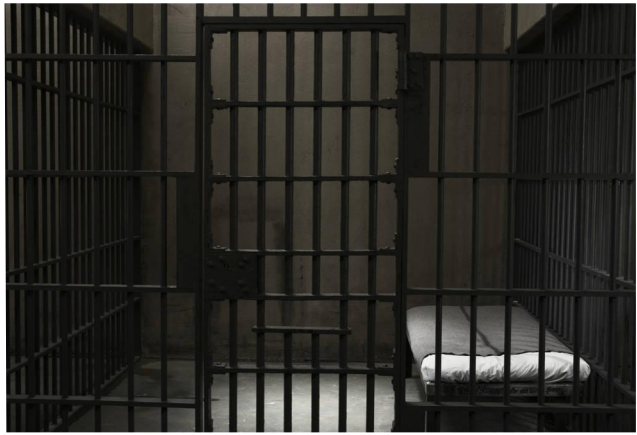
Responsible Autonomy

Carles Sierra
Artificial Intelligence Research Institute (IIIA-CSIC)
Barcelona

AI is profoundly impacting our lives and our cities

self-driving cars

medical diagnosis



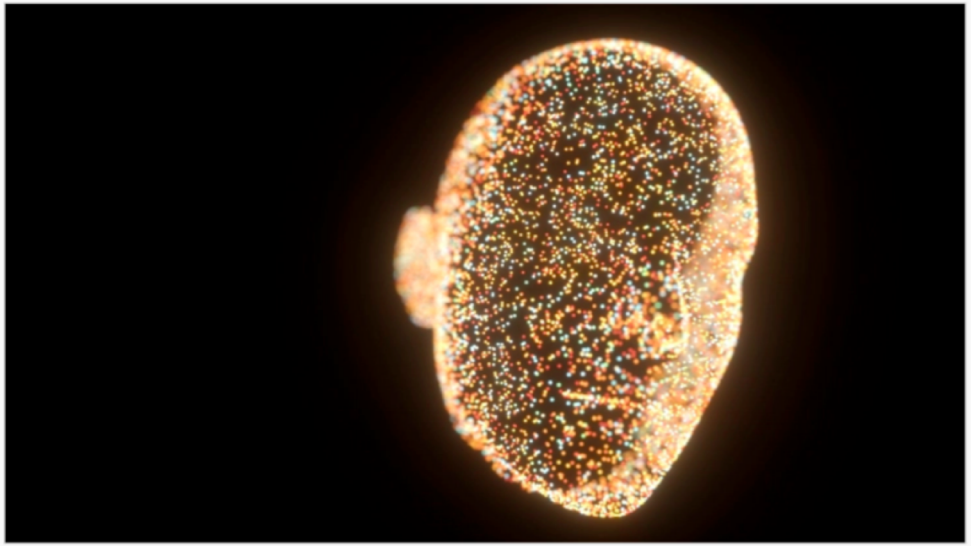
parole decisions

Ethical Concerns

CRUNCH NETWORK

Ethics — the next frontier for artificial intelligence

Posted Jan 22, 2017 by Don Basile (@TheDonBasile)



Don Basile
CONTRIBUTOR

Don Basile is an entrepreneur and venture capitalist with more than 20 years of executive experience in technology, healthcare and

AI's next frontier requires ethics built through policy. Will Donald Trump deliver?

With one foot in its science fiction past and the other in the new frontier of science and tech innovations, AI occupies a unique place in our

Ethical Concerns

The Guardian

The Cambridge Analytica Files

Cambridge Analytica: how did it turn clicks into votes?

Whistleblower Christopher Wylie explains the science behind Cambridge Analytica's mission to transform surveys and Facebook data into a political messaging weapon

Ethical Concerns

Concern over Singapore's anti-fake news law

Karishma Vaswani
Asia business correspondent
@BBCKarishma

4 April 2019 Share

Does the law hand too much power to the Singapore government?

This week Singapore's government proposed its anti-fake news law in parliament - the Protection from Online Falsehoods and Manipulation Bill.

The government says the law is necessary to protect Singaporeans from fake news

NEWS

EU tells social media giants to combat fake news or face new regulations

The EU's executive arm has outlined guidelines requesting social media companies to self-regulate the spread of fake news. The companies could be forced to combat the problem if they don't.

Social media companies such as Facebook or Twitter must stop fake news online or risk exposing themselves to new EU regulations, the bloc said on Thursday.

The move has come amid fears Russia could follow up its alleged attempt to sway the 2016 US

Ethical Concerns

> “Knight Capital's automated trading system
> is much less intelligent than Google DeepMind's AlphaGo,
> but the former lost $460 million in just forty-five minutes.
> AlphaGo hasn't and can't hurt anyone.”

Professor Dan Weld
University of Washington

Ethical Concerns

Not JUST privacy, security, & manipulation!

We are also concerned about basic features and functionality.

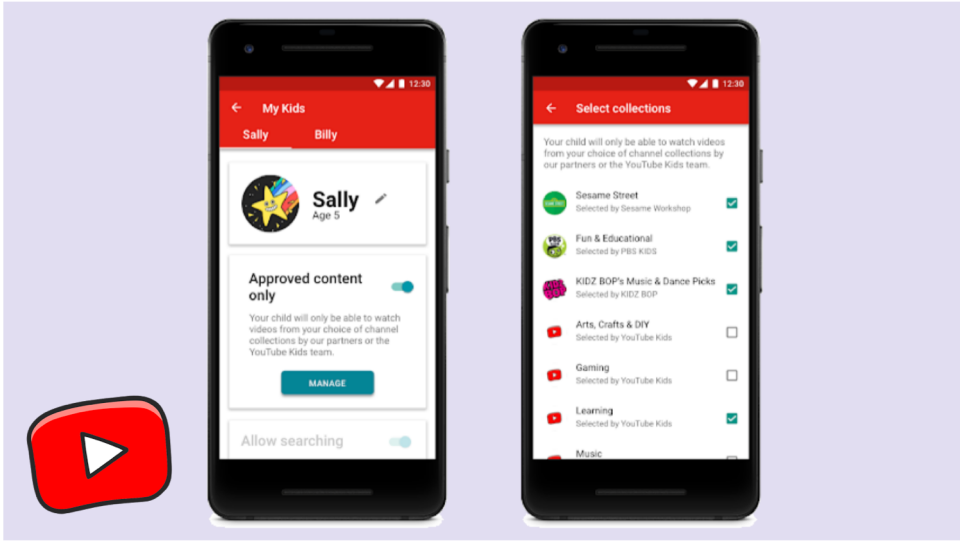
VARIETY

HOME > DIGITAL > NEWS

APRIL 26, 2018 6:07AM PT

After Complaints, YouTube Kids App Will Finally Let Parents Fully Lock Down What Their Children Can Watch

By TODD SPANGLER



CREDIT: YOUTUBE

More than three years after launching the tyke-targeted YouTube Kids app — which has turned out to not as clean and well-lit as YouTube had initially touted — the video giant is going to introduce features to help parents handpick exactly what content their children are allowed t

Privacy settings

Can we build Responsible Autonomous Systems?
Can we put humans in control?

Can we build Responsible Autonomous Systems?
Can we put humans in control?

AGENDA:

- Multiagent Responsible technologies
- Ethical code and self-regulated communities.
- A Roadmap to Responsible Autonomy.
- Value-Alignment
- Wrap-up

Multiagent Responsible Technologies

Responsible Research

“ **research and innovation must respond to the needs and ambitions of society, reflect its values, and be responsible** ”

European Commission on
Responsible Research & Innovation

Responsible Technologies

“

technologies that respond to
the needs and ambitions of society, reflect its values,
and put people in control.

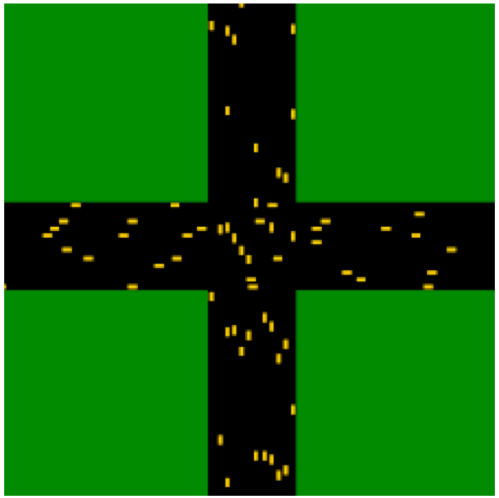
”

proposed definition for
Responsible Technologies

To put people in control, because AI must be social

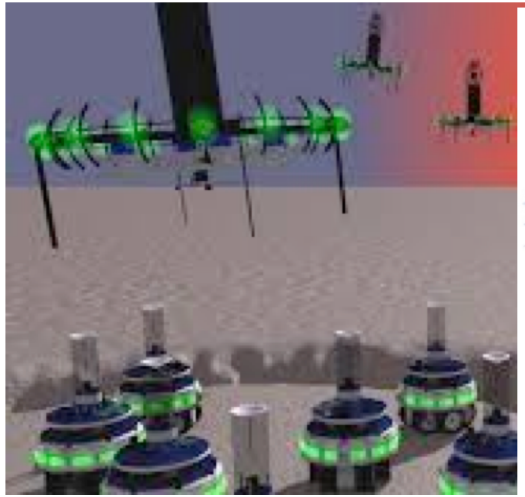
Billions of AI systems will interact among themselves and with humans. Our future society will be a colossal Multiagent System, a huge **sociotechnical community.**

Traffic



Kurt Dresner and Peter Stone

Multi-robot



IRIDIA Lab

IoT

MAS: meeting point for AI (technology) and Humanities (people).

From individual rationality to social intelligence we need:

- Communicative interaction
- Social Co-ordination
- Agreement technologies
- Social networks
- Social choice
- Agent-based modelling
- Social simulation

Matthew Yee-King, Roberto Confalonieri, Dave de Jonge, Katina Hazelden, Carles Sierra, Mark d'Inverno, Leila Amgoud, Nardine Osman:
Multiuser museum interactives for shared cultural experiences: an agent-based approach. AAMAS 2013: 917-924

But how to guarantee responsible behaviour when entities are autonomous?

- Responsible behaviour is a social convention. No universals; it is context dependent. It relates to the particular shared values of the community members.
- No individual behaviour guarantee can be obtained when systems are fully autonomous, but we can design **sociotechnical communities** so that unacceptable behaviour generates **repair actions** and **punishements**. (This is the **legal approach**.) And, desirable behaviour is geared via **incentives**. (This is the **economic approach**.)

