Prosociality and Fairness in Intelligent Agents

Nirav Ajmeri

School of Computer Science
University of Bristol
https://niravajmeri.github.io

October 2023

What is Ethics?

The field of ethics involves systematizing, defending, and recommending concepts of right and wrong behavior
[Fieser, The Internet Encyclopedia of Philosophy]

Classical ethics: Founded on economics and politics

The formation of the individual character (ethos) is intrinsically related to the others, as well as to the tasks of administration of work within the family (oikos), which eventually, expands into the framework of the public space (poleis)
[Ethically Aligned Design, IEEE]

Ethos
Oikos
Poleis

Prosociality

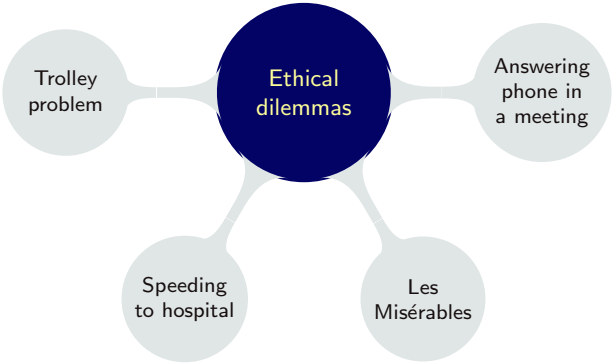
Prociality is the extent to which a person's context-driven behaviour relates

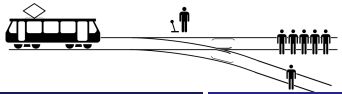
- to an individual beneficiary (e.g., a stranger or someone selected based on reciprocity) or
- a positive societal outcome (e.g., aggregate welfare or helping those worst off)

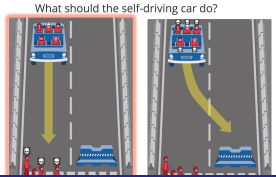
[Adler, 2019; van Lange, 1999]

https://verywellmind.com

Ethical Dilemmas: No (Obviously) Good Choices

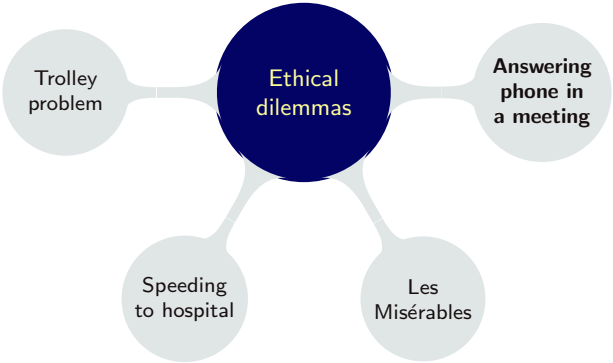






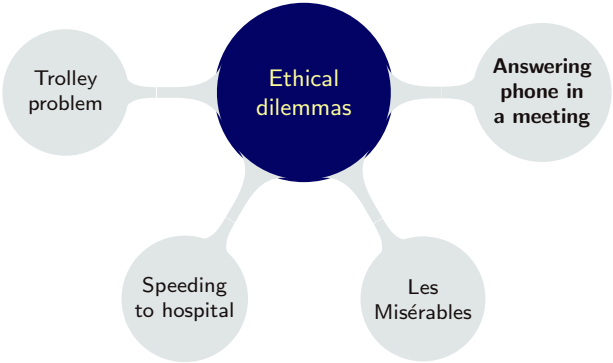
Ethical Dilemmas: No (Obviously) Good Choices

