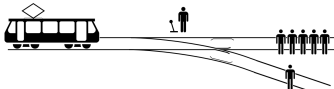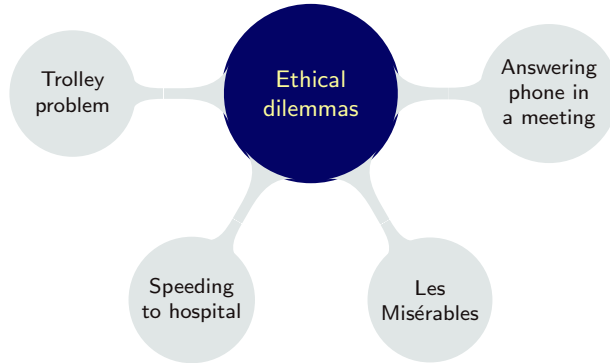# Socially Intelligent Agents to Support Ethical Decision Making

Nirav Ajmeri
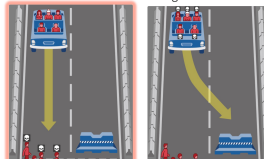
Intelligent Systems Lab
Department of Computer Science
University of Bristol
https://niravajmeri.github.io

May 2023

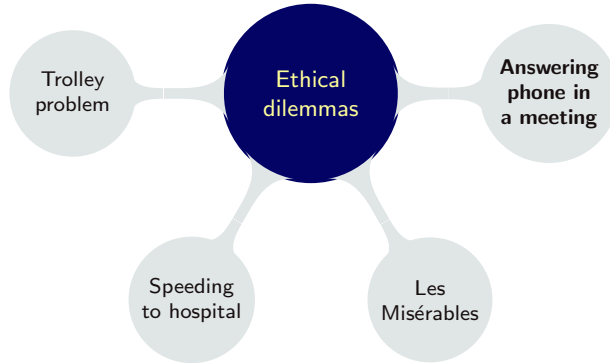# Ethical Dilemmas: No (Obviously) Good Choices

What should the self-driving car do?

Ethical Dilemmas: No (Obviously) Good Choices

Ethical dilemmas arise not only in hypothetical or extreme scenarios but also in mundane scenarios
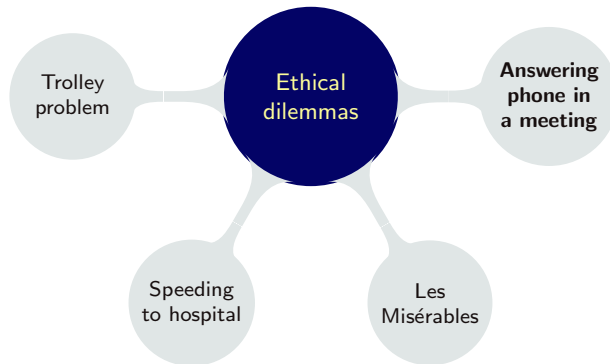
Ethical Dilemmas: No (Obviously) Good Choices

Ethical dilemmas arise not only in hypothetical or extreme scenarios but also in mundane scenarios

Ethics is inherently a multiagent concern

Privacy as Social Expectation
Example: Phone Ringer

US Senator's phone rings during important meeting - can you guess his embarrassing ringtone?

11:56, 17 APR 2015   BY KARA O'NEILL

It's bad enough when your phone goes off in a silent meeting room, but it's even worse when your ringtone is as embarrassing as this one

A US Senator was left red-faced after his phone went off during a finance meeting - but it was his choice of ringtone that really raised some eyebrows.

https://www.youtube.com/watch?v=r0tZU2_X1-Y

- Intrusion
- Disapprobation
- Disclosure

[Westin, 1967; Solove, 2006]

Source: https://twitter.com/TheSimpsons/status/441000198995582976

Tradeoffs: Values of Power, Pleasure, and Benevolence

Preliminary Concepts

Sociotechnical system is a cyberphysical system where multiple stakeholders (humans, organizations, and agents) interact

Social norm governs the interactions between two stakeholders

- Commitment: *Meeting attendee are committed to keep their phones on silent*
- Prohibition: *Students are prohibited by the university to use phones in examination halls*

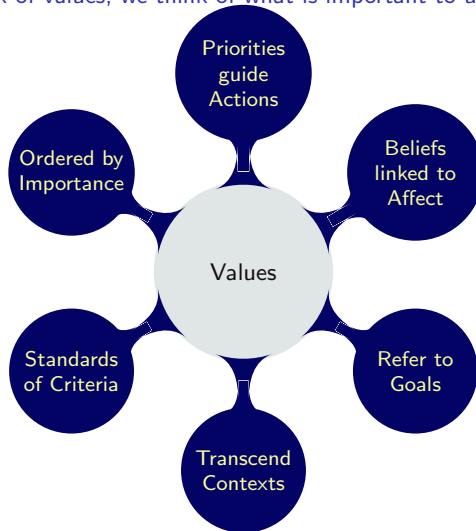Deviation is a perceived violation of a norm