But how to guarantee responsible behaviour when entities are autonomous?

- Responsible behaviour is a social convention. No universals; it is context dependent. It relates to the particular shared values of the community members.
- No individual behaviour guarantee can be obtained when systems are fully autonomous, but we can design **sociotechnical communities** so that unacceptable behaviour generates **repair actions** and **punishements**. (This is the **legal approach**.) And, desirable behaviour is geared via **incentives**. (This is the **economic approach**.)

Let's get inspiration from how we humans model responsible behaviour.

Legal Relations

	(1)	(2)	(3)	(4)
JURAL OPPOSITES	Right	Privilege	Power	Immunity
	No-right	Duty	Disability	Liability

	(1)	(2)	(3)	(4)
JURAL CORRELATIVES	Right	Privilege	Power	Immunity
	Duty	No-right	Liability	Disability

Wesley Newcomb Hohfeld.
Fundamental Legal Conceptions as Applied in Judicial Reasoning, 23 YALE L.J. 16 (1913).

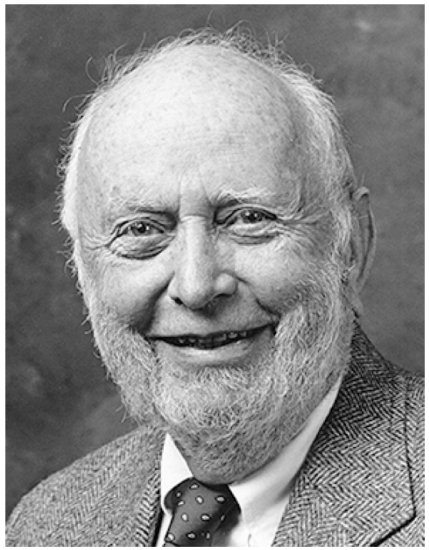
Legal Knowledge Representation in Hohfeld

- Basic deontic operators
- Power
- Multi agency
- Time

Hohfeld's Fundamental Legal Conceptions	LEGAL RELATIONS (LR)* * Defined terms in the LEGAL RELATIONS Language (LRL) are in upper case.		
	Unconditional (Deontic)	CONDITIONAL	
		Capacitive	Other CONDITIONAL
duty	DUTY(s,a,b)		CONDITIONAL(c,DUTY(s,a,b))
right	RIGHT(s,b,a)		CONDITIONAL(c,RIGHT(s,b,a))
privilege	PRIVILEGE(s,a,b)		CONDITIONAL(c,PRIVILEGE(s,a,b))
no-right	NO_RIGHT(s,b,a)		CONDITIONAL(c,NO_RIGHT(s,b,a))
	There are 1588 other different deontic LRs.	There are an infinite number of other capacitive LRs.	There are an infinite number of other noncapacitive CONDITIONAL LR, i.e., CONDITIONAL(c,LR).
power		POWER(D2(x,b),LR)	CONDITIONAL(c,POWER(D2(x,b),LR))
liability		LIABILITY(LR,D2(x,b))	CONDITIONAL(c,LIABILITY(LR,D2(x,b)))
disability		DISABILITY(D2(x,b),LR)	CONDITIONAL(c,DISABILITY(D2(x,b),LR))
immunity		IMMUNITY(LR,D2(x,b))	CONDITIONAL(c,IMMUNITY(LR,D2(x,b)))

Legal Relation Language by Layman E. Allen. Applied Deontic Logic.

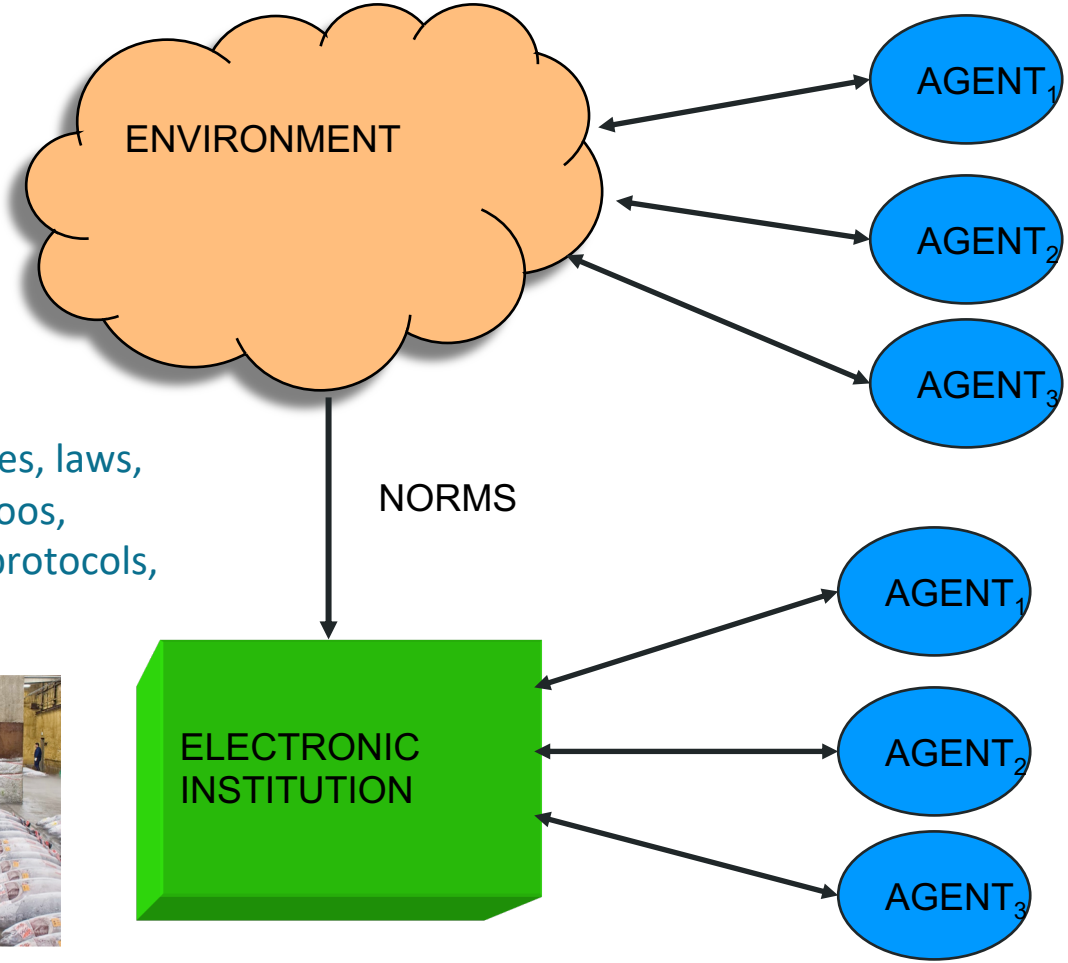
New Institutional Economics



Douglass North: "Transaction costs, institutions, and economic performance." (1992)

"humanly devised constraints that structure political, economic and social interactions".

Electronic Institutions

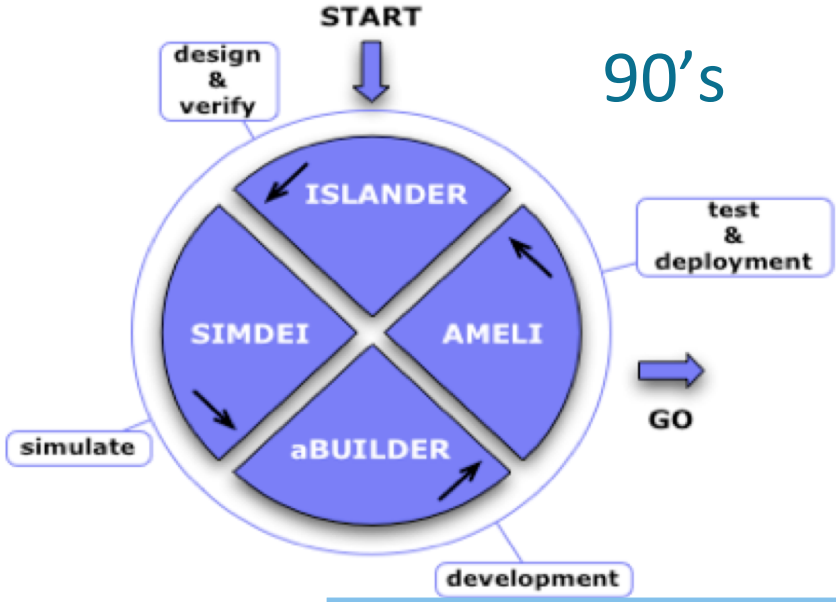


Formal rules, laws, rights, taboos, customs, protocols, …

Electronic institutions

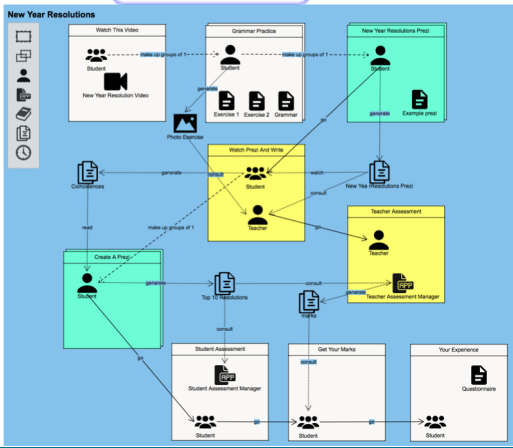
- Populated by heterogeneous agents, developed by different people, using different languages and architectures
- Self-interested agents
- Participants change over time and are unknown in advance

The city of Uruk



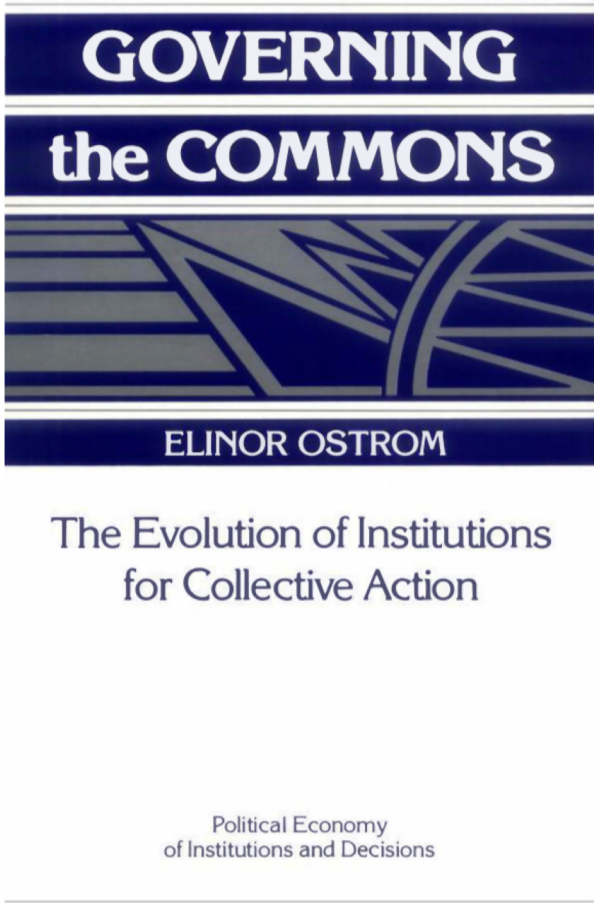
90's

10's



Mark d'Inverno, Michael Luck, Pablo Noriega, Juan A. Rodríguez-Aguilar, Carles Sierra:
Communicating open systems. Artif. Intell. 186: 38-94 (2012)

Sustainable Collective Action. Self-Governing Institutions.



