



Machines that now what matters: Engineering Awareness of Social Values

Carles Sierra

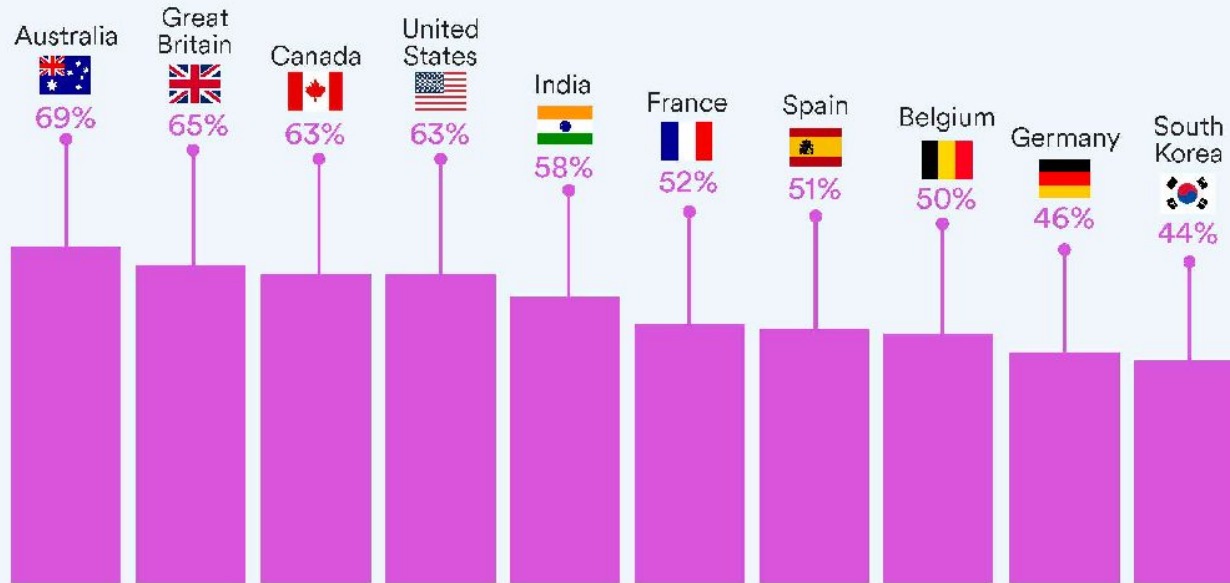
Artificial Intelligence Research Institute (IIIA-CSIC) Barcelona