Social context is the circumstance under which the interaction takes place

Values are guiding principles of humans

- Ethics: subsumed in the theory of values
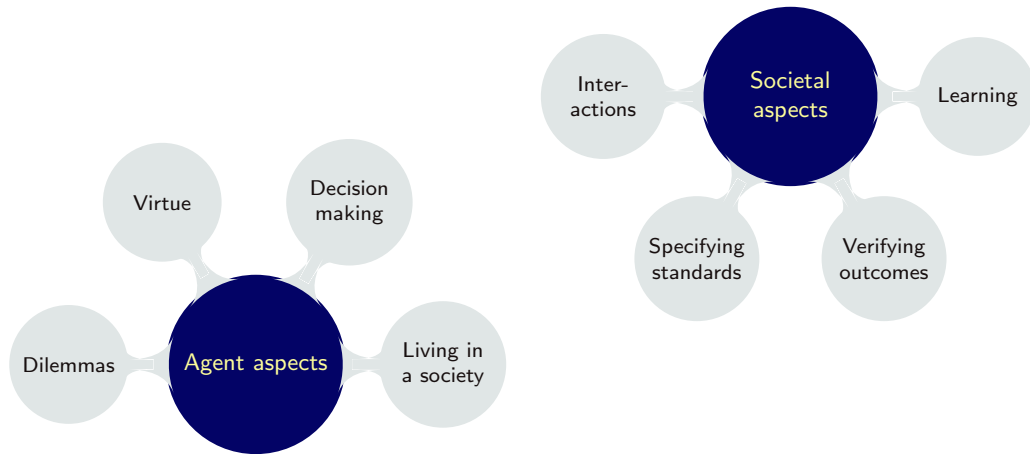- Privacy: a value with an ethical import

The Nature and Features of All Values

[Schwartz, 2012]: When we think of values, we think of what is important to us in life
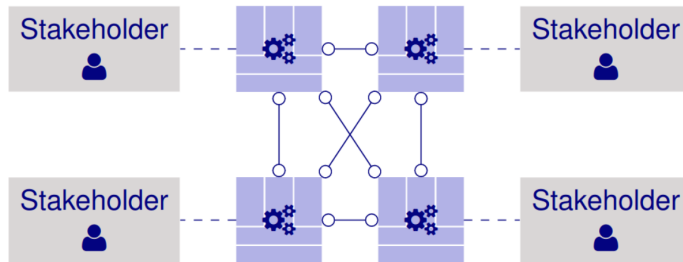
Ethics in Sociotechnical Systems

Ethics is an inherently multiagent concern, yet current approaches focus on single agents

Ethics in Society with SIPAs
SIPA: Socially intelligent (personal) agent

- A multiagent system is a microsociety
- Each agent reflects the autonomy of its (primary) stakeholder

Socially Intelligent Personal Agent (SIPA)

A SIPA adapts to social context and supports meeting social expectations

- Ethical: Seeks to balance needs of
  - Primary user (also a stakeholder), who directly interacts with the agent
  - Other stakeholders, who are affected by the agent's actions

Challenges

- Identifying values of interest to an agent or a MAS
- Incorporating values in an agent model
- Understanding values in context and communicate values
- Reasoning about values to revise norms

A SIPA: Schematically
What must a SIPA represent and reason about to participate ethically in a multiagent system?
A SIPA's decision making takes into account its stakeholders, primary and secondary

[AAMAS 2020] N. Ajmeri, H. Guo, P. K. Murukannaiah, and M. P. Singh. Elessar: Ethics in Norm-Aware Agents. In Proc. AAMAS, 1–9.
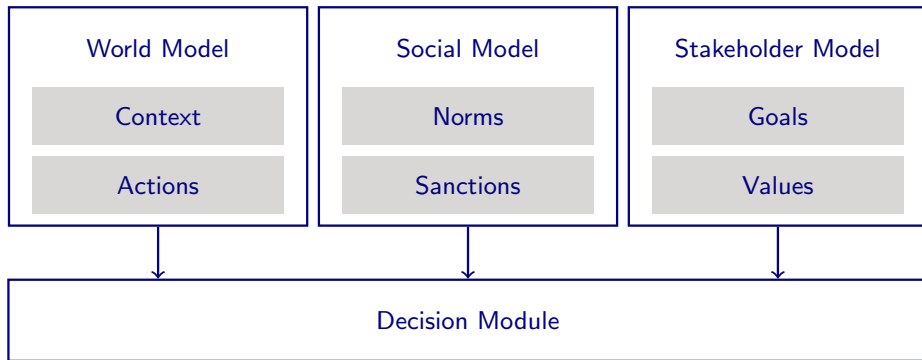
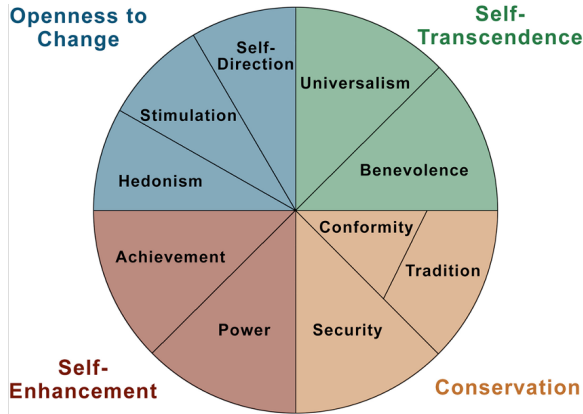# What Values should an Agent Elicit, Learn, or Align with?
[Schwartz, 2012]

Identifying Values of Interest an Agent or an MAS

Axies employs NLP for data-driven identification of values

Posted by u/jamesSkyder 28 days ago

13 **Demonstrators Rally in London to Protest Against COVID-19 Lockdown Measures**

......

owencroft 2 points · 28 days ago

I do wonder what they're protesting about. Like going on about freedom, what do they mean?