L'Horta watering communities

- May 29, 1435, 84 irrigators approved formal regulations on how to share water.
- Some rules had been in use from much earlier.
- Rules talk about maintenance, fines, officials, and use of water depending on the environment.
- They are an example of situatedness.

Figure 3.1. Location of Spanish *huertas*.

Human communities are often successful

Ostrom's principles and the *Horta*

Boundaries: irrigation rights come with the land.
Appropiation and provision: proportional to size of land.
Collective choice: election of officials in the court.
Monitoring: 'turno' system makes monitoring high and easy.
Sanctions: surprisingly low frequency. 0,8%.
Conflict: weekly meetings.
Rights to organise: no external interference

Table 3.1. *Design principles illustrated by long-enduring CPR institutions*

1. Clearly defined boundaries
   Individuals or households who have rights to withdraw resource units from the CPR must be clearly defined, as must the boundaries of the CPR itself.
2. Congruence between appropriation and provision rules and local conditions
   Appropriation rules restricting time, place, technology, and/or quantity of resource units are related to local conditions and to provision rules requiring labor, material, and/or money.
3. Collective-choice arrangements
   Most individuals affected by the operational rules can participate in modifying the operational rules.
4. Monitoring
   Monitors, who actively audit CPR conditions and appropriator behavior, are accountable to the appropriators or are the appropriators.
5. Graduated sanctions
   Appropriators who violate operational rules are likely to be assessed graduated sanctions (depending on the seriousness and context of the offense) by other appropriators, by officials accountable to these appropriators, or by both.
6. Conflict-resolution mechanisms
   Appropriators and their officials have rapid access to low-cost local arenas to resolve conflicts among appropriators or between appropriators and officials.
7. Minimal recognition of rights to organize
   The rights of appropriators to devise their own institutions are not challenged by external governmental authorities.

For CPRs that are parts of larger systems:

8. Nested enterprises
   Appropriation, provision, monitoring, enforcement, conflict resolution, and governance activities are organized in multiple layers of nested enterprises.

Ethical code and self-regulated communities.

What is an ethical code

- The norms that regulate the behaviour of communities. They are of different sorts
  - Legal (institutional) norms. Imposed.
  - Community norms. Based on shared values, collective behaviour.
  - Individual norms. Based on individual preferences and values.
- Behaviour and the environment impact the fulfilment of needs and the adherence to values. The ethical code must be dynamic. Change is triggered by unsatisfied needs and evolving values.

Legal Norms. Hammurabi code. 1754 BCE.

Retributive justice. TFT.

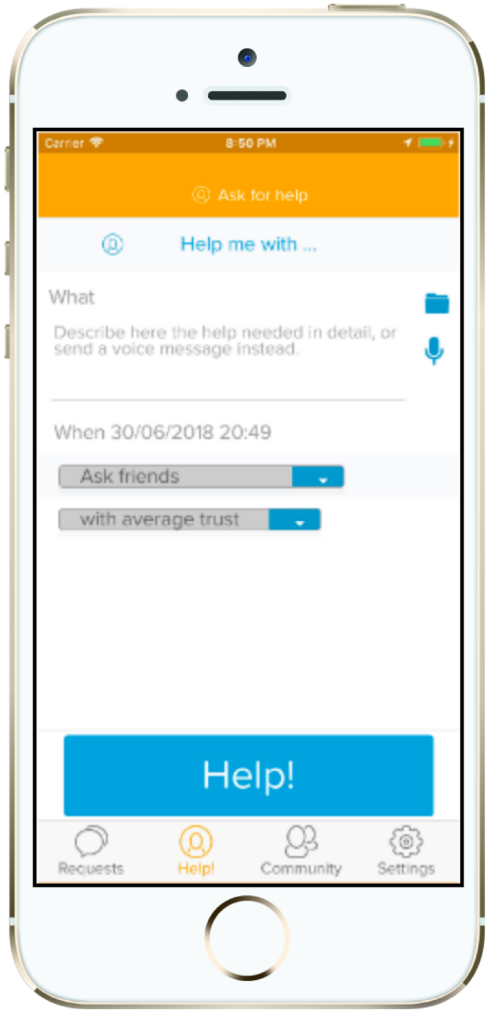
Law 196: If a man destroy the eye of another man, they shall destroy his eye. If one break a man's bone, they shall break his bone. If one destroy the eye of a freeman or break the bone of a freeman he shall pay one gold mina. If one destroy the eye of a man's slave or break a bone of a man's slave he shall pay one-half his price.

Community norms

- Each farm on a canal receives water in a rotation order.
- If a farmer fails to open his headgate when the water arrives there, he misses his turn and must wait for the next rotation.
- Each farmer decides how much water to take.
- The households to receive timber form teams and equaly divide the work.
- Workers will make equaly sized piles of logs.
- A lottery determines which pile goes to which household.

Individual norms

- Don’t show me messages during my afternoon nap
- Don’t show me messages from people that are not in my contact list.
- *Don’t show me requests coming from men.*



Formalisms for normative systems.

If-Then rules (e.g. Hammurabbi)

Conditional Deontic Logic with Deadlines

Event Calculus

Hybrid Metric Interval Temporal Logic

Social Integrity Constraints

Object Constraint Language

Constraint rule-based

Normative Temporal Logic

Constraint rule-based

Punishment – We must punish those agents when issuing a winning bid they cannot pay for. More precisely, the rule punishes an agent A1 by decreasing its credit of 10% of the value of the good being auctioned. The oav predicate on the LHS of the rule represents the current credit of the offending agent. The rule also adds an obligation for the auctioneer to restart the bidding round and the constraint that the new offer should be greater than 120% of the old price.

$$\left(\begin{array}{c} X = \left\{ \begin{array}{l|l} \alpha_0 & \alpha_1 \,\&\, (T_0 > T_1) \,\& \\ & \mathtt{not}(\alpha_2 \,\&\, (T_2 > T_1)) \end{array} \right\} \& \\ oav(A_1, credit, C) \,\& \\ (\mathtt{size}(X) = 1) \,\&\, (C < P) \,\& \\ C2 = C - P * 0.1 \end{array} \right) \rightsquigarrow \left(\begin{array}{c} \mathtt{del}(oav(A_1, credit, C)), \\ \mathtt{add}(oav(A_1, credit, C2)), \\ \mathtt{add}(\alpha_3) \end{array} \right)$$

$$where \left\{ \begin{array}{l} \alpha_0 = utt(dutch, w_4, inform(A_1, buyer, Au, auct, bid(It, P), T_0)) \\ \alpha_1 = utt(dutch, w_3, inform(Au, auct, all, buyer, offer(It, P), T_1)), \\ \alpha_2 = utt(dutch, w_3, inform(Au, auct, all, buyer, offer(It, P), T_2)) \\ \alpha_3 = obl(dutch, w_5, inform(Au, auct, all, buyer, offer(It, P * 1.2), T_3)) \end{array} \right.$$

Andrés García-Camino, Juan A. Rodríguez-Aguilar, Carles Sierra, Wamberto Weber Vasconcelos:
Constraint rule-based programming of norms for electronic institutions.
Autonomous Agents and Multi-Agent Systems 18(1): 186-217 (2009)

Normative Temporal Logic. SNL.

module *toggle* controls x
  init
  $\ell_1 : \top \rightsquigarrow x' := \top$
  $\ell_2 : \top \rightsquigarrow x' := \bot$
  update
  $\ell_3 : x \rightsquigarrow x' := \bot$
  $\ell_4 : (\neg x) \rightsquigarrow x' := \top$

normative-system *id*
  χ_1 disables $\ell_{1_1}, \ldots, \ell_{1_k}$
  ...
  χ_m disables $\ell_{m_1}, \ldots, \ell_{m_k}$

Thomas Ågotnes, Wiebe van der Hoek, Juan A. Rodríguez-Aguilar, Carles Sierra, Michael J. Wooldridge: On the Logic of Normative Systems. IJCAI 2007: 1175-1180

Normative Temporal Logic. SNL.

module *toggle* controls x
  init
  $\ell_1 : \top \rightsquigarrow x' := \top$
  $\ell_2 : \top \rightsquigarrow x' := \bot$
  update
  $\ell_3 : x \rightsquigarrow x' := \bot$
  $\ell_4 : (\neg x) \rightsquigarrow x' := \top$