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

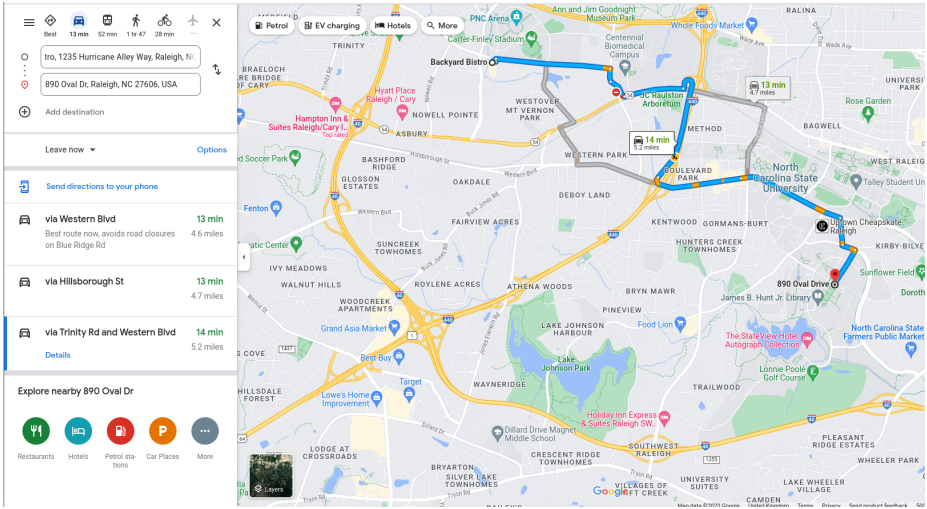


Ethical Dilemmas: No (Obviously) Good Choices

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

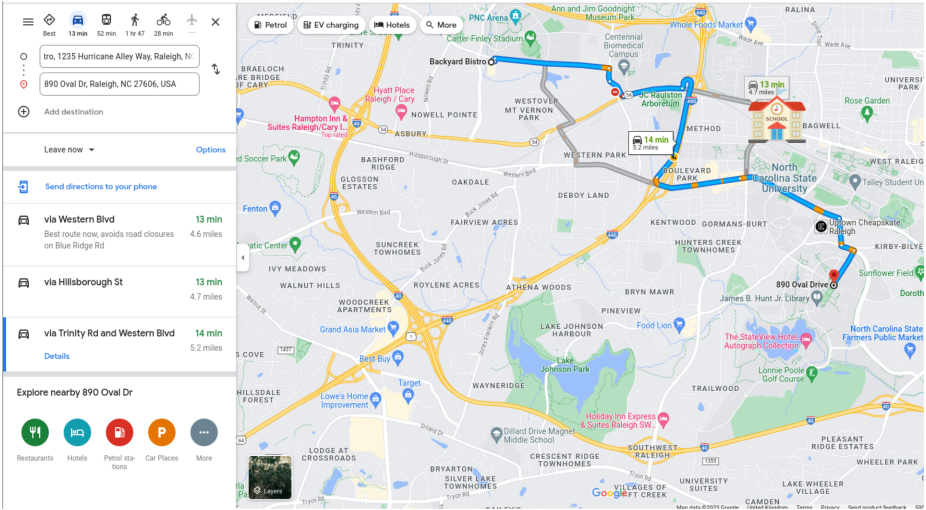


Ethics is inherently a multiagent concern



Source: https://maps.google.com

Tradeoffs: Safety (others?) and Time (yours?)



Source: https://maps.google.com

Tradeoffs: Safety (others?) and Time (yours?)

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: *Conference attendees are committed to wear face masks at all times*
- Prohibition: *Students are prohibited to use their phones in examination halls*

Deviation is a perceived violation of a norm

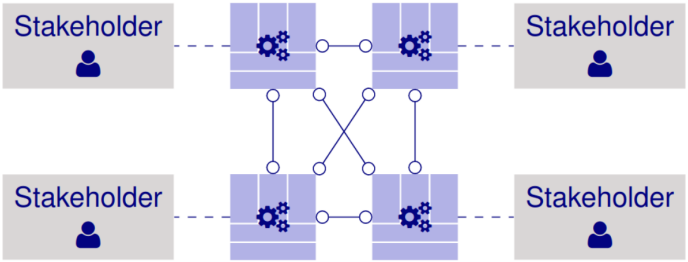
Social context is the circumstance under which the interaction takes place