Lord_Bingham -2 points · 28 days ago

Well done them! Good to see people standing up for freedom.

ANormalPersonOnline 7 points · 28 days ago

Yeah! Freeeedom! Why do my lungs hurt?

...... Source: https://www.reddit.com/r/CoronavirusUK/comments/iisk44/demonstrators_rally_in_london_to_protest_against/

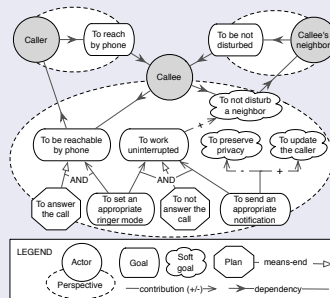# Incorporating Values in an Agent Model

Agent models provide **technical abstractions** to represent values

Example abstractions

- **Actor**: A social, physical, or software agent
- **Goal**: A strategic interest of an actor
- **Plan**: An abstraction of action
- **Belief**: An actor's representation of the world
- **Dependency**: A relationship between actors

An actor model of an Intelligent Ringer

Incorporating Values in an Agent Model
Agent models can be refined and mapped to agent capabilities

- To be reachable:
  Welfare of others $\Uparrow$
- To work uninterrupted:
  Ambition $\Uparrow$
- Welfare of others $\succ$ Ambition?

Xipho can yield a specification of value preferences grounded in contexts, e.g.,

> Relationship $= ?R_1 \wedge$ Neighbor's context $=$
> $?N_1 \rightarrow$ Welfare of others $\succ$ Ambition

A contextual model of Intelligent Ringer
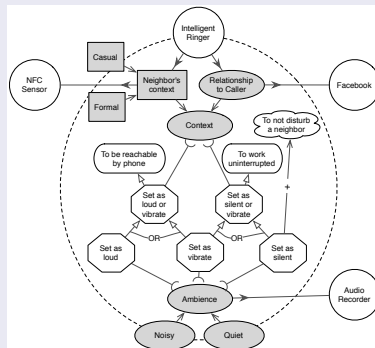
From Personal Values to Social Norms

Consider an example of values in a location sharing app

Prefers privacy

Prefers pleasure and recognition

Prefers Frank's safety

Prohibited by to share location

Committed to share location
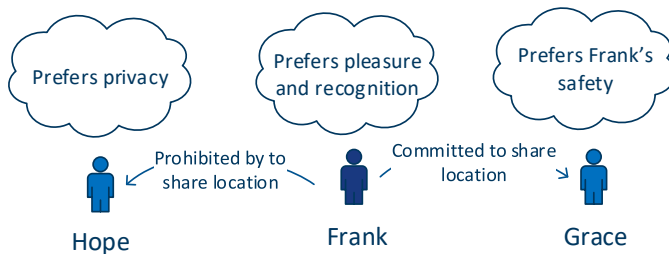
Hope

Frank

Grace

Frank's dilemma: Which sharing policy to select?

Share with all: Pleasure for Frank ⇑

Share only with Grace: Safety for Grace ⇑

Share with no one: Privacy for Hope ⇑

Choosing an Ethical Action using Values and Norms

How can SIPAs aggregate value preferences of their stakeholders to select an ethical action?
A SIPA's secondary stakeholders can change with the context
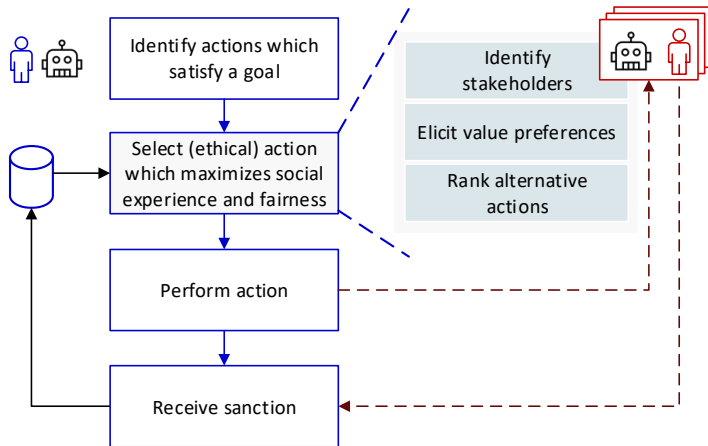
Choosing an Ethical Action using Values and Norms

SIPAs adapt a multicriteria decision making method (VIKOR) to select ethically appropriate action—balancing *utilitarianism* and *egalitarianism*
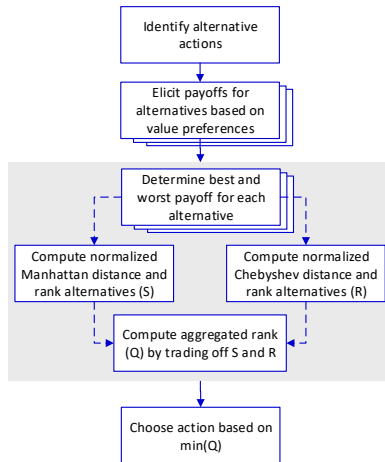
Restaurant Example: Where to Go?