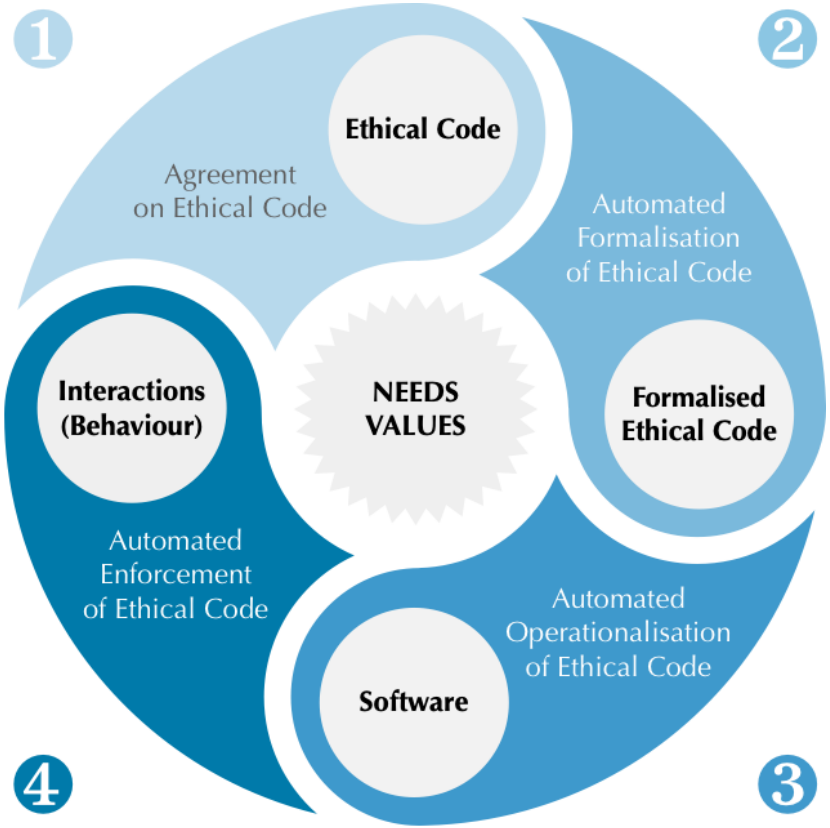
normative-system *id*
  χ_1 disables $\ell_{1_1}, \ldots, \ell_{1_k}$
  ...
  χ_m disables $\ell_{m_1}, \ldots, \ell_{m_k}$

But

Maybe more expressivity is needed, based on Hohfeld, blending *Deontic*, *power*, *multiagent*, and *temporal* concepts.

Thomas Ågotnes, Wiebe van der Hoek, Juan A. Rodríguez-Aguilar, Carles Sierra, Michael J. Wooldridge: On the Logic of Normative Systems. IJCAI 2007: 1175-1180

Responsible autonomy life-cycle



Some illustrative examples

1 Birth of Norms:

Members of the Anthropology Class of 2019 agree on a new norm:

Winning norm:
If someone uploads a photo, then only they can add tags.

■ ■ ■

Maya

Voting Trigger:
It seems each one has presented their view and discussed it. Let us vote.

■ ■ ■

Anna

Norm Suggestion:
What about restricting who can tag. Maybe the owner of the photo?

Anna

Argument:
I think disabling tagging is too strict.

Mark

Norm Suggestion:
I suggest to disable tagging!

Dave

Opinion:
Me too! My photos page is cluttered!

Anna

Evolution Trigger:
I am not happy that anyone can tag anyone else in a photo. I suggest we change this rule.

2 Norm Formalisation (automated):

The norm in [restricted] natural language is formalised.

$$\text{upload\_photo}(\text{Someone}, \text{Photo}) \Longrightarrow \neg\ \text{tag}(\text{SomeoneElse}, \text{Photo}, \text{TaggedPerson}) \wedge \text{SomeoneElse} \neq \text{Someone}$$

3 Norm Operationalisation (automated):

The formal norm is operationalised.

```
alert("You cannot tag this photo. Only
the owner can tag this photo.");
```

4 Norm Enforcement (automated):

The photo cannot be tagged by anyone other than the owner.

You cannot tag this photo.
Only the owner can tag the photo.

Ok

1 Birth of Norms:

Members of the Anthropology Class of 2019 modify a norm:

Winning norm:
If the tagged person does not accept to be tagged, then the tag is not added.

■ ■ ■

Maya

Voting Trigger:
It seems no one objects. Please confirm by voting.

■ ■ ■

Maya

Argument:
I look horrible in some photos! I should be consulted first.

Anna

Argument:
This seems reasonable.

Dave

Norm Suggestion:
I suggest not to tag anyone without their consent!

Mark

Opinion:
Fully agree!

Dave

Evolution Trigger:
I am not happy I am being tagged without my consent. I suggest we change this.

2 Norm Formalisation (automated):

The norm in [restricted] natural language is formalised.

$$\neg \text{ accept\_tag(TaggedPerson, Photo)} \Longrightarrow \neg \text{ add\_tag(Photo, TaggedPerson)}$$

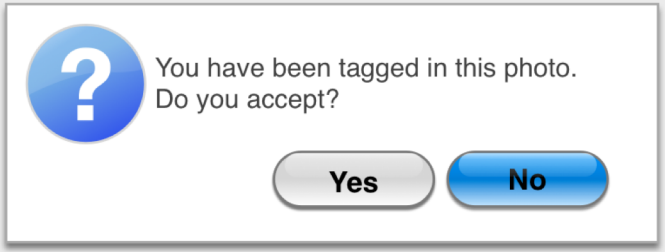
3 Norm Operationalisation (automated):

The formal norm is operationalised.

```
if (confirm('You have been tagged in
this photo. Do you accept?')) {
    addTag();
} else {
    deleteTag();
}
```

4 Norm Enforcement (automated):

The photo is not tagged before the user being tagged accepts.

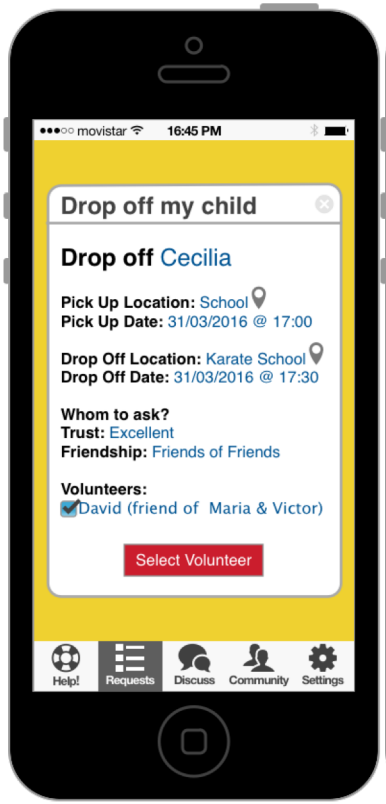


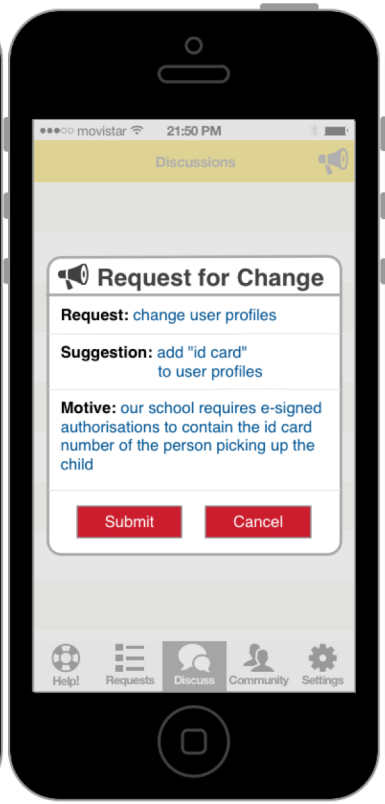
Single mothers community in uHelp.

