Value Awareness & Multiagent Systems

Nardine Osman
Artificial Intelligence Research Institute (IIIA-CSIC)

Conexión AIHUB Escuela 2022
Palma de Mallorca
July 4, 2022
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 are what govern/regulate behaviour.

- Don’t shout
- No pets
- No diving
- Don’t run
- Don’t swim alone
- No rough play
- No peeing in pool
- No littering
- Use the stairs
- Use restrooms
- Children only with parents
- Use cap and goggles
- Use swimsuit
- Use slippers
- Shower before pool
- Watch your children
Norm Representation

Mostly based on **deontic concepts**
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!'.
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

- Norm Implementation
  - Norm Regimentation
  - Norm Enforcement
    - Mediation
    - Hard-wiring
      - Self Enforcement
      - Second-Party Enforcement
      - Third-Party Enforcement
        - Retaliation
        - Reciprocation
        - Social Enforcement
        - Institutional Enforcement

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
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
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)
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 believes in general and not just for themselves to be able to lead a good life or realize a good society.”

Van de Poel & Royakkers (2011, p. 72). Values are “lasting convictions or matters of concern, about which people care and aspire to change their relationships to them, not just for themselves but for others as well, for the sake of improving the quality of life”.

<table>
<thead>
<tr>
<th>Author/Year</th>
<th>Citations</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lewin (1952, p. 41)</td>
<td>“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.”</td>
<td></td>
</tr>
<tr>
<td>Guth &amp; Tagiuri (1965, p.124-125)</td>
<td>“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”.</td>
<td></td>
</tr>
<tr>
<td>Hutcheon (1972, p. 184)</td>
<td>“…values are not the same as ideals, norms, desired objects, or espoused beliefs about the 'good', but are, instead, operating criteria for action...”</td>
<td></td>
</tr>
<tr>
<td>Rokeach (1973, p. 5)</td>
<td>“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”</td>
<td></td>
</tr>
<tr>
<td>Schwartz (1994, p.20)</td>
<td>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”.</td>
<td></td>
</tr>
<tr>
<td>Feather (1996, p. 222)</td>
<td>“I regard values as beliefs about desirable or undesirable ways of behaving or about the desirability or otherwise of general goals.”</td>
<td></td>
</tr>
<tr>
<td>Braithwaite &amp; Blamey (1998, p.364)</td>
<td>“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”.</td>
<td></td>
</tr>
<tr>
<td>Friedman et al. (2006, p. 349)</td>
<td>“A value refers to what a person or group of people consider important in life”</td>
<td></td>
</tr>
<tr>
<td>van de Poel &amp; Royakkers (2011, p. 72)</td>
<td>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”</td>
<td></td>
</tr>
</tbody>
</table>
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)
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
Values in the Social Sciences

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 ... 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

Value

- biological needs
- social coordination
- group survival & welfare

Goal

represents

Function

serves

accomplishes
Proposal.

Abstract values are grounded into permanent goals, that agents actively pursue.
So from a computational perspective, what is the difference between goals that ground values and traditional AI goals?

- Permanency
- Degree of satisfaction
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
- ...

Examples of Value-Grounding Goals
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)
We propose values to be grounded into \textit{permanent goals}.

\textbf{Preferences over values} need to be specified, to help with conflicts.

\textbf{Dynamic} nature of values must be acknowledged and accounted for.
The Norm–Value Relationship
So what are the computational implications of having values grounded into goals?

- Supports decision making over agent actions
- Support norm selection/generation
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

Values → Norms
- legitimise
- promote

Norms → Values
- steer towards

Values → Outcomes
- evaluate

Outcomes → Values
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
Value-aware AI

Noun [U] 
/ˈvæl.juəˈweər əˌweər ə ˈwɛər ə/ 

an AI system that understands and abides by a value system, explains 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
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 are understood as preferences over behaviour, or preferences over the states of the world.
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: $\text{Prf}_v(s, s')$
A value-based preference depends on the satisfaction of state properties relevant to the value: $\Phi_v$.

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

- Male’s salary > Female’s salary
- Male’s salary = Female’s salary

Defining Value-Based Preferences
A value-based preference depends on the satisfaction of state properties relevant to the value: $\Phi_\nu$

\[
\text{Prf}_\nu(s, s') = f \left( \mathbb{P}(s \models \Phi_\nu), \mathbb{P}(s' \models \Phi_\nu) \right)
\]
What about preferences over sets of values & for groups of people?