COOPIS Conference, Marbella, 2025

AI at the ethical frontier

Global opinions: Where people say AI makes them nervous, 2023

Source: Ipsos, 2023 | Chart: 2024 AI Index report

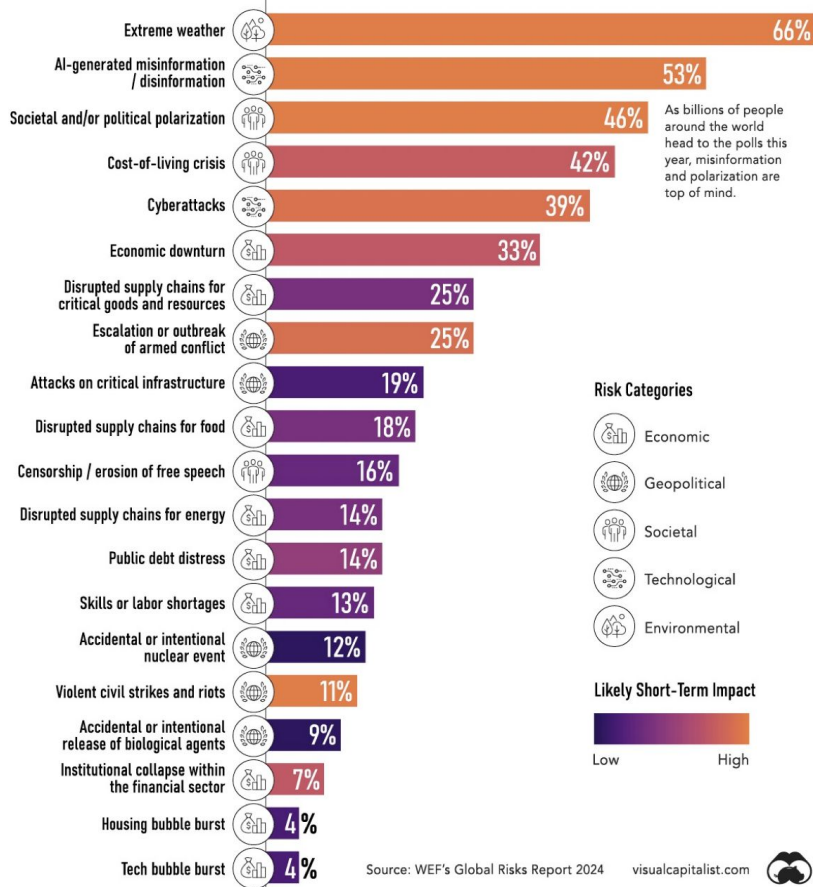


THE TOP GLOBAL RISKS IN 2024

The World Economic Forum surveyed 1,490 leaders on the top global risks in 2024 and their potential scale of impact.

Q

Please select up to five risks that you believe are most likely to present a material crisis on a global scale in 2024.



Source: WEF's Global Risks Report 2024

visualcapitalist.com



Support The Guardian

The Guardian

News | **Opinion** | **Sport** | **Culture** | **Lifestyle**

■ ■ ■

target variables
gather predictions
bored housewives Facebook likes
computer build 87 million
survey data human psychology
political affiliation results over-sampled incentive
profiling factors spreadsheet 120 data
data voting intention target Cambridge Analytica
real name target Information
favourite predictive power
predictive adverts score
algorithm accurately predict
training set
people massive matrix
open book consumer
electoral register data
more information
women
men
AI
data
voters
data
differences
Christopher Wylie
fill out a survey
ensemble model
harvest
possibilities
being able to predict
answer
data

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



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 EU regulation, the bloc said on Thursday.

We use cookies to improve our service for you. You can find more information in our data protection declaration.

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

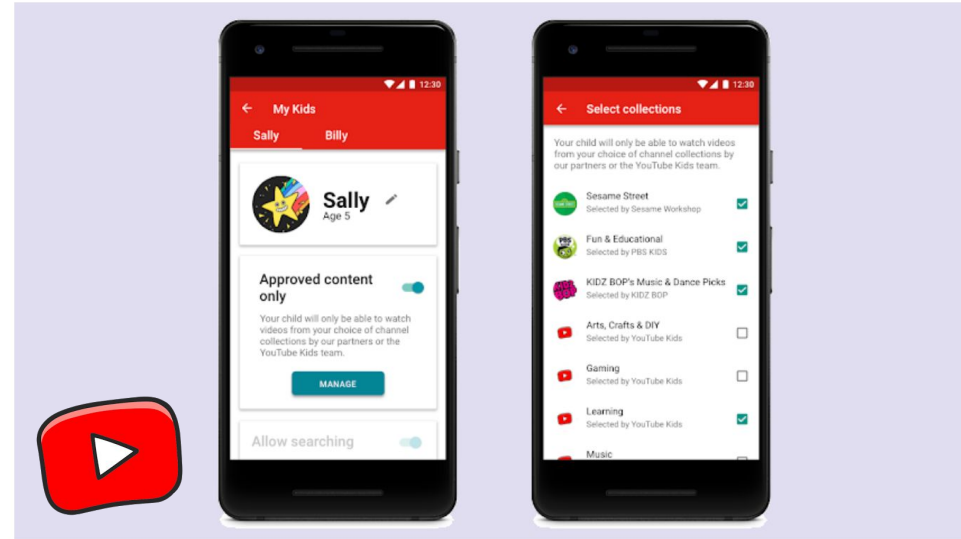
Ethical Concerns

Not JUST
privacy,
security, &
manipulation!

We are also
concerned about
basic features
and functionality.