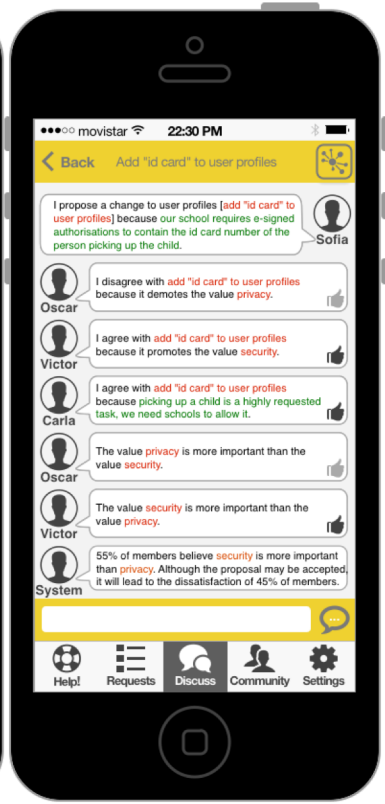
Download on the App Store

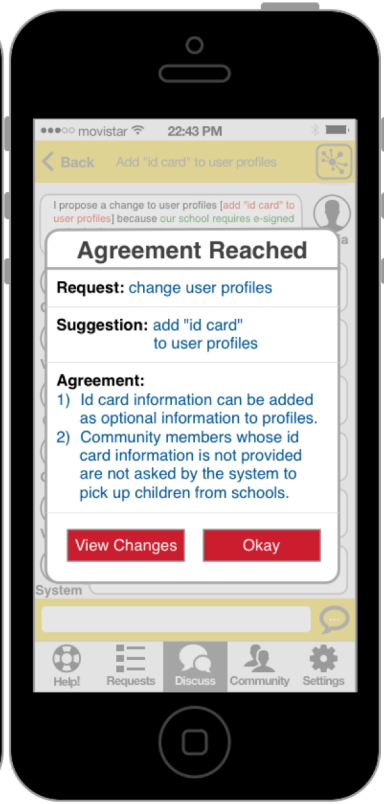
uHelp

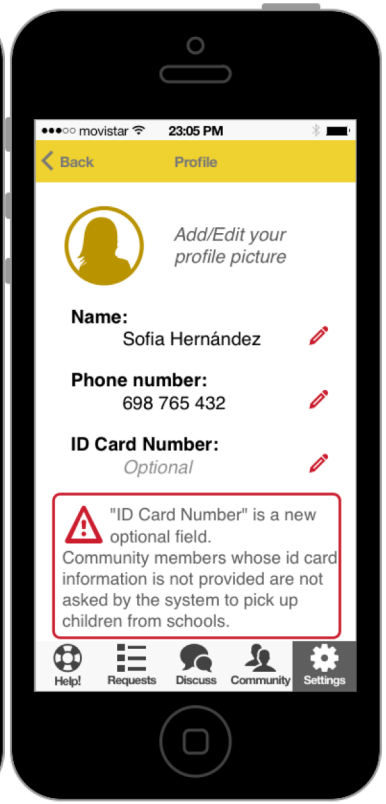
GET IT ON Google play





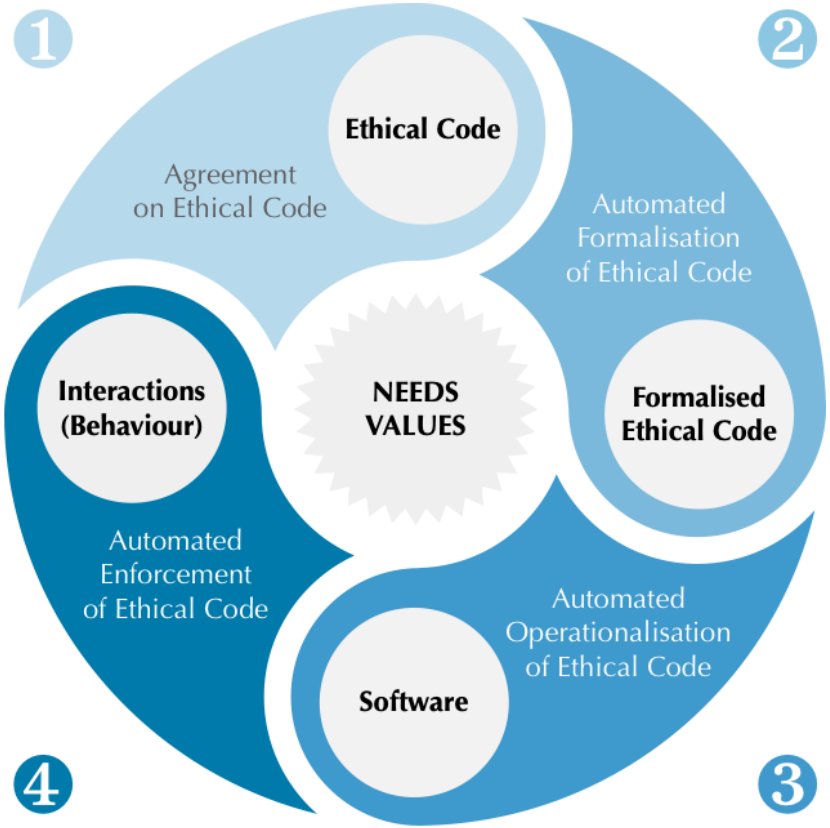






A Roadmap to Responsible Autonomy. Combination of techniques.

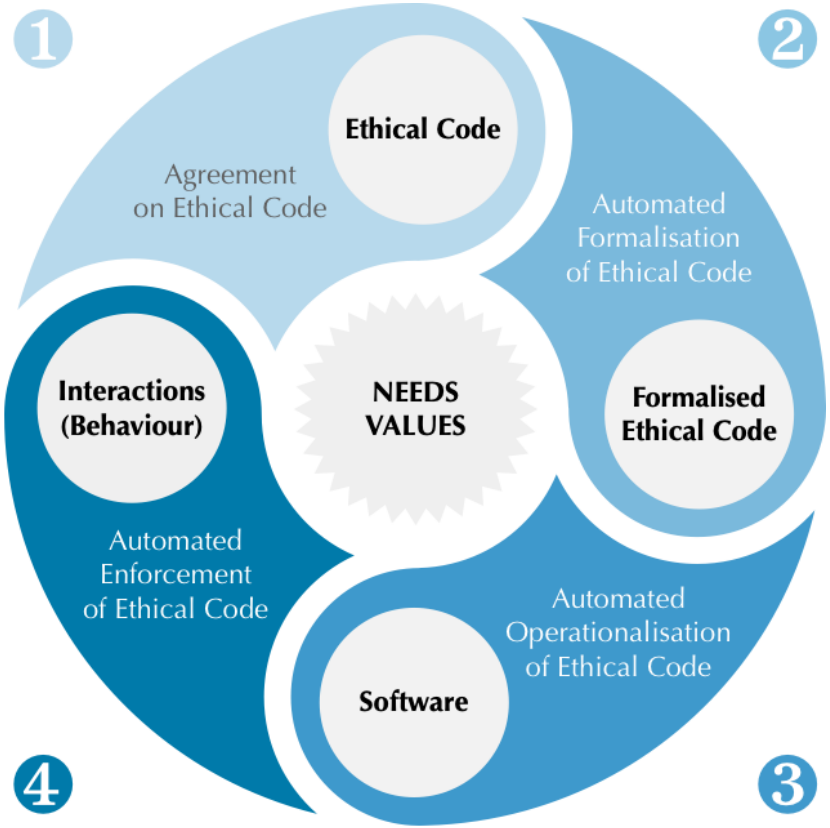
The Roadmap


1
Agreement
on Ethical Code
Ethical Code
2
Automated
Formalisation
of Ethical Code
Formalised
Ethical Code
3
Automated
Operationalisation
of Ethical Code
Software
4
Automated
Enforcement
of Ethical Code
Interactions
(Behaviour)
NEEDS
VALUES

The Roadmap

Agreement Technologies

Argumentation
Negotiation
Trust & Reputation
Computational Social Choice
Value alignment



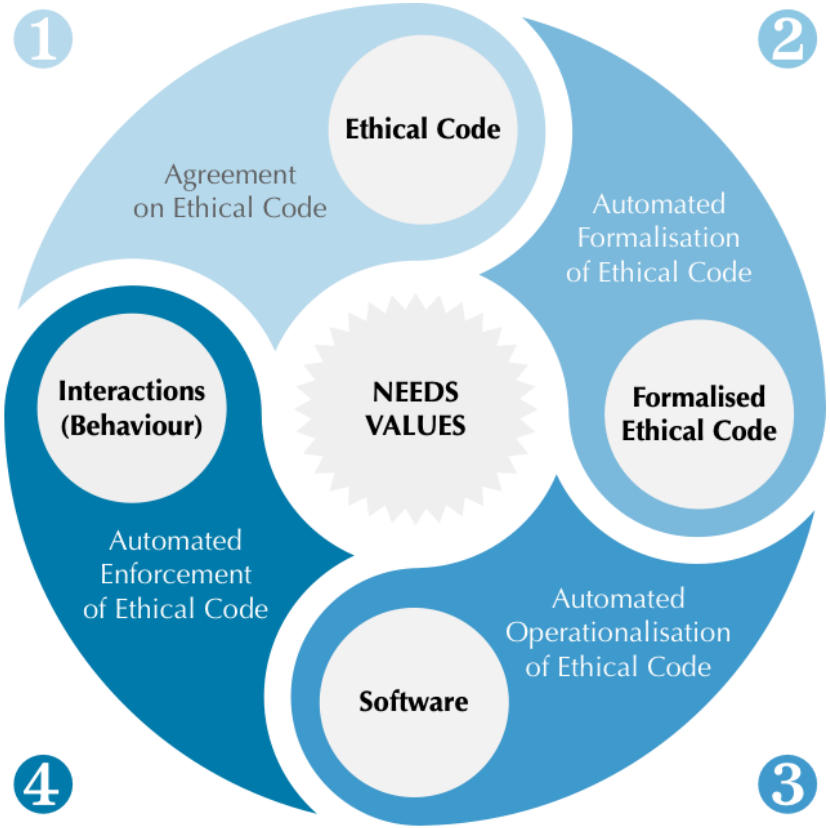
The Roadmap

Agreement Technologies

Learning

Learn when to change norms
Learn the best norms
Learn norm consequences

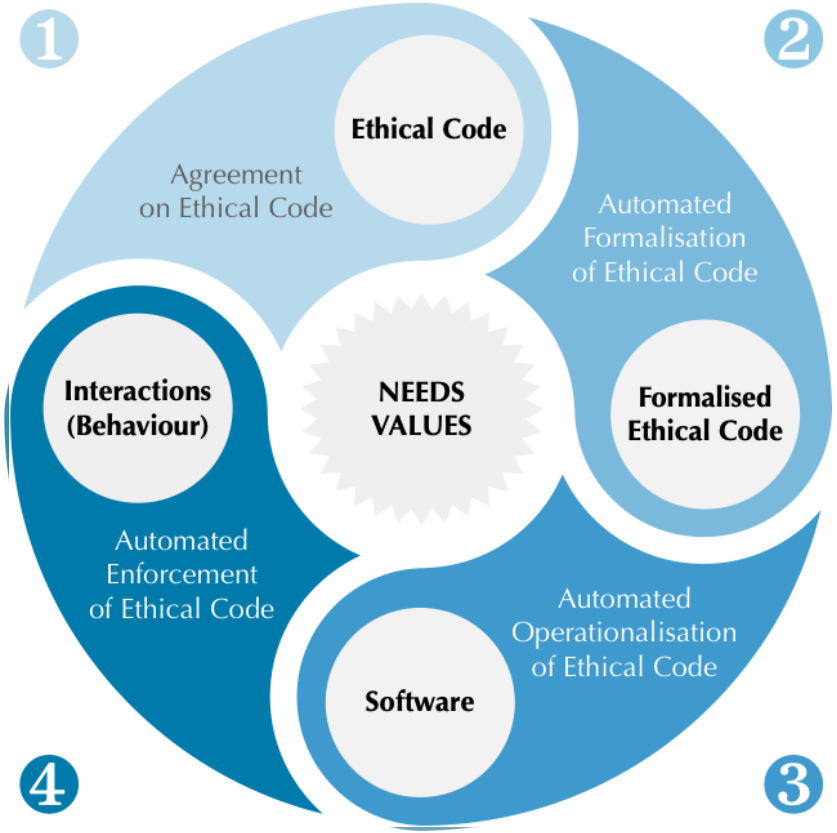
ML / CBR / simulations / sentiment analysis / analogical reasoning / coherence theory / norm synthesis



