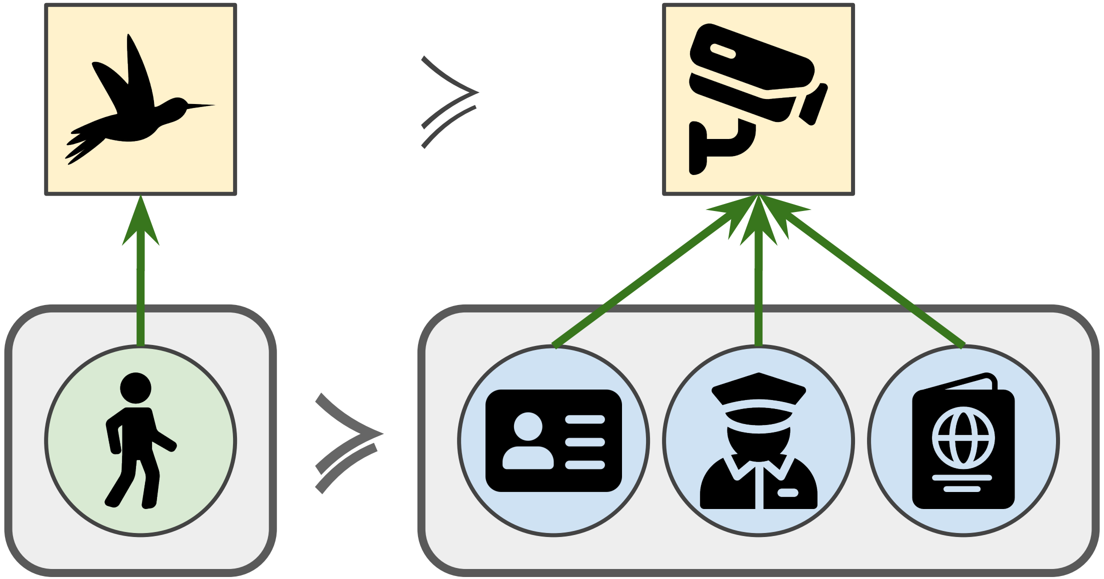
Step 1. Preference Induction

Get preferences over <u>some</u> norm systems.

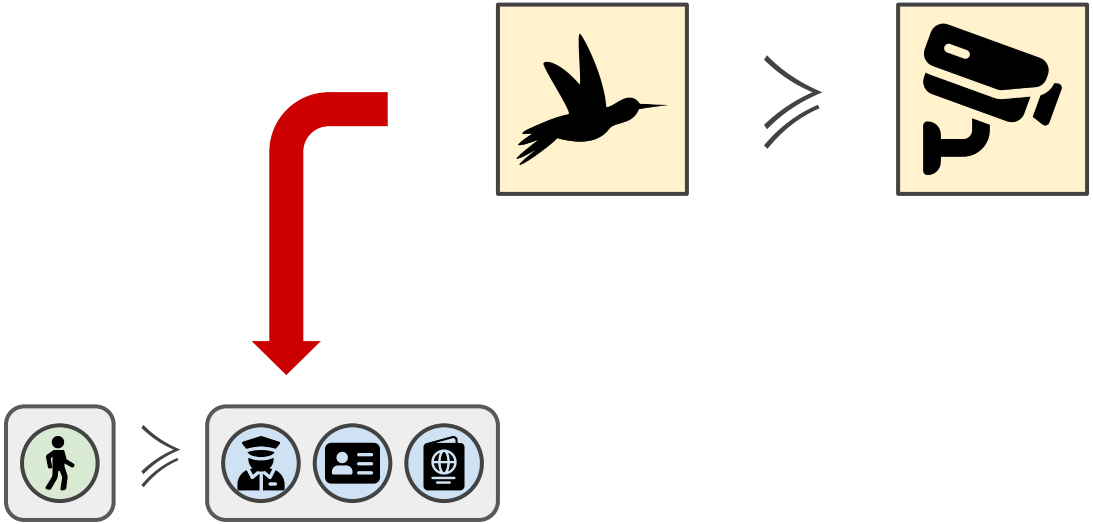


Step 1. Preference Induction

Get preferences over <u>some</u> norm systems.