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 to

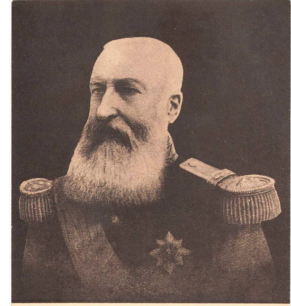
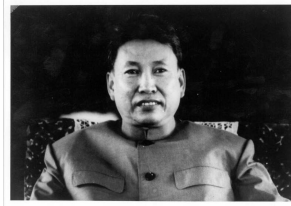
Can we (AI developers) put humans in control?

Should we?

Moral Humans

Psychopaths, an amoral mind.

- **Enjoying other's suffering/loses**
- **Insensibility to signs of physical pain in others**
- **Lack of fear and insensibility to punishment**
- **Instrumental, gratuitous violence**
- **Lack of remorse, guilt, shame**



Breakdown in the brain network subserving moral judgment in criminal psychopathy

Pujol J, Batalla I, Contreras-Rodriguez O, Harrison BJ, Pera V, Hernández-Ribas R, Real E, Soriano-Mas C, Deus J, López-Solà M, Pifarre J, Menchón JM and Cardoner N (2011), *SCAN*, 7, 8, 917-923.



18. Mr. Jones is working on a section of the rail track where two separate tracks converge. A runaway trolley is heading towards Mr. Jones's position. On the tracks travelling to the left there is a group of five railway workmen. On the tracks to the right there is a single workman. If Mr. Jones does nothing, the trolley will proceed to the left and kill the group of five men. If he changes the train's direction, the trolley will divert to the right killing only the one workman.

Voice prompt: *Would you switch on the dashboard?*

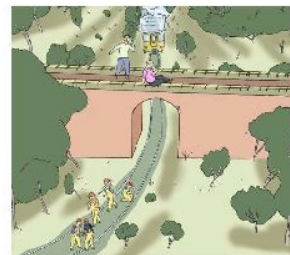
Psychopaths' Responses:

Yes: 77.3%; No: 22.7% (Omissions= 0)

Controls' Responses:

Yes: 90.9%; No: 9.1% (Omissions= 0)

Group comparison: $\chi^2 = 1.5$; $p = 0.216$



8. Mr. Jones sees a trolley can that is moving at high speed towards five workmen on the rail track. Mr. Jones is standing on a footbridge above the tracks. Next to him there is a very large and tall man. If Mr. Jones pushes the man off the bridge, he will die but his body will stop the trolley and the workmen will be saved. If he does not, all the workmen will die.

Voice prompt: *Would you push the man off the bridge?*

Psychopaths' Responses:

Yes: 60%; No: 40% (Omissions= 2)

Controls' Responses:

Yes: 36.4%; No: 63.6% (Omissions= 0)

Group comparison: $\chi^2 = 2.3$; $p = 0.126$



6. Dr. Jones has five patients, each of whom is close to dying from organ failures. She also has another patient who is mostly healthy. The only way that she can save the five others is to transplant this man's organs into their bodies but against his will if she does this, the healthy man will die, but the other five patients will live.

Voice prompt: *Would you transplant the man's organs?*

Psychopaths' Responses:

Yes: 57.1%; No: 42.9% (Omissions= 1)

Controls' Responses:

Yes: 13.6%; No: 86.4% (Omissions= 0)

Group comparison: $\chi^2 = 9.0$; $p = 0.003$



13. Mr. Jones's plane has crashed in the Himalayas. The only survivors are one other man, a young boy and himself. To live they must find their way to a small town on the other side of the mountain. They trek for three days in the extreme cold. The young boy falls and breaks his leg, critically reducing his chances of survival. The other man suggests to Mr. Jones to sacrifice the boy and eat his remains in order to survive. If Mr. Jones accepts the proposal they will have enough strength to make it to the small town. If he does not, the boy will eventually die and they will too.

Voice prompt: *Would you kill the young boy?*

Psychopaths' Responses:

Yes: 50%; No: 50% (Omissions= 2)

Controls' Responses:

Yes: 22.7%; No: 77.3% (Omissions= 0)

Group comparison: $\chi^2 = 3.4$; $p = 0.065$



11. Mr. Jones goes to the hospital to visit a sick friend. There he meets a young man who explains to Mr. Jones that his father has been admitted to the hospital and only has one more week to live. He explains that his father has a substantial life insurance policy that will expire at midnight and offers Mr. Jones \$12,000 to kill him. If Mr. Jones accepts the offer, he will have to kill the old man but he will receive the money. If he does not, the insurance will expire and neither of them will receive a cent.