The Roadmap

Agreement Technologies

Learning



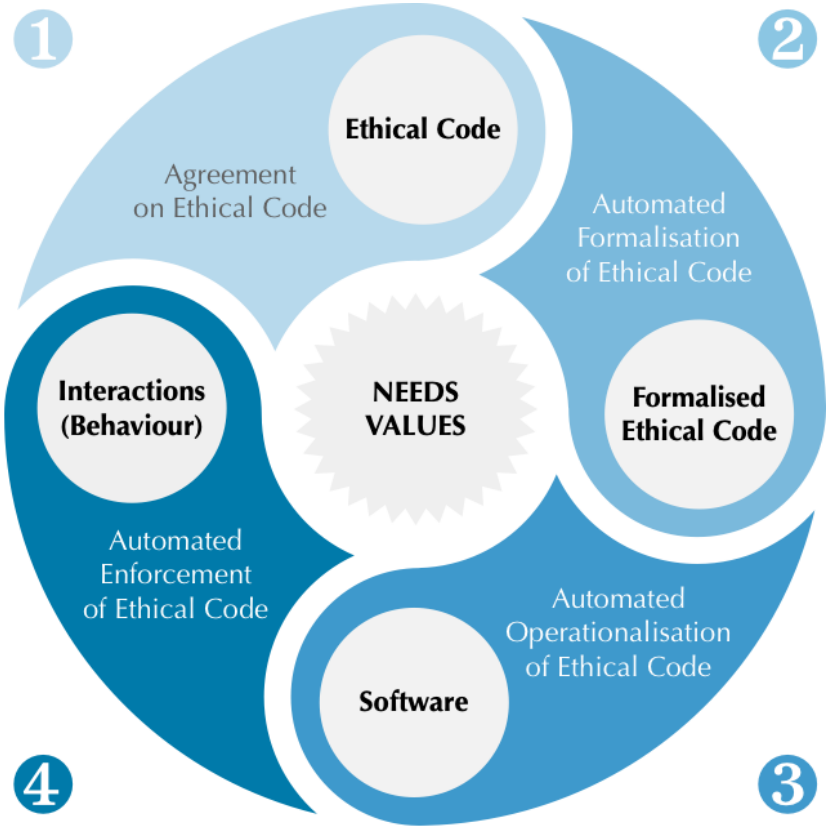
Logic for Norms

First Order Logic
Modal Logic
Deontic Logic
…
Hohfeldian Logic

The Roadmap

Agreement Technologies

Learning



Logic for Norms

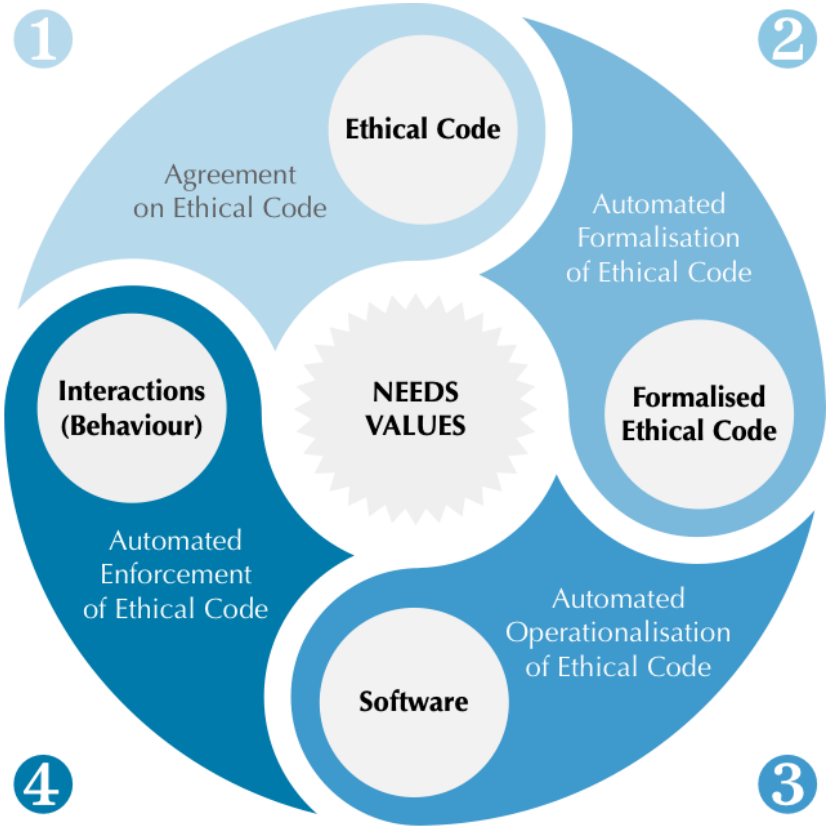
Natural Language Processing

Recognising norms
Extracting modalities & their parameters

The Roadmap

Agreement Technologies

Learning



Logic for Norms

Natural Language Processing

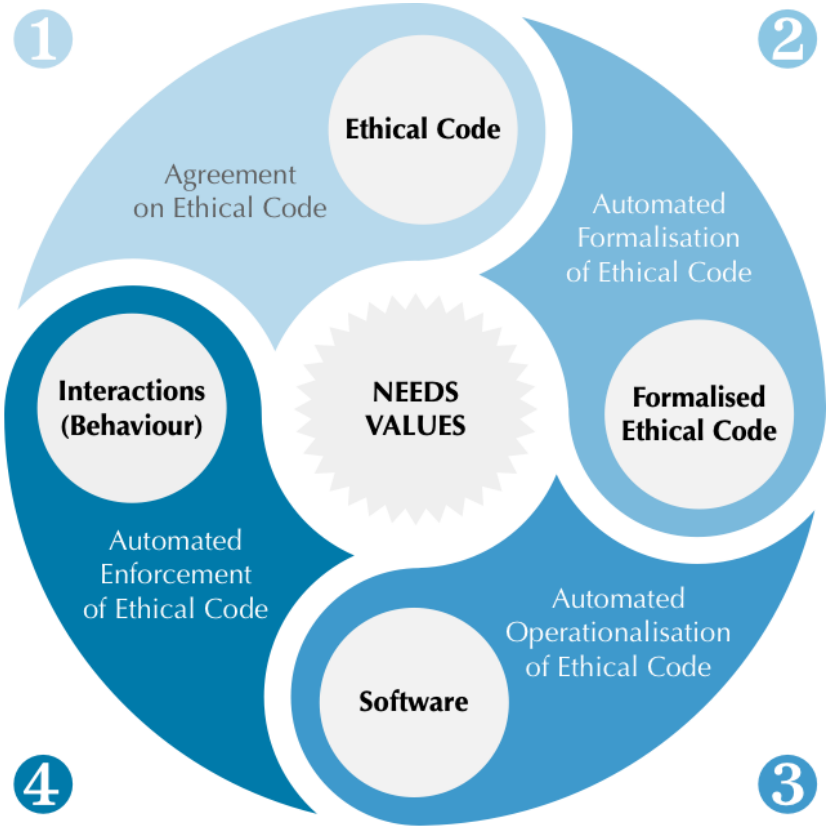
Normative Systems

Processes
If-Then statements
Constraints
SNL
...

The Roadmap

Agreement Technologies

Learning



Logic for Norms

Natural Language Processing

Normative Systems

Formal Verification

Model Checking /
Automated Theorem Proving

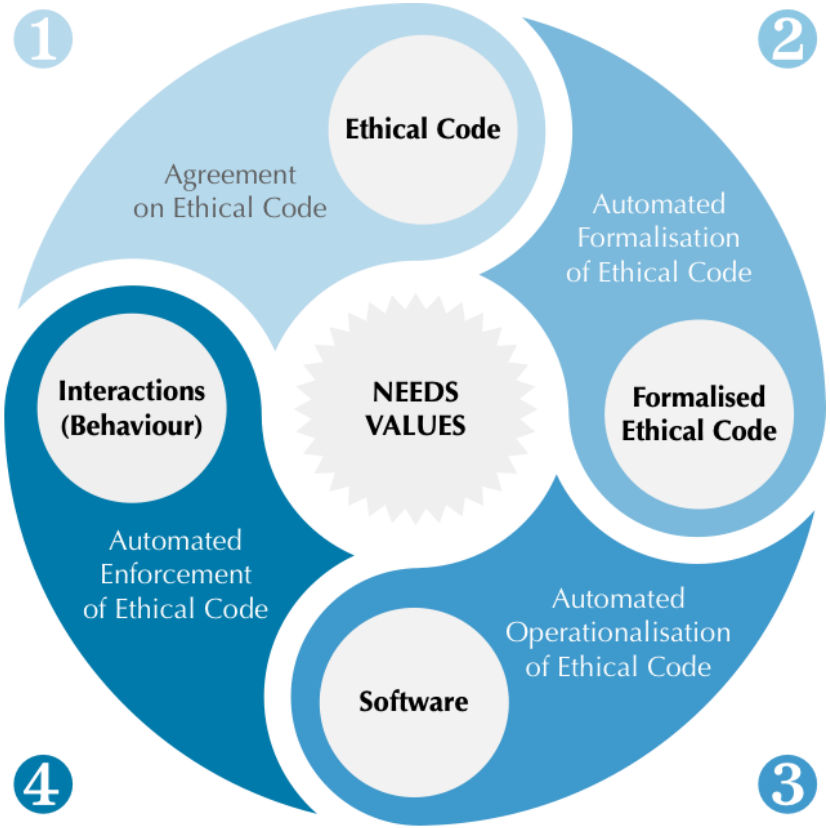
The Roadmap

Agreement Technologies

Learning

Norm Enforcement

Providing incentives to comply
"Punish" defects



Logic for Norms

Natural Language Processing

Normative Systems

Formal Verification

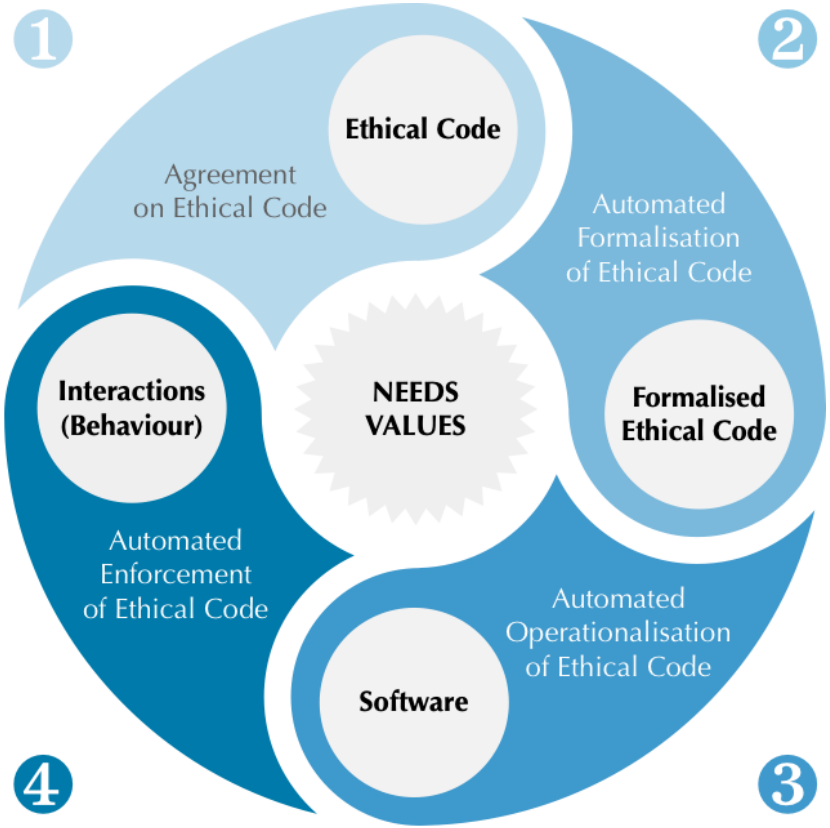
The Roadmap

Agreement Technologies

Learning

Norm Enforcement

Automated GUIs



Logic for Norms

Natural Language Processing

Normative Systems

Formal Verification

Every component is difficult.
One element of the roadmap:
Value Alignment - one of the main issues in Responsible AI today