Contrasting various ethical principles

	Alex	Dan	Jess
Pancake restaurant	10	10	2
Pasta restaurant	7	7	7
Pizza restaurant	5	5	10

Taxonomy of Ethical Principles
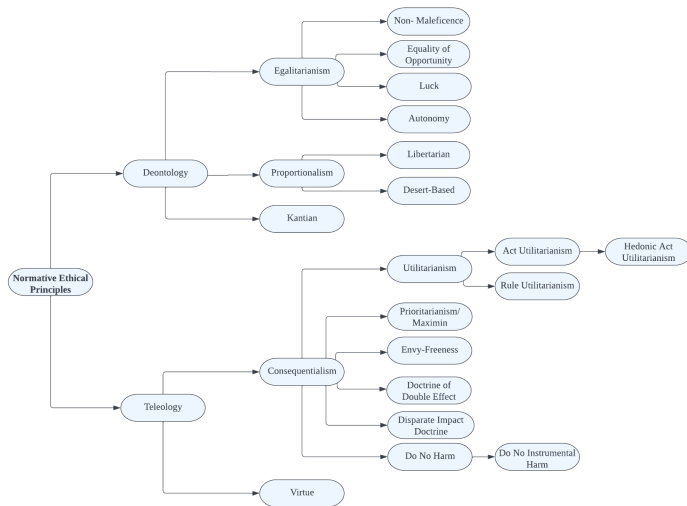
[arXiv 2022] J. Woodgate and N. Ajmeri. Principles for Macro Ethics of Sociotechnical Systems: Taxonomy and Future Directions. arXiv preprint arXiv:2208.12616

VIKOR Calculations

Policy Alternatives	Frank's Values				Hope's Values				S_y	R_y	Q_y
	Ple	Pri	Rec	Saf	Ple	Pri	Rec	Saf			
y_1 All	10	5	10	5	5	0	5	5	3.5	3	0.75
y_2 Common	5	5	5	10	5	0	5	5	0.4	3	1
y_3 Grace	0	5	0	0	5	15	5	5	**0.3**	**1**	**0**
w_x	1	1	1	1	1	3	1	1			
f_x^*	1	0	1	1	0	1	0	0			
f_x^-	0	0	0	0	0	0	0	0			

$$k = 0.5, \ w_{Hope-privacy} = 3$$

Explaining an Action using Values and Norms

Deviating SIPAs explain their deviations by sharing elements of their contexts
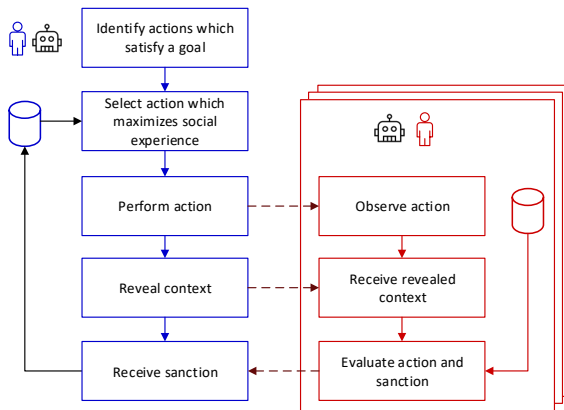
[IJCAI 2022] R. Agrawal, N. Ajmeri, and M. P. Singh. Socially Intelligent Genetic Agents for the Emergence of Explicit Norms. In Proc. IJCAI, 10–16.
[IJCAI 2018] N. Ajmeri, H. Guo, P. K. Murukannaiah, and M. P. Singh. Robust Norm Emergence by Revealing and Reasoning about Context: Socially Intelligent Agents for Enhancing Privacy. In Proc. IJCAI, 28–34.

Generating Explanations

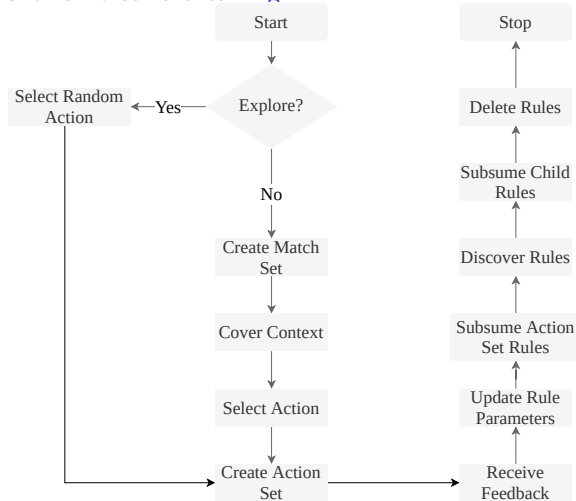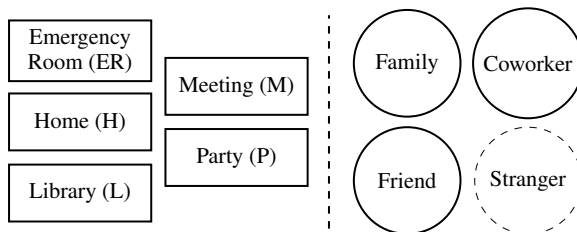SIPAs use genetic algorithm and reinforcement learning

[IJCAI 2022] R. Agrawal, N. Ajmeri, and M. P. Singh. Socially Intelligent Genetic Agents for the Emergence of Explicit Norms. In Proc. IJCAI, 10–16.