Voice prompt: *Would you kill the old man?*

Psychopaths' Responses:

Yes: 77.7%; No: 22.3% (Omissions= 0)

Controls' Responses:

Yes: 0%; No: 100% (Omissions= 0)

Group comparison: $\chi^2 = 5.6$; $p = 0.018$



7. Mr. Jones is a young architect who is visiting one of his construction sites with his boss. His boss is a despicable man who makes everyone miserable, including Mr. Jones. If Mr. Jones pushes him off the building he will die and Mr. Jones will be interviewed by the police, but if he does not his boss will continue running other people's lives.

Voice prompt: *Would you push him off the building?*

Psychopaths' Responses:

Yes: 51.8%; No: 48.2% (Omissions= 0)

Controls' Responses:

Yes: 0%; No: 100% (Omissions= 0)

Group comparison: $\chi^2 = 8.3$; $p = 0.004$



Answers

- Psychopaths: 57,1% Yes

- Control Group: 13,6% Yes

Control group were nurses!!!!

Humans in control? Yes, but as communities.

- We as individuals take different decisions in front of similar situations
- We as individuals interpret social and moral values differently
- Therefore, we need a social contract with technology to enable Communities to rule.

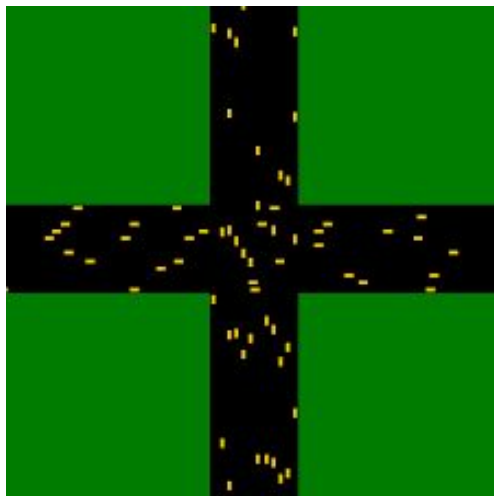
What rights do we *agree* to surrender, as a society or group, to technology in exchange for the protection of our remaining rights and maintenance of the social order?

Towards value engineering

It has to be US! We need to put people (pl.) in control, because AI must be social

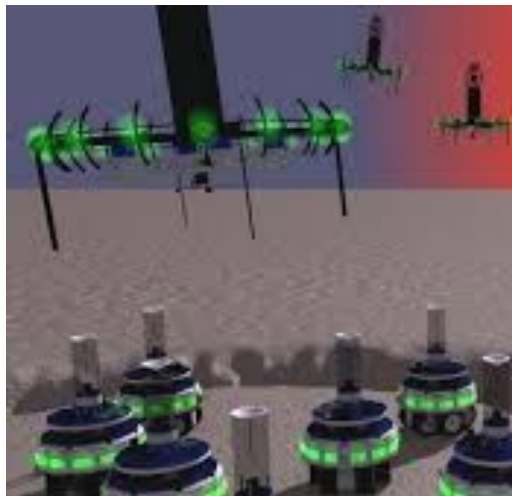
Billions of AI systems will interact among themselves and with humans. Our future society will be a colossal Multiagent System, a huge **sociotechnical community**.

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](#):