Step 1. Preference Induction



Step 2. Preference Grounding

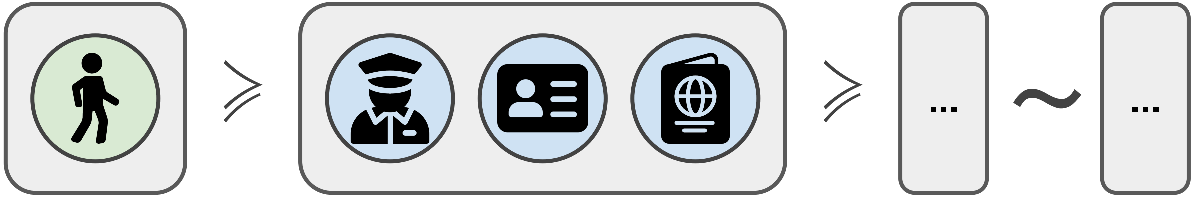
Ground these preferences to preferences over single norms.

We use **Lex-cel**, a novel method to ground preferences from sets of objects to objects.
It satisfies properties that make the grounding fair.

Bernardi et al. (2019)

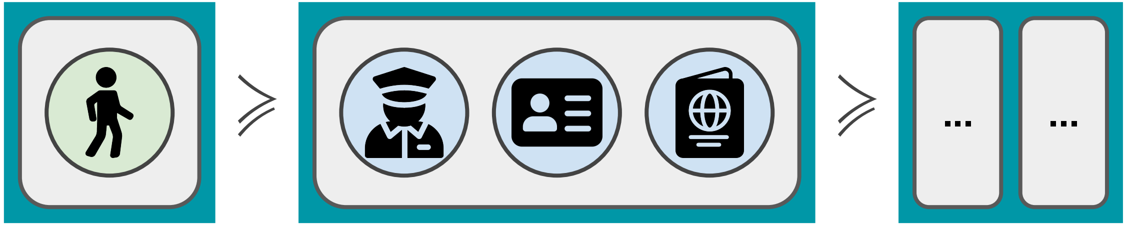
Step 2. Preference Grounding

1. Extend the preferences



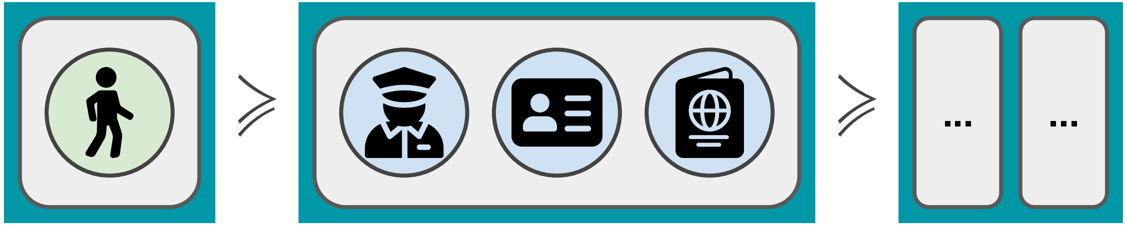
Step 2. Preference Grounding

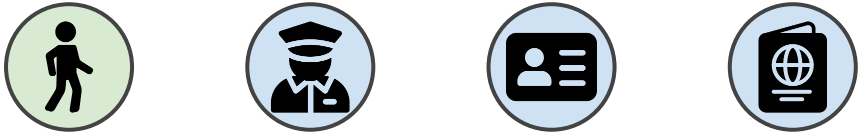
2. Extract Equivalence Classes



Step 2. Preference Grounding

3. For each norm, compute occurrence in equivalence classes

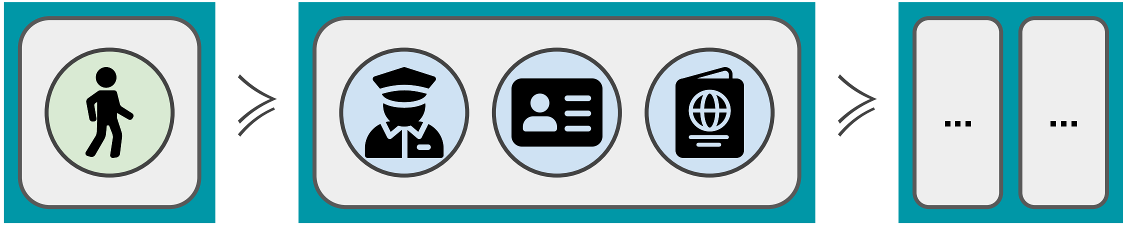


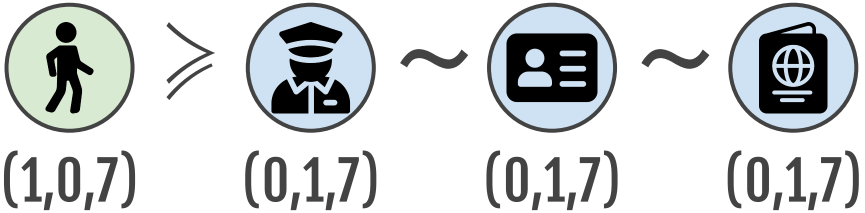


(1,0,7) (0,1,7) (0,1,7) (0,1,7)