Sets of Values & Groups of People

\[
\Prf^\alpha_v(s,s') \xrightarrow{\Prf^G}(s,s') \xrightarrow{\Prf^\alpha}(s,s')
\]

\[
\Prf^\alpha(s,s') = \frac{\sum_{\alpha \in G} \Prf^\alpha_v(s,s')}{|G|}
\]

\[
\Prf^G(s,s') = \frac{\sum_{\alpha \in G} \Prf^\alpha_v(s,s')}{|V|}
\]

\[
\Prf^\alpha_v(s,s') = \frac{\sum_{\alpha \in G} \Prf^G(s,s')}{|G|}
\]
Norms define the possible worlds.
Norms define the possible worlds.

Changing the norms result in changing the possible worlds.
The degree of alignment of a norm $n$ with a value $v$ for agent $\alpha$ is the **accumulation of preferences** along the transitions.
The degree of alignment of a norm $n$ with a value $v$ for agent $\alpha$ is the **accumulation of preferences** along the transitions.

And we consider **all possible paths**.
The degree of alignment of a norm $n$ with a value $v$ for agent $\alpha$ 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!!!
The degree of alignment of a norm $n$ with a value $v$ for agent $\alpha$ is the **accumulation of preferences** along the transitions.

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

$$\text{Align}_{n,v}^{\alpha}(S,A,T) = \sum_{p \in \text{paths}} \sum_{d \in [1, \text{length}(p)]} \Pr_{v}^{\alpha}(p_1[d], p_F[d]) \frac{\text{length}(p)}{\sum_{p \in \text{paths}} \text{length}(p)}$$

**Monte Carlo sampling is applied to address efficiency**
The relative alignment of norm $n_1$ w.r.t. norm $n_2$ is defined as the **difference** in their **alignments**!

$$\text{RA} \text{lgn}^\alpha_{n_1/n_2,v}(S, A, T) = \text{Algn}^\alpha_{n_1,v}(S, A, T) - \text{Algn}^\alpha_{n_2,v}(S, A, T)$$
With the right **aggregation functions**, and just like preferences, we can talk about alignment / relative alignment over **sets of values** & for **groups of people**?
Agents’ actions (cooperate (c) & defect (d)) results in certain gains. Let the relevant state parameters describe accumulated gains: (x,y)

<table>
<thead>
<tr>
<th></th>
<th>$\beta$ co-operates</th>
<th>$\beta$ defects</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$ co-operates</td>
<td>6,6</td>
<td>0,9</td>
</tr>
<tr>
<td>$\alpha$ defects</td>
<td>9,0</td>
<td>3,3</td>
</tr>
</tbody>
</table>
Value-based preferences.

1. States with higher equality in accumulated gain are preferred.

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

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

4. States with higher personal gain are preferred.

Example: Prisoner’s Dilemma
Example: Prisoner’s Dilemma

Value-based preferences.

1. \( \text{Prf}(s, s') = \frac{|x - y|}{\max\{x, y\}} - \frac{|x' - y'|}{\max\{x', y'\}} \)

2. \( \text{Prf}(s, s') = \left(1 - \frac{|y' - x'|}{\max\{x', y'\}}\right) \cdot \frac{x' - x}{\max\{x', x\}} \)

3. \( \text{Prf}(s, s') = \frac{x' - x}{2(\max\{x', x\})} - \frac{y' - y}{2(\max\{y', y\})} \)

4. \( \text{Prf}(s, s') = \frac{x' - x}{\max\{x', x\}} \)
Value-based preferences.

1. States with higher equality in accumulated gain are preferred.
2. States with higher equality in accumulated gain are preferred, only if my personal gain is not lower.
3. States with higher personal gain are preferred, only if equality is not lower.
4. States with higher personal gain are preferred.

Example: Prisoner’s Dilemma

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.
Value-based preferences.

1. States with higher equality in accumulated gain are preferred.

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

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

4. States with higher personal gain are preferred.

Example: Prisoner’s Dilemma

Norms.

(a) $n_0$ applied

(b) $n_1$ applied

(c) $n_2$ applied
Which norms are better aligned with an agent’s interpretation of ‘equality’?