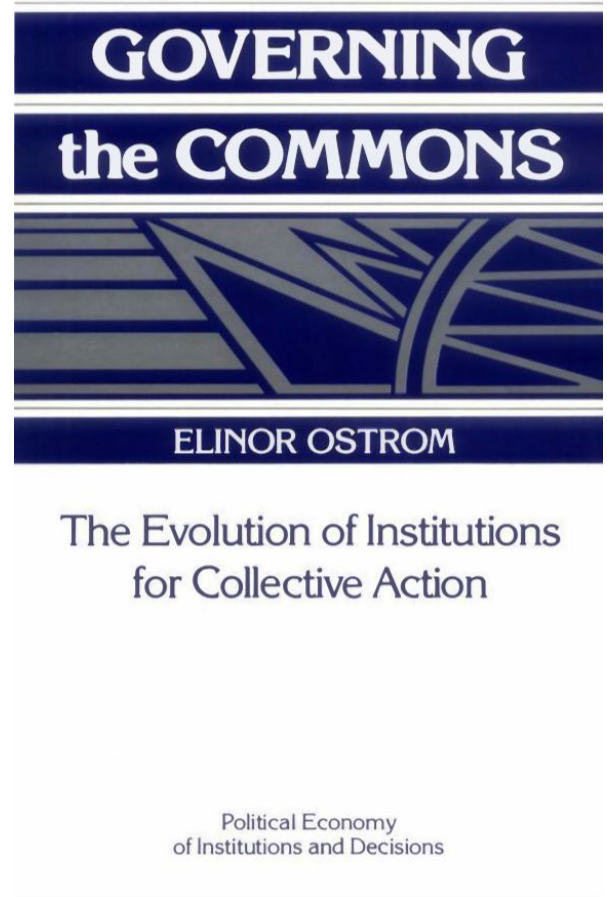
Multiuser museum interactives for shared cultural experiences: an agent-based approach. [AAMAS 2013](#): 917-924

But how to guarantee value respect when entities are autonomous?

- Values are social constructs. No universals; they are context dependent.
- No individual behaviour guarantee can be obtained when systems are **fully** autonomous.
- However, we can design **sociotechnical communities** so that unacceptable behaviour generates **repair actions** and **punishments**. (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.

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.

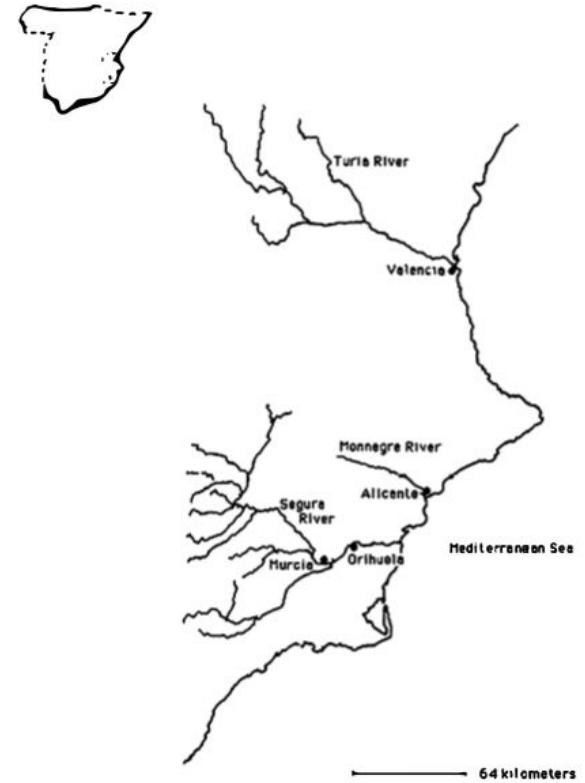


Figure 3.1. Location of Spanish *Invernas*.

Human communities are often successful

Ostrom's principles and the *Horta*

Boundaries: irrigation rights come with the land.

Appropriation 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.

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 equally divide the work.
- Workers will make equally sized piles of logs.
- A lottery determines which pile goes to which household.