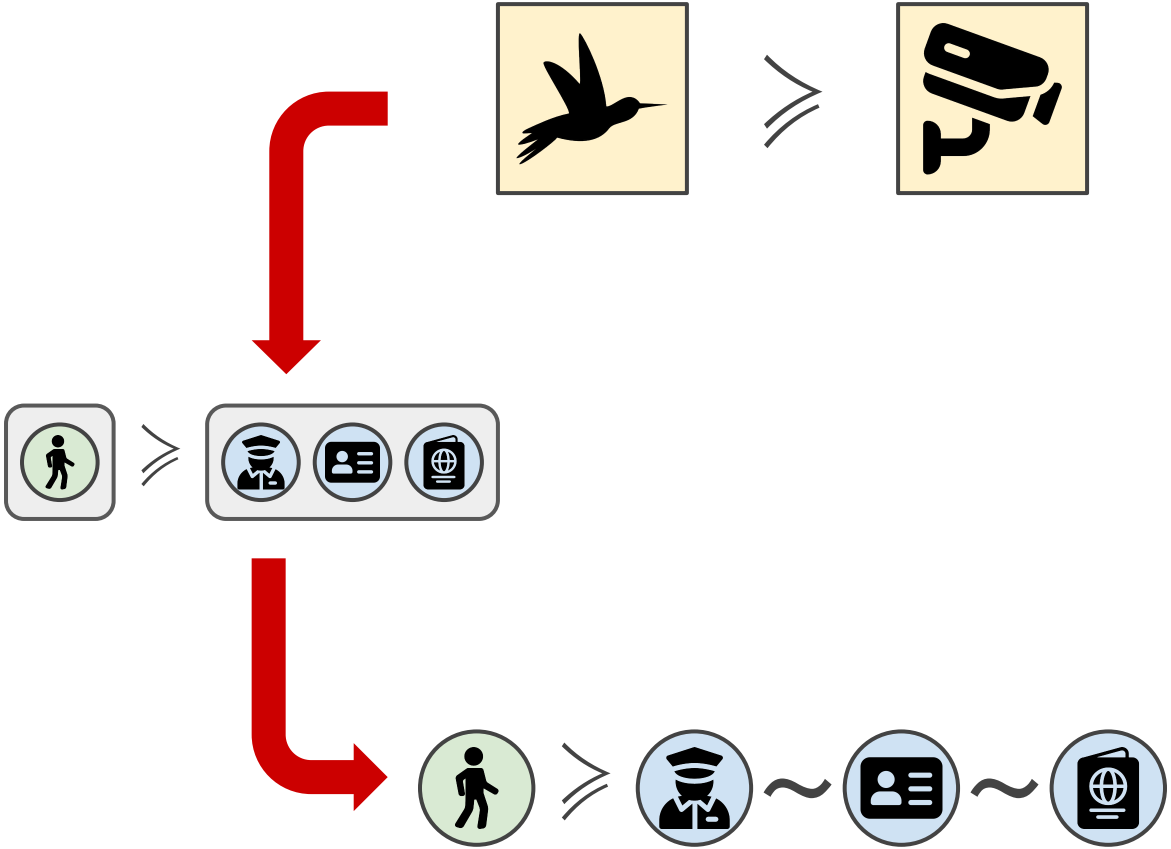
Step 2. Preference Grounding

4. Compare the norms lexicographically





Step 2. Preference Grounding



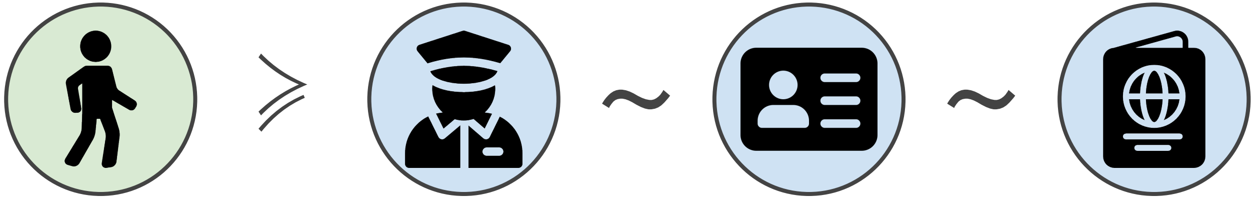
Step 3. Preference Lifting

Lift the preferences over norms to preferences over ALL norm systems.

We design a novel **anti-Lex-cel** lifting mechanism that reverses input & output.