3 norms: $n_0$, $n_1$, $n_2$

4 interpretations of ‘equality’: 1, 2, 3, 4
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’: ➊, ➋, ➌, ➍

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

<table>
<thead>
<tr>
<th>α’s actions</th>
<th>β’s actions</th>
<th>Relative Alignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>①</td>
<td>{c}</td>
<td>$n_1 &gt; n_0 \sim n_2$</td>
</tr>
<tr>
<td>②</td>
<td>{c}</td>
<td>$n_0 \sim n_1 \sim n_2$</td>
</tr>
<tr>
<td>③</td>
<td>{c}</td>
<td>$n_0 \sim n_1 \sim n_2$</td>
</tr>
<tr>
<td>④</td>
<td>{c}</td>
<td>$n_0 &gt; n_2 &gt; n_1$</td>
</tr>
<tr>
<td>①</td>
<td>{d}</td>
<td>$n_1 &gt; n_0 \sim n_2$</td>
</tr>
<tr>
<td>②</td>
<td>{d}</td>
<td>$n_0 \sim n_1 \sim n_2$</td>
</tr>
<tr>
<td>③</td>
<td>{d}</td>
<td>$n_0 \sim n_1 \sim n_2$</td>
</tr>
<tr>
<td>④</td>
<td>{d}</td>
<td>$n_0 \sim n_1 \sim n_2$</td>
</tr>
</tbody>
</table>
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’: \( 1, 2, 3, 4 \)

All norms \( (n_0, n_1, n_2) \) are equally aligned for moderate supporters of equality (2,3).
Which norms are better aligned with an agent’s interpretation of ‘equality’?

3 norms: \( n_0, n_1, n_2 \)

4 interpretations of ‘equality’: ➊, ➋, ➌, ➍

Example: Prisoner’s Dilemma

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 (➊, ➋, ➌, ➍).
Value Alignment

Montes & Sierra (2021 a)
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.
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.
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.
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

<table>
<thead>
<tr>
<th></th>
<th>$\beta$ co-operates</th>
<th>$\beta$ defects</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$ co-operates</td>
<td>6,6</td>
<td>0,9</td>
</tr>
<tr>
<td>$\alpha$ defects</td>
<td>9,0</td>
<td>3,3</td>
</tr>
</tbody>
</table>
Example: Prisoner’s Dilemma

Norms. Sierra et al. (2019)’s game

<table>
<thead>
<tr>
<th></th>
<th>$\beta$ co-operates</th>
<th>$\beta$ defects</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$ co-operates</td>
<td>6,6</td>
<td>0,9</td>
</tr>
<tr>
<td>$\alpha$ defects</td>
<td>9,0</td>
<td>3,3</td>
</tr>
</tbody>
</table>

\[
S = (x_\alpha, x_\beta)
\]

\[
S_1 = (x_\alpha + 6, x_\beta + 6)
\]

\[
S_2 = (x_\alpha, x_\beta + 9)
\]

\[
S_3 = (x_\alpha + 9, x_\beta)
\]

\[
S_4 = (x_\alpha + 3, x_\beta + 3)
\]
**Example: Prisoner’s Dilemma**

**Norms.** Sierra et al. (2019)’s game

<table>
<thead>
<tr>
<th></th>
<th>$\beta$ co-operates</th>
<th>$\beta$ defects</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$ co-operates</td>
<td>6,6</td>
<td>0,9</td>
</tr>
<tr>
<td>$\alpha$ defects</td>
<td>9,0</td>
<td>3,3</td>
</tr>
</tbody>
</table>

**Values.** Equality

\[
Prf_{equality}(s, s') = 1 - 4 \cdot GI(s') = 1 - 2 \cdot \frac{|x'_\alpha - x'_\beta|}{x'_\alpha + x'_\beta}
\]
Norms. Sierra et al. (2019)’s game

Values. Equality

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

Example: Prisoner’s Dilemma

\[
\text{Prf}_{\text{equality}}(s, s') = 1 - 4 \cdot GI(s') = 1 - 2 \cdot \frac{|x'_\alpha - x'_\beta|}{x'_\alpha + x'_\beta}
\]
Under random profiles, alignment is highest when both players have similar cooperation probabilities.

Under heterogeneous profiles, tit-for-tat results in stable alignment.
We have assessed value-aligned strategies.

Can we assess value-aligned norms?
Example: Prisoner’s Dilemma

Norms. Sierra et al. (2019)’s game

<table>
<thead>
<tr>
<th></th>
<th>$\beta$ co-operates</th>
<th>$\beta$ defects</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$ co-operates</td>
<td>6,6</td>
<td>0,9</td>
</tr>
<tr>
<td>$\alpha$ defects</td>
<td>9,0</td>
<td>3,3</td>
</tr>
</tbody>
</table>
Norms. Sierra et al. (2019)’s game

New Norms. ➊ Ban on 2 consecutive defections:
If a player defect twice in a row, then they are obliged to cooperate next.