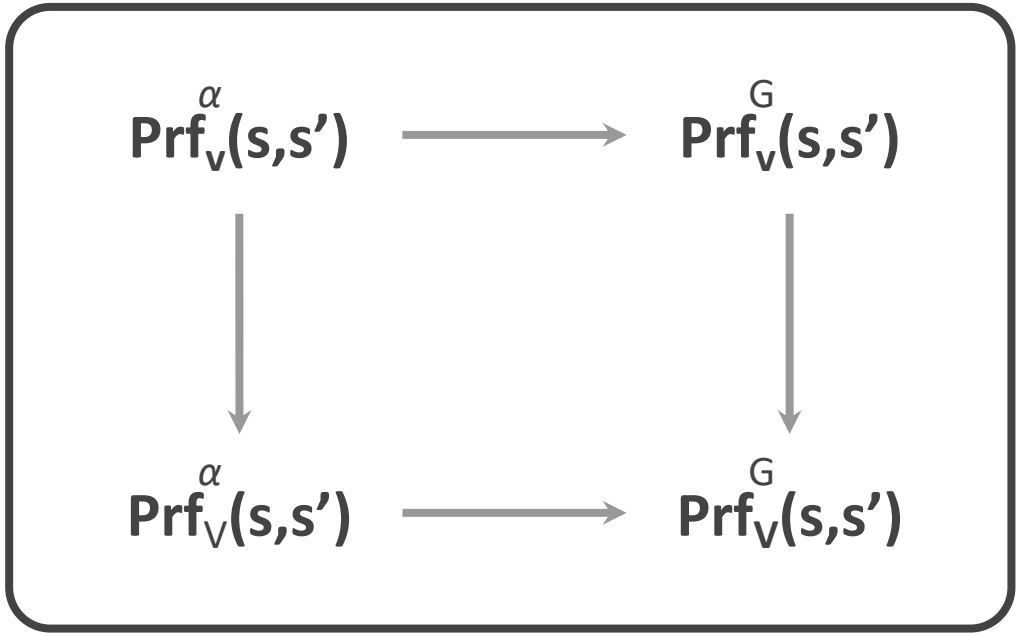
Carles Sierra, Nardine Osman, Pablo Noriega, Jordi Sabater Mir and Antoni Perello-Moragues

Value alignment: a formal approach
RAIA Workshop, AAMAS 2019

Values as preferences

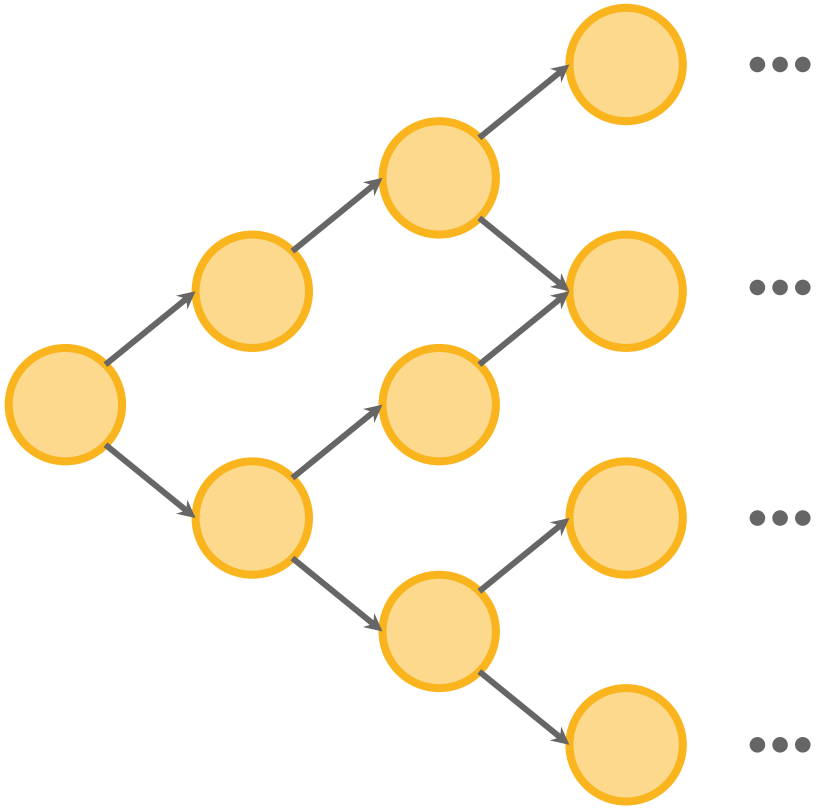
Values are understood as preferences over behaviour, or preferences over the states of the world: $\mathbf{Prf_v(s,s')}^{\alpha}$

Aggregation of value-based preferences



Value alignment problem: the concept

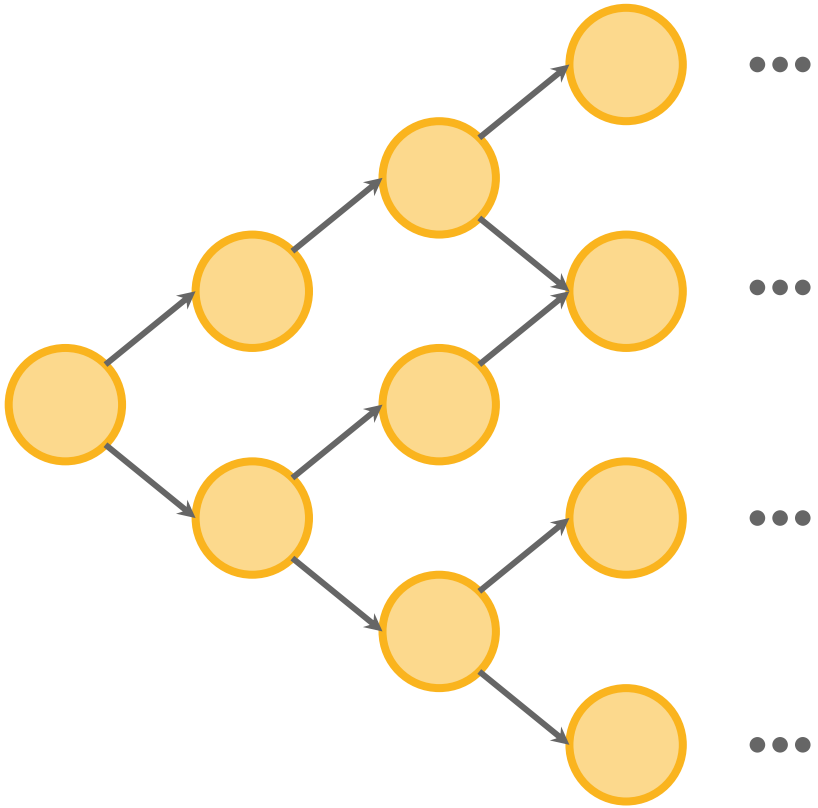
One is aligned with a value if their actions move them towards preferred states.



Value alignment problem: the concept

One is aligned with a value if their actions move them towards preferred states.

Actions get one to preferred states

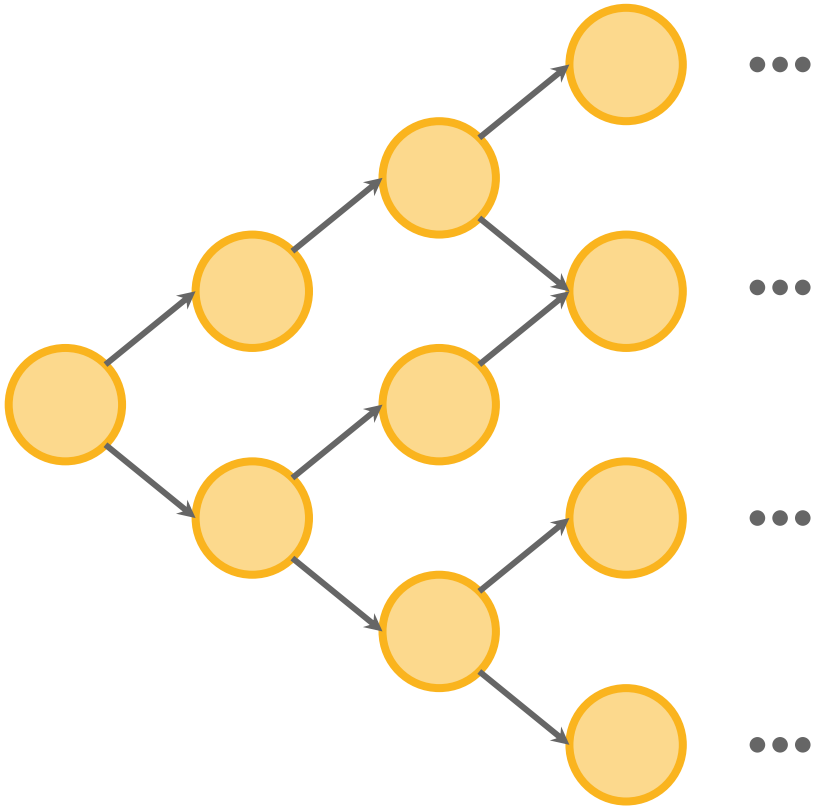


Value alignment problem: the concept

One is aligned with a value if their actions move them towards preferred states.

Actions get one to preferred states

Norms govern one's behaviour



Value alignment: alignment of norms with values

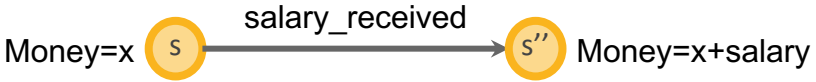
The transitions between states is governed by norms.