Two illustrative examples in a digital world

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.

```
upload_photo(Someone, Photo)  $\implies$ 
   $\neg$  tag(SomeoneElse, Photo, TaggedPerson)
   $\wedge$  SomeoneElse  $\neq$  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):

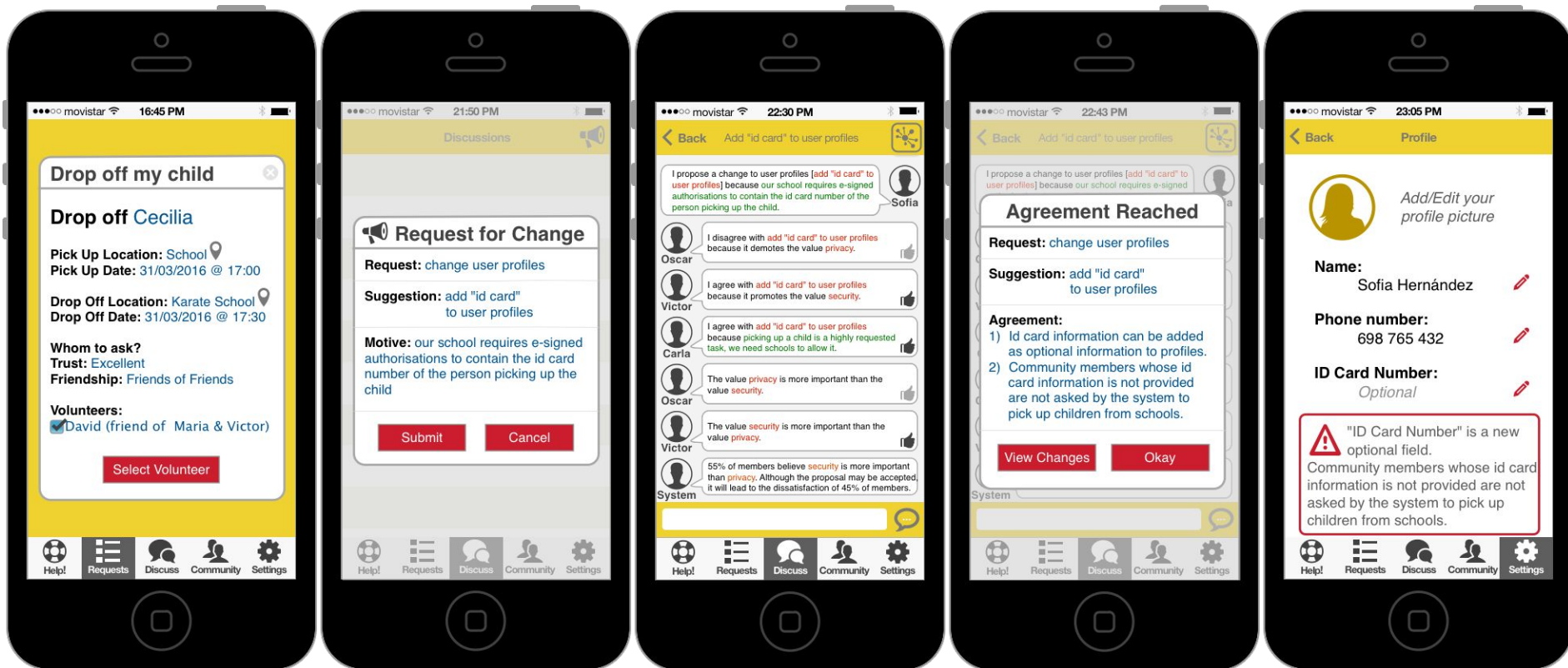
The photo cannot be tagged by anyone other than the owner.



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

Ok

Single mothers community in uHelp.



From Values to Norms

Context



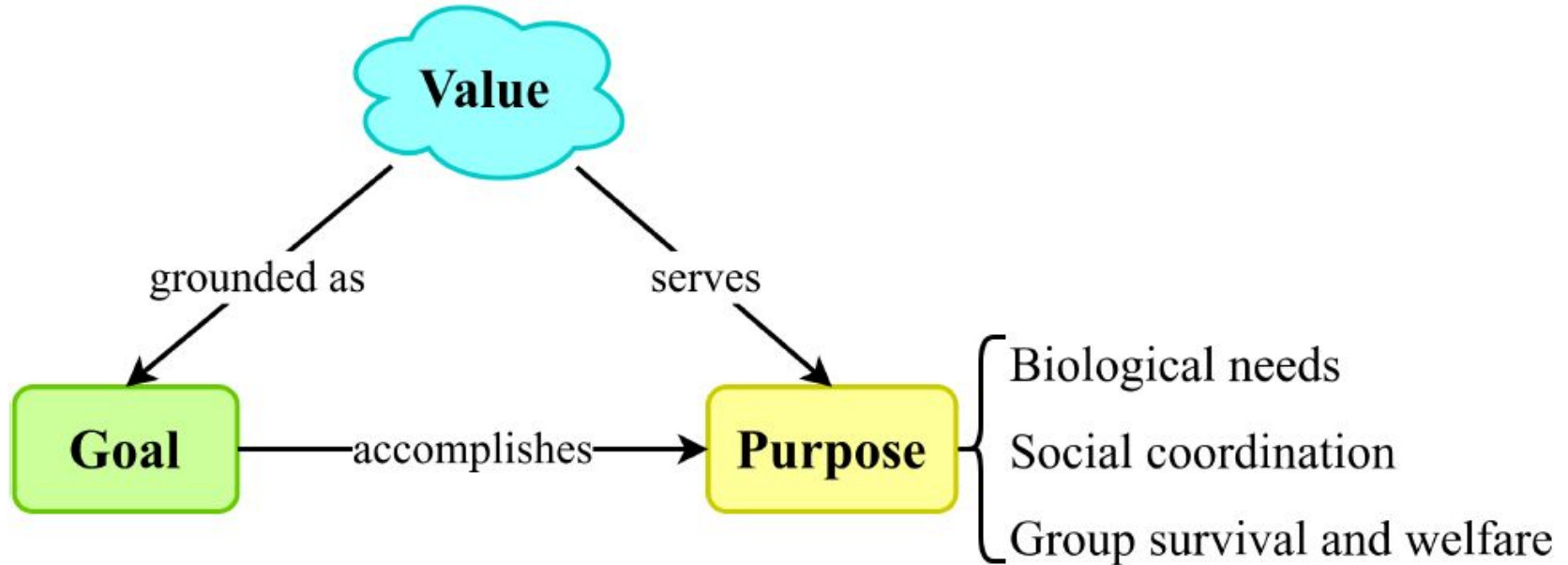
The future is a **massive multiagent system**:

- Humans
- Software agents

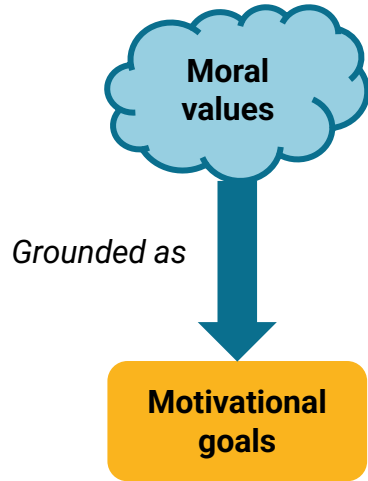
We (humans) are responsible for **embedding values** into autonomous agents.

→ Programmers need to turn **values into code**

Turning Values into Code. Schwartz's theory



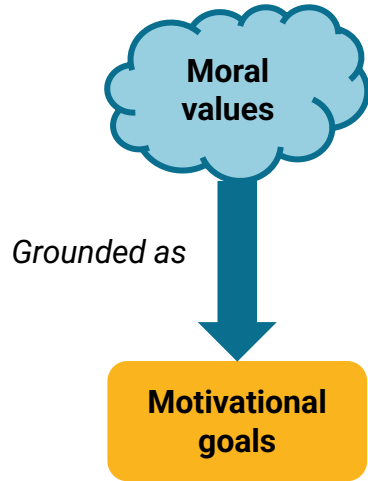
Turning Values into Code



Based on Schwartz's Theory of Basic Human Values, abstract values are **grounded into permanent goals** that agents pursue:

1. Proxies for values
2. How values manifest in the real/virtual world
3. State of affairs + history of actions
4. Permanent status
5. Value hierarchy → goal hierarchy

Turning Values into Code



Once values are turned into permanent goals, the autonomous agent can **evaluate the state of the world** as it relates to every value instilled in it.

The **human designer is in charge** of deciding how a value is grounded:

→ **control over the meaning of values**

Promoting Values with Norms

Technical norms encompass:

Do's and don'ts over the actions that an autonomous agent can take,