Evaluation: The Ringer Environment
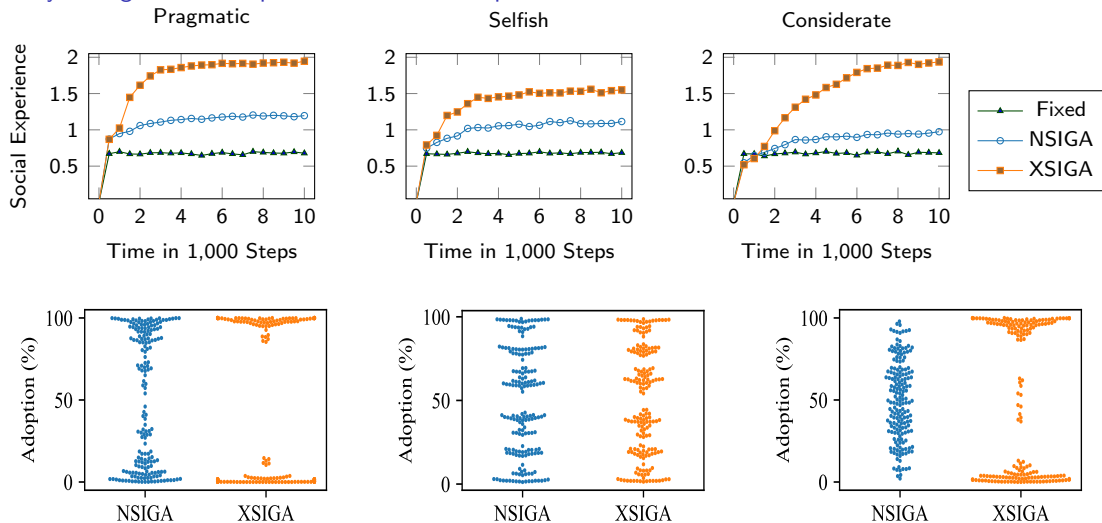

Agent societies:

- Pragmatic
- Considerate
- Selfish
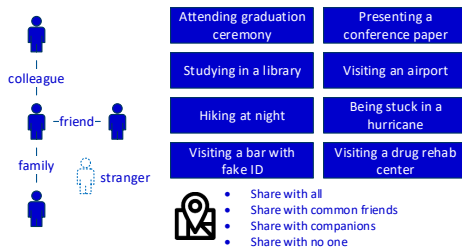
Learning strategies:

- Fixed
- Sanctioning
- Sharing Context (SIPA) or Explanations (XSIGA)

Results: Social Experience and Norm Adoption

SIPAs yield higher social experience and norm adoption than baselines

Evaluation: The Context-Sharing Environment

Attending graduation ceremony	Presenting a conference paper
Studying in a library	Visiting an airport
Hiking at night	Being stuck in a hurricane
Visiting a bar with fake ID	Visiting a drug rehab center

colleague

–friend–

family

stranger

- Share with all
- Share with common friends
- Share with companions
- Share with no one

Simulated societies:

- Mixed
- Cautious
- Conscientious
- Casual

Privacy attitude:
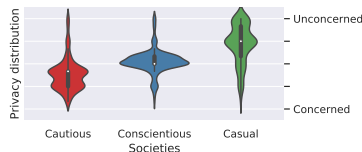
Decision-making strategies:
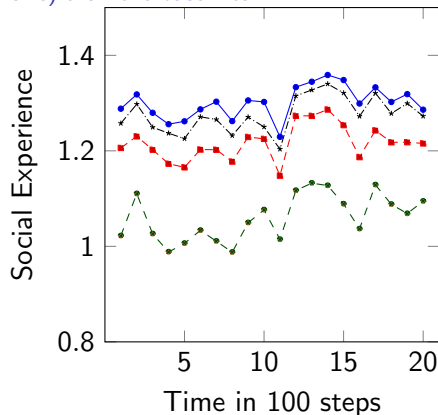
$S_{Elessar}$: Policy based on VIKOR

$S_{primary}$: Primary user's preference

$S_{conservative}$: Least privacy-violating

$S_{majority}$: Most common

Experience: Experiment with Mixed Privacy Attitudes

Result: Elessar SIPAs (which reason about value preferences) yield higher social experience ($p < 0.01$; Glass' $\Delta > 0.8$ indicating large effect size) than the baselines