Value alignment: alignment of norms with values

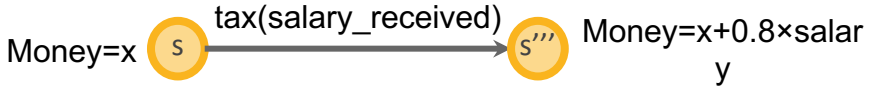
The transitions between states is governed by norms.

Norms change the world: states and transitions.

E.G.



a world with no tax

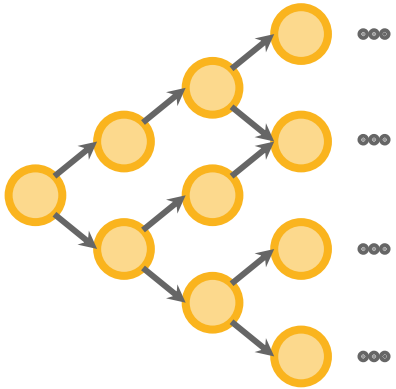


a world with 20% taxes

Value alignment: a definition

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

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

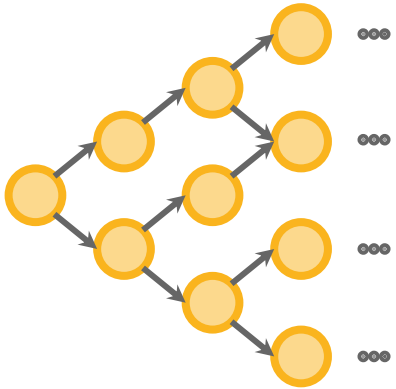


Value alignment: a definition

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.

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

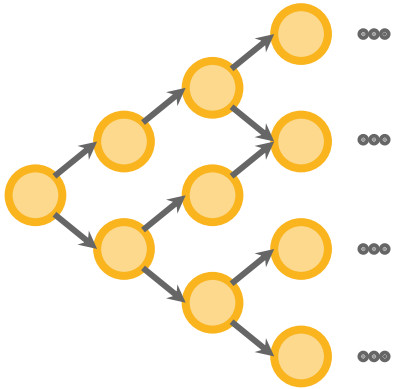


Value alignment: a definition

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.

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

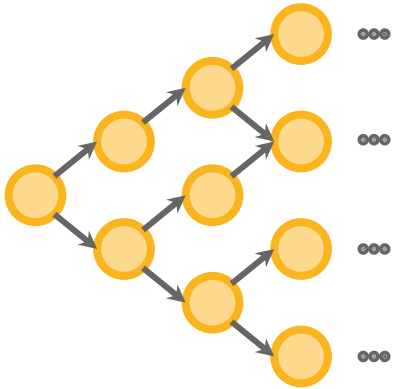


Value alignment: a definition

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

For large spaces, we can follow a Monte Carlo sampling approach, where *x* is the number of sampled paths, and *l* the path length:

$$\mathsf{Align}^{\alpha}_{n,v}(\mathcal{S},\mathcal{A},T) = \frac{\sum_{p \in paths'} \sum_{d \in [1,l]} \mathsf{Prf}^{\alpha}_{v}(p_I[d], p_F[d])}{x * l}$$



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

Prisoner's Dilemma

Value-based preferences.

- States with higher equality in accumulated gain are preferred:

1. $\mathrm{Prf}(s, s') = \dfrac{|x - y|}{\max\{x, y\}} - \dfrac{|x' - y'|}{\max\{x', y'\}}$

- States with higher equality in accumulated gain are preferred only if my personal gain is not lower:

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

- States with higher personal gain are preferred only if equality is not lower:

3. $\mathrm{Prf}(s, s') = \dfrac{x' - x}{2(\max\{x', x\})} - \dfrac{y' - y}{2(\max\{y', y\})}$

- States with higher personal gain are preferred, regardless of equality:

4. $\mathrm{Prf}(s, s') = \dfrac{x' - x}{\max\{x', x\}}$

Prisoner's Dilemma

Value-based preferences.

- States with higher equality in accumulated gain are preferred:

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

- States with higher equality in accumulated gain are preferred only if my personal gain is not lower:

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

- States with higher personal gain are preferred only if equality is not lower:

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

- States with higher personal gain are preferred, regardless of equality:

4. $\mathrm{Prf}(s,s') = \frac{x'-x}{\max\{x',x\}}$

Norms.

- **The no taxing - n_0:**
  No taxes are to be payed.

- **The incremental taxing - n_1:**
  No taxes to be paid when the gain is 0 or 3,
  3 to be paid as taxes when the gain is 6,
  and 5 to be paid as taxes when the gain is 9.

- **The fixed taxing - n_2:**
  1/3 of the gains of each game is to be paid as taxes.

Prisoner's Dilemma

Value-based preferences.

- States with higher equality in accumulated gain are preferred:

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

- States with higher equality in accumulated gain are preferred only if my personal gain is not lower:

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

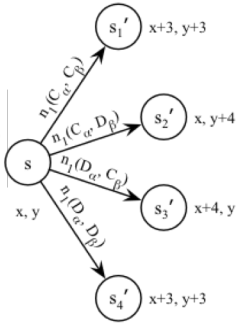
- States with higher personal gain are preferred only if equality is not lower:

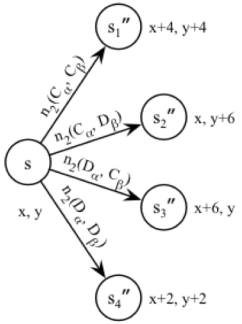
$$\text{3}\quad \text{Prf}(s,s') = \frac{x'-x}{2(\max\{x',x\})} - \frac{y'-y}{2(\max\{y',y\})}$$

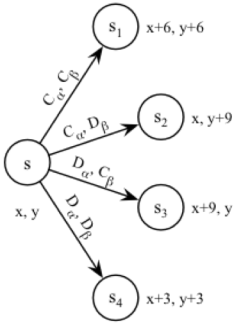
- States with higher personal gain are preferred, regardless of equality:

$$\text{4}\quad \text{Prf}(s,s') = \frac{x'-x}{\max\{x',x\}}$$

Norms.







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

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}	$n_0 \sim n_1 \succ n_2$
❶❷❸❹	{c,d}	{c,d}	$n_0 \sim n_1 \sim n_2$

Prisoner's Dilemma

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

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}	$n_0 \sim n_1 \succ n_2$
❶❷❸❹	{c,d}	{c,d}	$n_0 \sim n_1 \sim n_2$

Prisoner's Dilemma

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

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}	$n_1 \succ n_0 \sim n_2$
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}	$n_0 \sim n_1 \succ n_2$
❶❷❸❹	{c,d}	{c,d}	$n_0 \sim n_1 \sim n_2$

Prisoner's Dilemma

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

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}	$n_1 \succ n_0 \sim n_2$
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}	$n_0 \sim n_1 \succ n_2$
❶❷❸❹	{c,d}	{c,d}	$n_0 \sim n_1 \sim n_2$

Prisoner's Dilemma

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

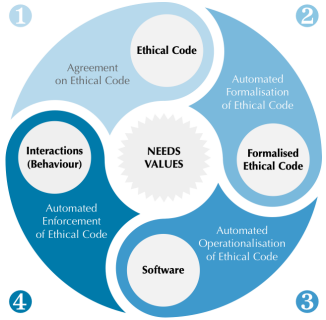
All norms (n_0 , n_1 , n_2) are equally aligned for moderate supporters of equality (❷,❸).

Except when *β* 's gains are higher (*β* always defecting).

	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}	$n_0 \sim n_1 \succ n_2$
❶❷❸❹	{c,d}	{c,d}	$n_0 \sim n_1 \sim n_2$

In conclusion

In conclusion...



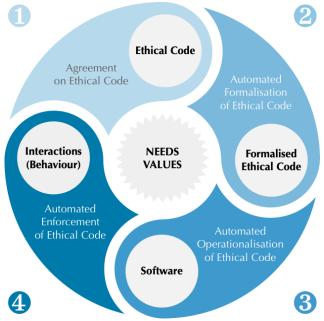
Motivated by some of the **ethical concerns**, I propose to:

(1) Develop a novel methodology and associated technology for the **design and development** of **responsible autonomy** that are based on people's **needs and values** and that **evolve** with people's evolving needs and values.

(2) **Give people control** over their technologies so they can decide amongst themselves on their needs and values, and how their technology should behave accordingly.

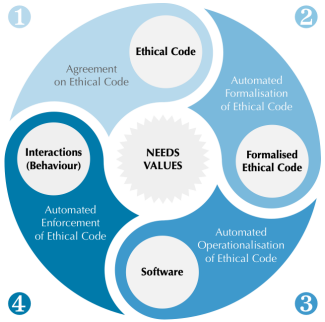
This methodology and technology aim at

- Empowering people to self-regulate their communities, interactions and objectives.
- Helping communities to satisfy Ostrom’s principles to guarantee sustainability.
- Supporting **explainabilty** and **transparency**.
- Providing tools for the analysis, coding and deployment of norms.



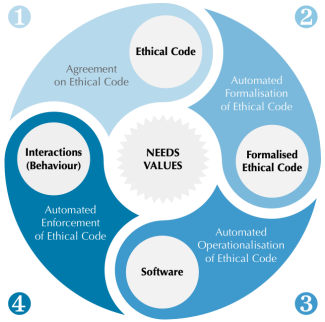
And generate plenty of open research questions

- When are two arguments similar?
- How to extract a normative position from text?
- How to deal with ethical conflict, i.e. conflicting norms?
- How to assess the impact of a normative change?
- How to learn norms from behaviour?
- How to synthesize code that implements norms?
- How to model incentives with norms?
- How to assess the sustainability of a normative system given a set of values shared by the humans?
- Is any set of norms acceptable?
- How to reconcile top-down and bottom-up generated norms?



And generate plenty of open research questions

- When are two arguments similar?
- How to extract a normative position from text?
- How to deal with ethical conflict, i.e. conflicting norms?
- How to assess the impact of a normative change?
- How to learn norms from behaviour?
- How to synthesize code that implements norms?
- How to model incentives with norms?
- How to assess the sustainability of a normative system given a set of values shared by the humans?
- Is any set of norms acceptable?
- How to reconcile top-down and bottom-up



A research program for the MAS community

Thank you

Carles Sierra
sierra@iiia.csic.es

WENET
INTERNET OF US