under what **circumstances**
with what **effects**

→ **Norms are the main value-promoting mechanism** ←

Promoting Values with Norms

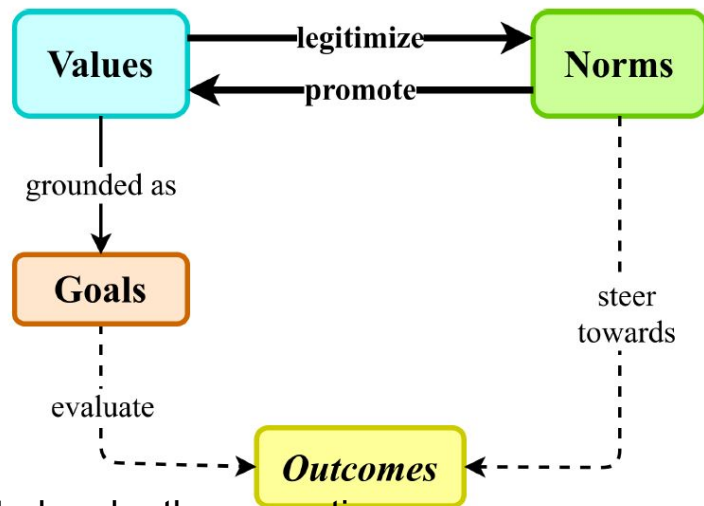
Norms can change the incentive structure and **steer the system** towards outcomes that are better viewed in terms of values (i.e. closer to the grounding goals).

When that is the case →

norms are aligned w.r.t. values

$$\text{Align}_{N,v} = \mathbb{E} \left[f_v \left(\text{out}(\mathcal{P}^N) \right) \right]$$

$\mathcal{P}^N$ is the random variable of the subset of paths restricted under the normative system N . f_v is the semantics function of value v .



Methodology

- 1) Define the variables that determine the state
- 2) Define the (parametric) norms that regulate the transitions
- 3) Define the set of values of interest
- 4) Apply an optimisation algorithm to find the parameters that maximise alignment
- 5) Analyse the shapley value of norms and value compatibility
- 6) Iterate from 2

Example

tax system, taxes get collected, grown and redistributed. Some agents cheat and try to evade.

3) Optimise

1) Values = {equality, fairness}

2) Normative system:

n_1 : how much should each agent pay

n_2 : how to redistribute

n_3 : how often should evaders be detected

n_4 : how much to fine evaders when detected

Value and semantics function	Optimal normative parameters	Optimal alignment $\text{Algn}_{N,v}^*$
Equality, eq. (5)	$collect = [20\%, 29\%, 26\%, 35\%, 27\%]$	0.95
	$redistribute = [20\%, 22\%, 19\%, 26\%, 13\%]$	
	$catch = 44\%$	
	$fine = 61\%$	