Fairness: Experiment with Mixed Privacy Attitudes

Result: Elessar SIPAs (which reason about value preferences) give significantly better ($p < 0.01$) fairness with large effect size (Glass' $\Delta > 0.8$) than the baseline methods
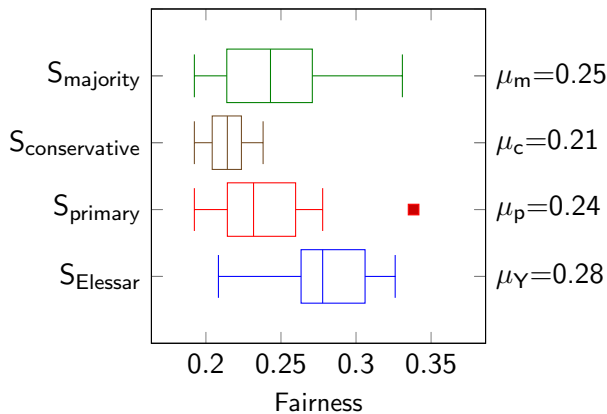
Summary

- Ethics inherently involves looking beyond one's self interests
- Ethical considerations apply in mundane settings—anywhere agents of multiple stakeholders interact
- Socially intelligent agents could help stakeholders navigate social norms of the society and support selecting ethically-appropriate actions

Ongoing Work
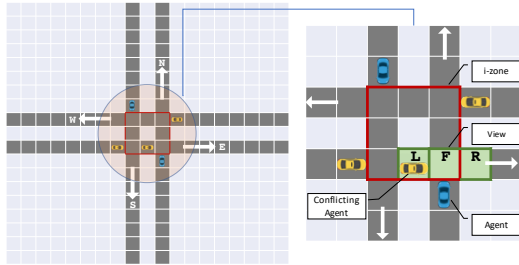
(a) Prosociality

(b) Ethical Principles

(c) Emotions

Opportunities and Directions

- How can we support decision making by an agent that takes into account the value preferences of the principals as well as the STS?
  - How to aggregate values preferences considering the current social context?
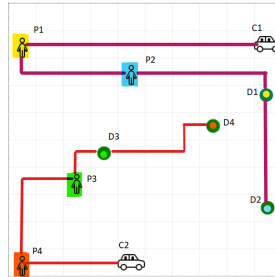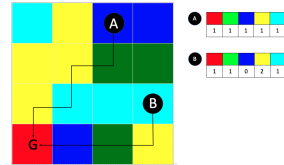 - How to incorporate ideas on emotions, guilt, consent, and prosociality?

- How can an agent elicit its users value preferences unintrusively?

- How can an agent explain its decisions to its users and to other agents and yet *preserve privacy*?

- Realistic data for realistic simulation?

[COINE 2023] D.E. Collins, C.J. Houghton, and N. Ajmeri. Social Value Orientation and Integral Emotions in Multi-Agent Systems Coordination. In Proc. COINE@AAMAS, 1–21.
[COINE 2022] S. Tzeng, N. Ajmeri, M.P. Singh. Fleur: Social Values Orientation for Robust Norm Emergence. In Proc. COINE@AAMAS, 185–200.
[arXiv 2022] A.O. Davies and N. Ajmeri. Realistic Synthetic Social Networks with Graph Neural Networks. arXiv preprint arXiv:2212.07843.

Thank You

Contact at nirav.ajmeri@bristol.ac.uk
https://niravajmeri.github.io
https://sites.google.com/view/ai-ethics/home
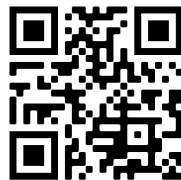
Appendix

Arnor: A Method to Model Social Intelligence

How can we model social intelligence in a SIPA to help it deliver a satisfactory experience to its stakeholders?

Goal modeling: identifying a SIPA's stakeholders, their goals, and plans

Context modeling: identifying the social contexts in which a SIPA's stakeholders interact

- Context helps in deciding which goals to bring about or plans to execute

Social expectation modeling: identifying norms and sanctions that govern stakeholders' goals and plans

Social experience modeling: identifying a SIPA's actions that improve social experience, i.e., choosing plans, goals, and norms

Evaluation: Developer Study

Participants: 30 developers

Mechanics: One factor; two alternatives

- Two groups (Arnor and Xipho, a prior method) balanced on skills developed RINGER SIPAs in six weeks
- Model, Implement, Test

Metrics:

- Coverage and correctness
- Time and difficulty to develop

Study Unit: RINGER SIPAs
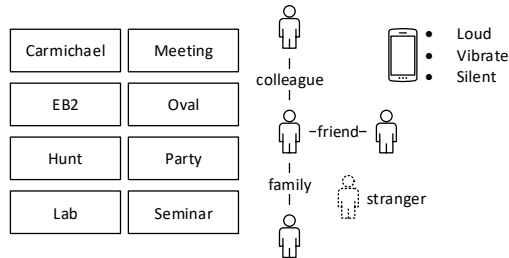
Developer Study

Result: Developers who follow Arnor spend less time and less effort to develop a SIPA, than those who follow Xipho (a previous approach)