Step 3. Preference Lifting

1. Extract equivalence classes

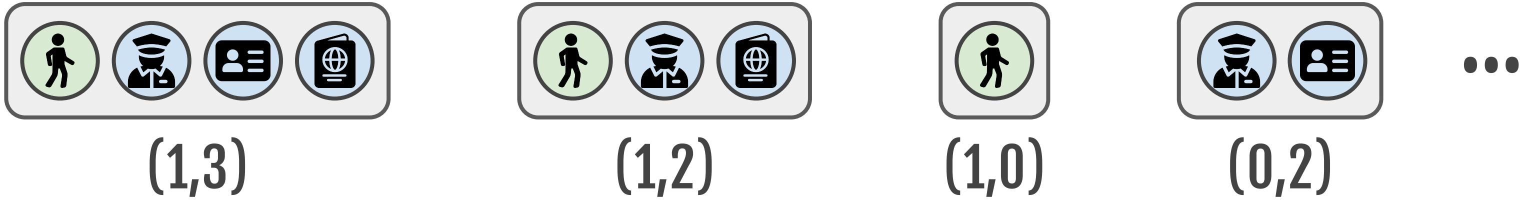


Step 3. Preference Lifting

1. Extract equivalence classes

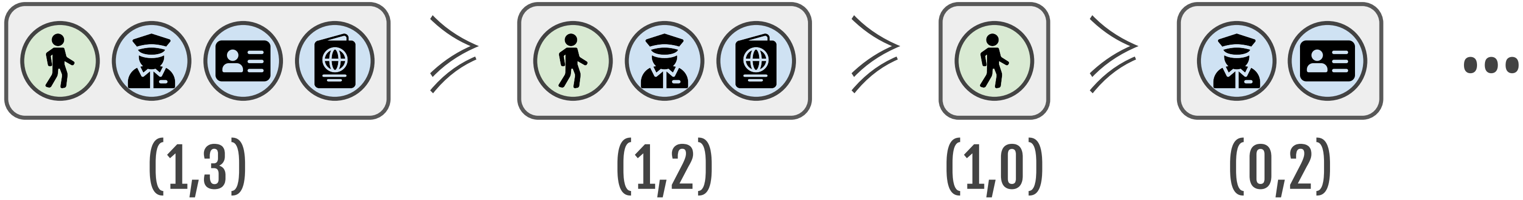
Step 3. Preference Lifting

2. For each norm system, compute occurrence of norms in eq. classes

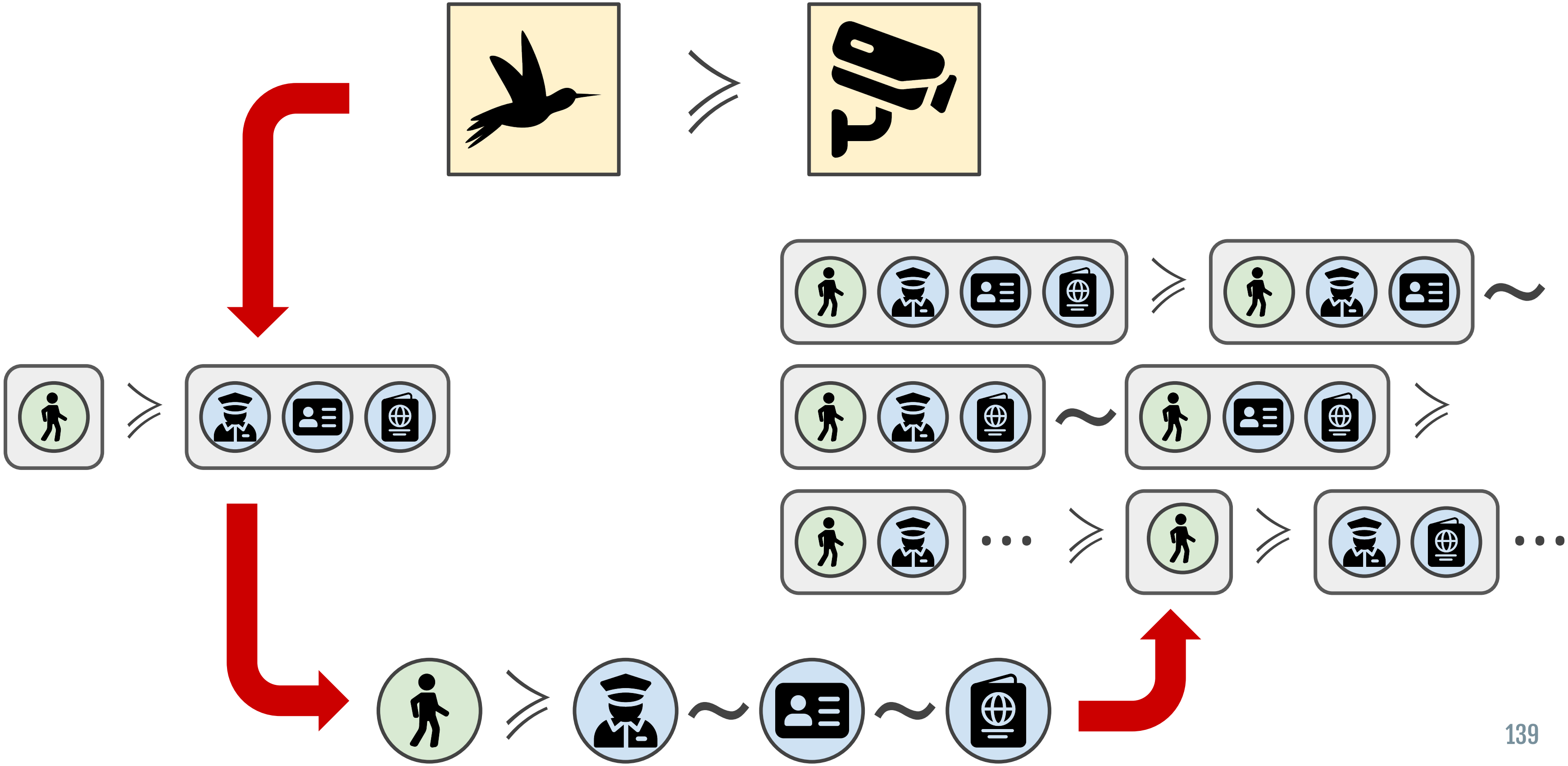


Step 3. Preference Lifting

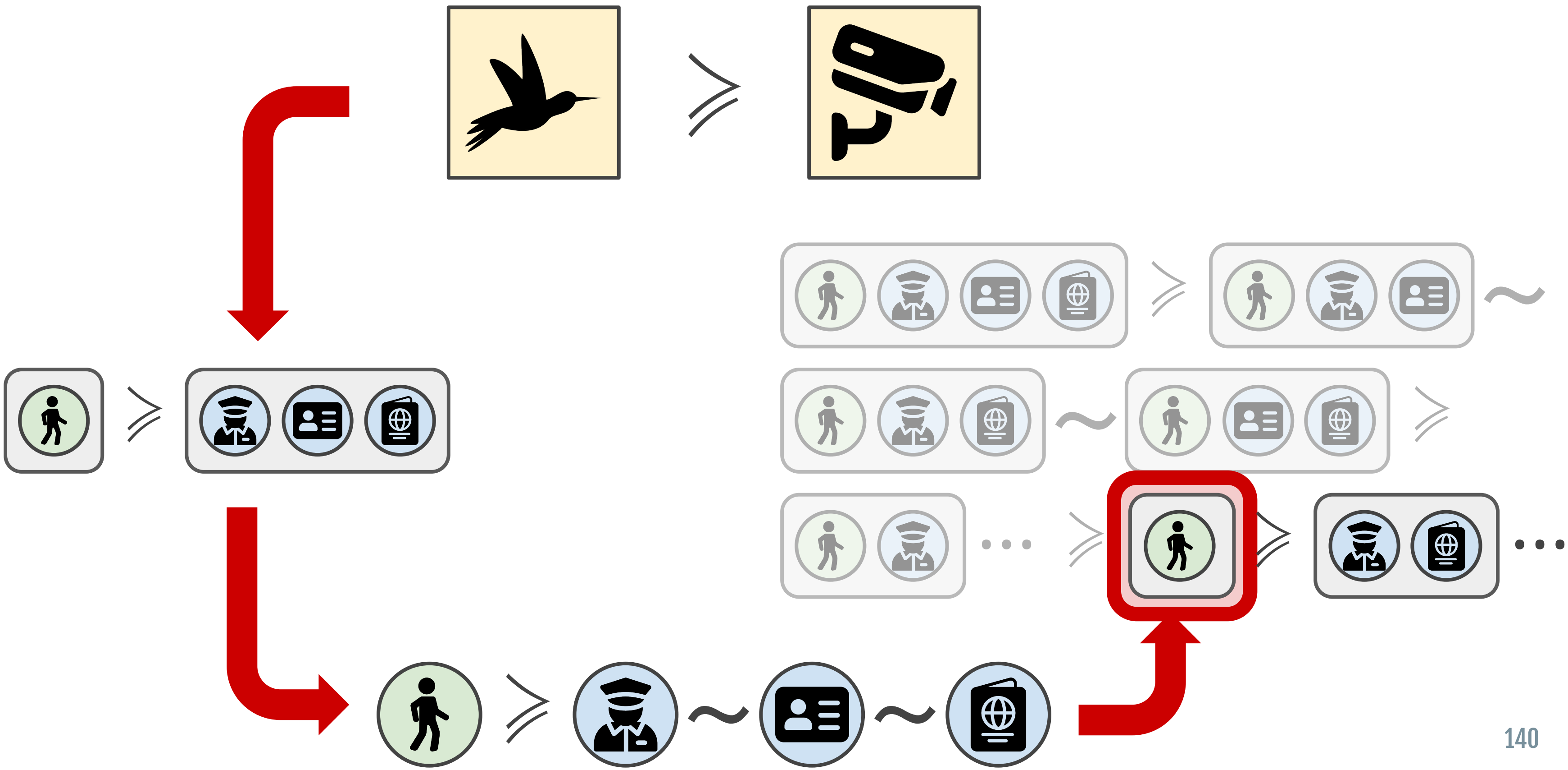
3. Compare the norm systems lexicographically



Step 3. Preference Lifting



Step 4. Discard Non-sound Norm Systems



Optimisation

We have to build preferences for $2^{|N|}$ norm systems and check for soundness!

This is computationally costly.

We translate the problem into an optimisation problem.

Optimisation

We have designed an **alignment** formula that gives the value alignment of any norm system satisfying:

$$\Omega \succcurlyeq \Omega' \Leftrightarrow al(\Omega) \geq al(\Omega')$$

The optimisation problem can then be encoded as a linear program, that also considers the constraints of norm relations.

$$\mathbf{max} \qquad al(\{n_1\})\, x_1 + \ldots + al(\{\mathrm{n}_{|N|}\})\, x_{|N|}$$

Take Home Message #4

Value Alignment Mechanisms

- Formal definition of value alignment.
- Mechanisms for the value-aligned selection of:
  - agent strategies
  - norms
  - norm parameters