Fairness, eq. (7)	$collect = [1\%, 30\%, 37\%, 72\%, 66\%]$	0.93
	$redistribute = [2\%, 23\%, 42\%, 24\%, 9\%]$	
	$catch = 45\%$	
	$fine = 56\%$	
Aggregation, eq. (8)	$collect = [2\%, 79\%, 56\%, 65\%, 59\%]$	0.66
	$redistribute = [2\%, 28\%, 25\%, 35\%, 10\%]$	
	$catch = 31\%$	
	$fine = 77\%$	

Example

Interactions among individual norms

In normative system N , how much is norm n_i responsible for compliance with value v ?

Shapley values

For equality: collecting taxes

For fairness: dealing with cheaters

Value	Norm	Shapley value
Equality	n_1	0.50
	n_2	0.03
	n_3	0.07
	n_4	0.01
Fairness	n_1	0.19
	n_2	0.45
	n_3	0.46
	n_4	0.42
Aggregation	n_1	0.00
	n_2	0.27
	n_3	0.25
	n_4	0.31

Example

		v_j		
		Equality	Fairness	Aggregation
v_i	Equality	-	-0.28	-0.26
	Fairness	0.60	-	0.56
	Aggregation	0.71	0.88	-

Interactions among values

Can normative system N promote more than one value simultaneously?

Value compatibility

Promoting *fairness indirectly leads to equality*, but *not the other way around*.

Methodology

- 1) Define the variables that determine the state
- 2) Define the (parametric) norms that regulate the transitions
- 3) Define the set of values of interest
- 4) Apply an optimisation algorithm to find the parameters that maximise alignment
- 5) Analyse the shapley value of norms and value compatibility
- 6) Iterate from 2

In conclusion

Value engineering aims at

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

And generates plenty of open research questions

- When are two values 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

TAILOR



Carles Sierra
sierra@iia.csic.es