Development Time

Time in hours

$\mu_A < \mu_X$, $p = 0.39$

Development Effort

Effort in (perceived) difficulty x hours

$\mu_A < \mu_X$, $p = 0.31$
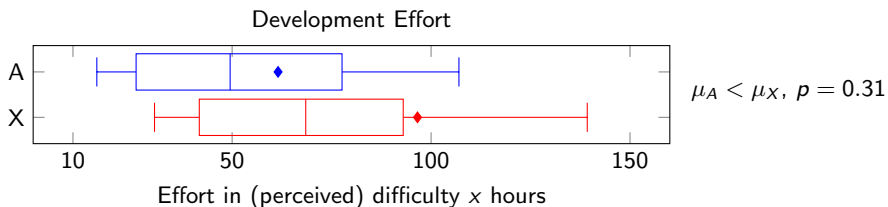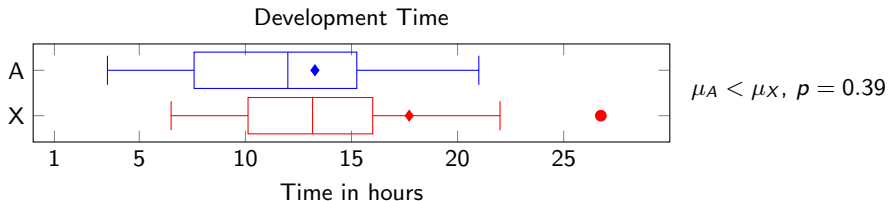
Evaluation: User Study (Simulations)

Developed RINGER SIPAs simulated in varying adaptation scenarios:

- Fixed norms
- Changing norms
- Changing context
- Changing sanction

Metrics:

- Adaptability coverage and correctness
- Norm compliance
- Proportion of positive sanctions

Simulation Experiments

Result: SIPAs developed using Arnor yield lower sanction proportions than SIPAs developed using Xipho (a previous approach)
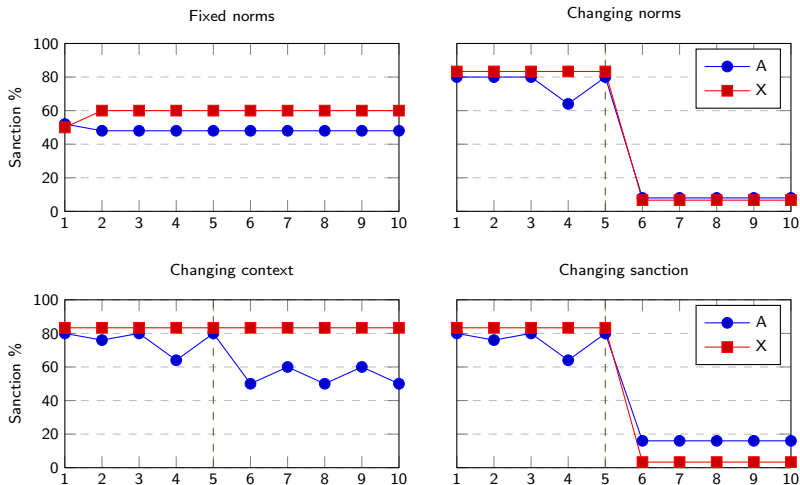
Evaluation: Social Simulations

Metric:

Social cohesion measures the proportion of agents that perceive actions as norm compliant. Higher the social cohesion, lower is the number of negative sanctions

Social experience measures the goal satisfaction delivered by an agent (computed by aggregating payoffs for all stakeholders)

Experiments on Pragmatic Agents (Varying Network Types)
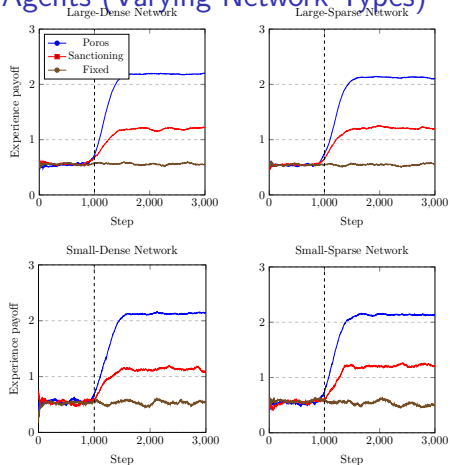
Social cohesion and social experience offered by Poros agents are significantly better than those offered by Fixed and Sanctioning agents

Experiments on Considerate and Selfish Agents
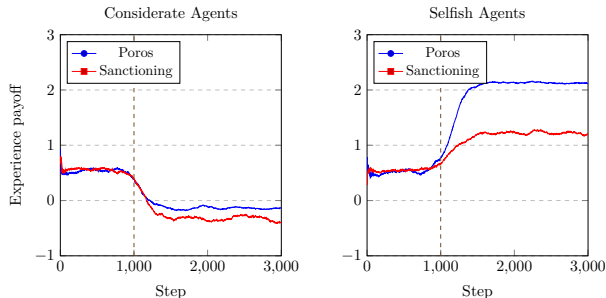
- The average social experience drops for considerate Sanctioning and Poros agents after they have gained enough confidence
- Plots for selfish agents are similar to those in the experiment with pragmatic agents, but with slightly lower stabilized values

VIKOR Summary

1. Determine the best and worst numeric payoffs, f_x^* and f_x^- for each value preference x over the alternative actions y to bring about a goal. That is, $f_x^* = \max_y f_{xy}$, $f_x^- = \min_y f_{xy}$.