References I

1. Bernardi, G., Lucchetti, R., Moretti, S. (2019) Ranking objects from a preference relation over their subsets. Social Choice and Welfare 52, 4 (01 Apr 2019), 589–606. https://doi.org/10.1007/s00355-018-1161-1
2. Cheng, AS. and Fleischmann, K. R. (2010) Developing a meta-inventory of human values. *Proceedings of the 73rd Annual Meeting of the American Society for Information Science and Technology*, Pittsburgh, PA, USA. https://doi.org/10.1002/meet.14504701232
3. Criado Pacheco, N. (2012) Using Norms To Control Open Multi-Agent Systems. PhD Thesis, UPV.
4. de Jonge D., Sierra C. (2016) SIMPLE: A Language for the Specification of Protocols, Similar to Natural Language. In: Dignum V., Noriega P., Sensoy M., Sichman J. (eds) *Coordination, Organizations, Institutions, and Norms in Agent Systems XI. COIN 2015*. Lecture Notes in Computer Science, vol 9628. Springer, Cham. https://doi.org/10.1007/978-3-319-42691-4_6
5. Montes N. (2020) Value Alignment in Multiagent Systems. MSc Thesis, UAB.
6. Montes N. and Sierra C. (2021 a) Value-Alignment Equilibrium in Multiagent Systems. In: Heintz F., Milano M., O'Sullivan B. (eds) *Trustworthy AI - Integrating Learning, Optimization and Reasoning. TAILOR 2020*. Lecture Notes in Computer Science, vol 12641. Springer, Cham. https://doi.org/10.1007/978-3-030-73959-1_17

References II

7. Montes N. and Sierra C. (2021 b) Value-Guided Synthesis of Parametric Normative Systems. *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 907–915. https://dl.acm.org/doi/10.5555/3463952.3464060
8. Rohan MJ. (2000) A Rose by Any Name? The Values Construct. *Personality and Social Psychology Review*, 4(3):255-277. https://doi.org/10.1207/S15327957PSPR0403_4
9. Russell, S.J. (2014) *Of Myths and Moonshine*, Edge, November 14, 2014. http://edge.org/conversation/the-myth-of-ai#26015
10. Schwartz, S. H. (2007). Value orientations: Measurement, antecedents and consequences across nations. In R. Jowell, C. Roberts, R. Fitzgerald, & G. Eva (Eds.), *Measuring attitudes cross-nationally: Lessons from the European Social Survey* (pp. 169–203). Sage Publications, Inc. https://doi.org/10.4135/9781849209458.n9
11. Schwartz, S. H. (2012) An Overview of the Schwartz Theory of Basic Values. *Online Readings in Psychology and Culture*, 2(1). https://doi.org/10.9707/2307-0919.1116

References III

12. Serramia, M., Lopez-Sanchez, M., Rodriguez-Aguilar, J. A., Rodriguez, M., Wooldridge, M., Morales, J., and Ansotegui, C. (2018) Moral Values in Norm Decision Making. *In Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '18)*. IFAAMAS, Richland, SC, 1294–1302. https://dl.acm.org/doi/10.5555/3237383.3237891
13. Serramia, M., Lopez-Sanchez, M., Rodriguez-Aguilar, J. A. (2020) A Qualitative Approach to Composing Value-Aligned Norm Systems. *In Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '20)*. IFAAMAS, Richland, SC, 1233–1241. https://dl.acm.org/doi/abs/10.5555/3398761.3398904
14. Sierra, C., Osman, N., Noriega, P., Sabater-Mir, J., Perello-Moragues, A. (2019) Value alignment: A formal approach. In *Responsible Artificial Intelligence Agents Workshop (RAIA) in AAMAS 2019.*
15. van de Poel, I. (2021) Design for value change. *Ethics Inf Technol* 23, 27–31. https://doi.org/10.1007/s10676-018-9461-9
16. Vázquez-Salceda J., Aldewereld H., Dignum F. (2004) Implementing Norms in Multiagent Systems. In: Lindemann G., Denzinger J., Timm I.J., Unland R. (eds) Multiagent System Technologies. MATES 2004. Lecture Notes in Computer Science, vol 3187. Springer, Heidelberg. https://doi.org/10.1007/978-3-540-30082-3_23

Thank you!

Nardine Osman
nardine@iiia.csic.es