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

Example: Prisoner’s Dilemma

<table>
<thead>
<tr>
<th></th>
<th>(\beta) co-operates</th>
<th>(\beta) defects</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\alpha) co-operates</td>
<td>6,6</td>
<td>0,9</td>
</tr>
<tr>
<td>(\alpha) defects</td>
<td>9,0</td>
<td>3,3</td>
</tr>
</tbody>
</table>

\[ s_1' (x_\alpha + 6, x_\beta + 6) \]
\[ s_3' (x_\alpha + 9, x_\beta) \]
\[ s_4' (x_\alpha + 3, x_\beta + 3) \]
Results

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!
A simple tax model.
Parametric Norms

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**
Equality.

\[
Prf_{\text{equality}}(s, s') = 1 - 4 \cdot GI(s') = 1 - 2 \cdot \frac{|x'_\alpha - x'_\beta|}{x'_\alpha + x'_\beta}
\]
Value-Aligned Norm Synthesis

Find the norm parameters that maximise alignment.

An Optimisation Problem.

\[
N^* = \arg\max_{N' \subseteq N} \text{Align}_{N', V}
\]

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

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

- Wealth evenly distributed at the end.
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.
What is the contribution of each individual norm to the overall alignment?

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

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

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

Serramia et al. (2018)
Value-Aligned Norm Selection

Values → Value aligned norm selection → Value-Aligned Norms

Candidate Norms → Value aligned norm selection → subset

Values promote Candidate Norms
Value-Aligned Norm Selection

Values

Candidate Norms

Value aligned norm selection

Value-Aligned Norms
The value system is a structure containing:

- a set of moral values
- preferences over these values (a ranking $\succeq$)
Value–Aligned Norm Selection

Values

Candidate Norms

Value aligned norm selection

Value–Aligned Norms
A **norm net** contains:

- a set of norms
- norm relations
A **norm net** contains:

- a set of norms
- norm relations:  □ **exclusivity**  
  □ **generalisation**
Value-Aligned Norm Selection

Values

Candidate Norms

Value aligned norm selection

Value-Aligned Norms
Objective of value-aligned norm selection is to find a norm system that:
- best aligns the values system
- is sound
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
Assess the **value alignment of a norm as a utility**, then the solution is the sound norm system that **maximises** its cumulative utility.
Norm utilities, that describe the value alignment of norms, depend on:

- Value utilities
- Norm-value utilities
\( v \geq v' \Rightarrow u(v) \geq u(v') \)
Value Utility

\[ u(v) = 1 + \sum_{v' \succ v} u(v') \]
Value Utility

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

Example.

\[ u(v_{sec}) = 1 \]
\[ u(v_{free}) = 2 \]
Norm-Value Utility

\[ u(n,v) \in [-1, 1] \]
### Example.

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>-0.2</td>
<td>-0.2</td>
</tr>
<tr>
<td>-1</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
</tr>
</tbody>
</table>
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.

\[ u(n) = \sum_v u(n,v) \cdot u(v) \]
Example.

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.2</td>
<td>-0.2</td>
<td>-0.2</td>
</tr>
<tr>
<td>-1</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
</tr>
</tbody>
</table>

\[
u(v_{sec}) = 1 \\
u(v_{free}) = 2\]
Example.

\[
\begin{array}{cccc}
1 & -0.2 & -0.2 & -0.2 \\
-1 & 0.7 & 0.7 & 0.7 \\
\end{array}
\]

\[
\begin{align*}
\text{\( u(v_{\text{sec}}) \)} &= 1 \\
\text{\( u(v_{\text{free}}) \)} &= 2
\end{align*}
\]

\[
\begin{align*}
\text{\( u(n_1) \)} &= 1 \\
\text{\( u(n_2) \)} &= 0.3 \\
\text{\( u(n_3) \)} &= 0.3 \\
\text{\( u(n_4) \)} &= 0.3
\end{align*}
\]
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.

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

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

Serramia et al. (2020)
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 some norm systems.
Get preferences over some norm systems.
Step 1. Preference Induction
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.
Step 2. Preference Grounding

1. Extend the preferences

[Diagram showing a series of icons and symbols connected by arrows, indicating a process.]
2. Extract Equivalence Classes
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

(1,0,7) ≥ (0,1,7) ~ (0,1,7) ~ (0,1,7)
Step 2. Preference Grounding
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
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
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.
We have designed an **alignment** formula that gives the value alignment of any norm system satisfying:

\[ \Omega \succeq \Omega' \iff al(\Omega) \geq al(\Omega') \]

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

\[ \text{max} \quad al(\{ n_1 \}) x_1 + \ldots + al(\{ n_{|N|} \}) x_{|N|} \]
Take Home Message #4

Value Alignment Mechanisms

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


Thank you!

Nardine Osman
nardine@iiiia.csic.es