2. For each alternative action y, compute the weighted and normalized Manhattan distance [Opricovic and Tzeng, 2004]:
$S_y = \sum_{x=1}^n w_x(f_x^* - f_{xy})/(f_x^* - f_x^-)$, where w_x is the weight for value preference x, which is subject to a stakeholder context and preferences over values. In particular, $S_y = 0$ when $f_x^* = f_x^-$.

3. Compute the weighted and normalized Chebyshev distance [Krause, 1973]:
$R_y = \max_x[w_x(f_x^* - f_{xy})/(f_x^* - f_x^-)]$, where w_x is the weight for value preference x.

4. Compute $Q_y = k(S_y - S^*)/(S^- - S^*) + (1 - k)(R_y - R^*)/(R^- - R^*)$, where
$S^* = \min_y S_y, S^- = \max_y S_y, R^* = \min_y R_y, R^- = \max_y R_y$, and k is a weight of the strategy to maximum group or individual experience. We set $k = 0.5$ to select a consensus policy.

5. Rank alternative actions, sorting by the values S, R, and Q, in increasing order. The results are three ranked lists of actions.

6. Choose the alternative based on min Q as the compromise solution if it is better than the second best alternative by a certain threshold or also the best ranked as per S and R.

VIKOR Calculations

Policy Alternatives	Frank's Values				Hope's Values				S_y	R_y	Q_y
	Ple	Pri	Rec	Saf	Ple	Pri	Rec	Saf			
y_1 All	10	5	10	5	5	0	5	5	3.5	3	0.75
y_2 Common	5	5	5	10	5	0	5	5	0.4	3	1
y_3 Andrew	0	5	0	0	5	15	5	5	**0.3**	**1**	**0**
w_x	1	1	1	1	1	3	1	1			
f_x^*	1	0	1	1	0	1	0	0			
f_x^-	0	0	0	0	0	0	0	0			

$$k = 0.5, \; w_{Hope-privacy} = 3$$

Simulated Places in the Simulation with Attributes Safe and Sensitive

Place	Safe	Sensitive
Attending graduation ceremony	–	No
Presenting a conference paper	–	No
Studying in library	Yes	–
Visiting airport	Yes	–
Hiking at night	No	–
Being stuck in a hurricane	No	–
Visiting a bar with fake ID	–	Yes
Visiting a drug rehab center	–	Yes

Example Numeric Utility Matrix for a Stakeholder

Place	Companion	Policy	Value			
			Pleasure	Privacy	Recognition	Security
Graduation	Family	All	1	0	1	0
Conference	Co-workers	None	0	1	0	0
Library	Friends	All	1	0	0	0
Airport	Friends	Common	0	1	0	0
Hiking	Alone	All	1	0	0	1
Hurricane	Family	All	1	0	0	1
Bar	Alone	None	0	2	0	0
Rehab	Friends	None	0	2	0	0

Evaluation: Crowdsourcing Study

Participants: 58 students enrolled in a mixed graduate and undergraduate-level computer science course

Privacy attitude survey: Level of comfort in sharing personal information [Schnorff et al., 2014]

Casual   Conscientious   Cautious
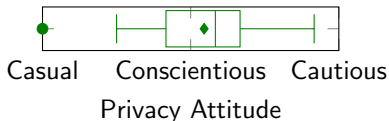
Privacy Attitude

Context sharing surveys: Select context sharing policy
- Phase 1. Based on context, including place and social relationship
- Phase 2. Based on context and values (pleasure, privacy, recognition, safety)

Metrics in Society with Mixed Privacy Attitudes

Strategy	Social	Best	Worst	Fairness	p
$S_{Elessar}$	**1.31**	3.07	**−0.57**	**0.28**	–
$S_{primary}$	1.23	3.01	−1.14	0.25	<0.01
$S_{conservative}$	1.07	3.07	−1.55	0.22	<0.01
$S_{majority}$	1.28	**3.08**	−1.15	0.24	<0.01

Metrics in Society with Majority Privacy Attitudes

Attitude / Strategy	Cautious				Conscientious				Casual			
	Social	Best	Worst	Fairness	Social	Best	Worst	Fairness	Social	Best	Worst	Fairness
$S_{Elessar}$	**1.25**	2.90	**−0.70**	**0.28**	**1.30**	2.93	**−0.46**	**0.30**	1.38	3.12	**−0.67**	**0.27**
$S_{primary}$	1.15	2.86	−1.07	0.26	1.22	2.91	−1.21	0.25	1.33	3.13	−1.03	0.24
$S_{conservative}$	0.93	2.89	−1.79	0.22	1.09	2.93	−1.42	0.23	1.23	3.13	−1.38	0.23
$S_{majority}$	1.20	**2.92**	−1.27	0.24	1.28	**2.94**	−0.86	0.27	**1.39**	3.13	−0.92	0.25

Location Sharing Survey: Policy Selection

Companion	Check-in Policy			
	Share with all	Common friends	Companions	No one
Alone	○	○	○	○
Colleague	○	○	○	○
Friend	○	○	○	○
Family member	○	○	○	○
Crowd	○	○	○	○