Values are guiding principles of humans

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

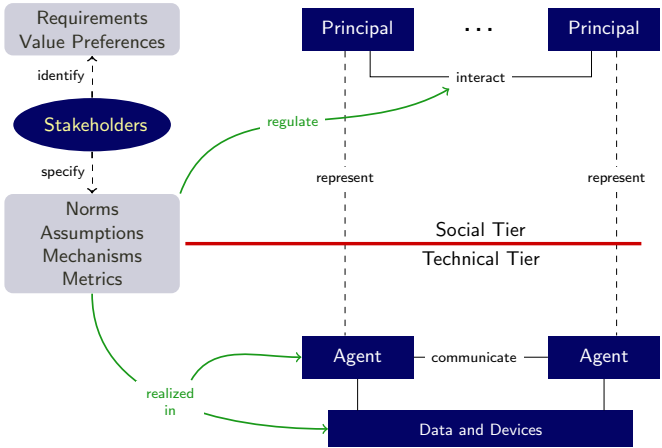
Ethics in Society with SIPAs

SIPA: Socially intelligent (personal) agent



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

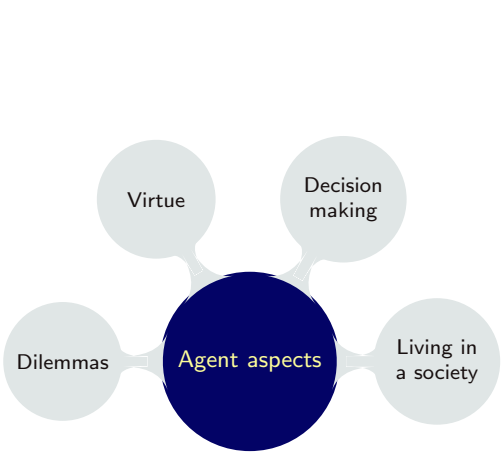
Schematic of a Sociotechnical System (STS)

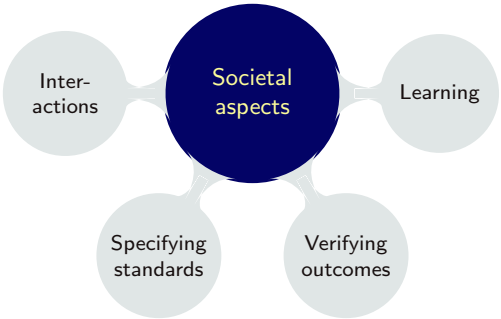


[IS 2016] Kafalı, Ajmeri, and Singh. 2016. Revani: Revising and Verifying Normative Specifications for Privacy. IEEE Intelligent Systems (IS) 31(5) 8–15.
[TOSEM 2020] Kafalı, Ajmeri, and Singh. 2020. Desen: Specification of Sociotechnical Systems via Patterns of Regulation and Control. ACM Transactions on Software Engineering and Methodology (TOSEM) 29(1) 7:1–7:50.

Ethics and Fairness in Sociotechnical Systems

Both are inherently multiagent concerns, yet current approaches focus on single agents





Socially Intelligent Personal Agent (SIPA)

A SIPA adapts to social context and supports meeting social expectations

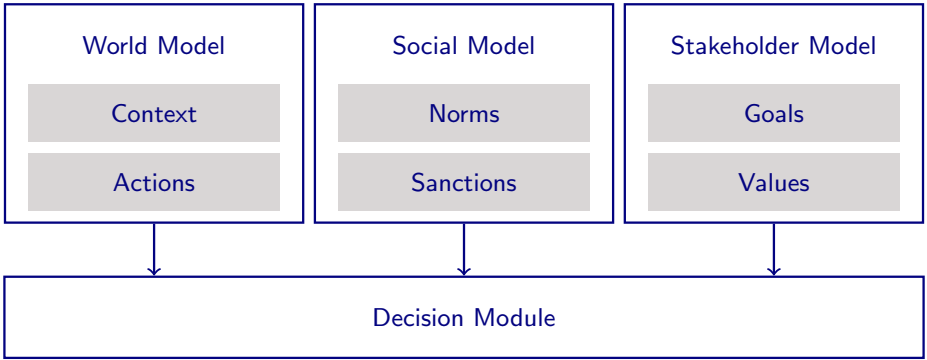
- Ethical: Seeks to balance needs of
  - Primary user (also a stakeholder), who directly interacts with the agent
  - Other stakeholders, who are affected by the agent's actions

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

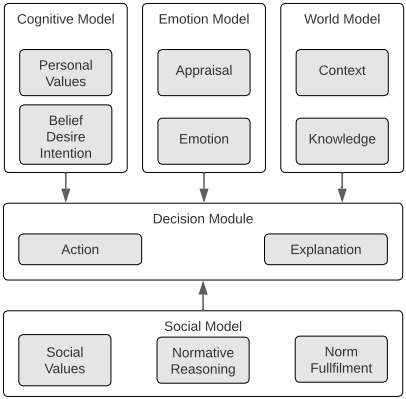
What must a SIPA represent and reason about to participate ethically in a multiagent system?
A SIPA's decision making takes into account its stakeholders, primary and secondary



[AAMAS 2020] Ajmeri et al., Elessar: Ethics in Norm-Aware Agents. Proc. AAMAS, 1–9.
[2023] Sz-Ting Tzeng. Understanding the Interplay of Social Signals and Values in Norm Emergence. Ph.D. Dissertation. NC State University

A SIPA: Schematically

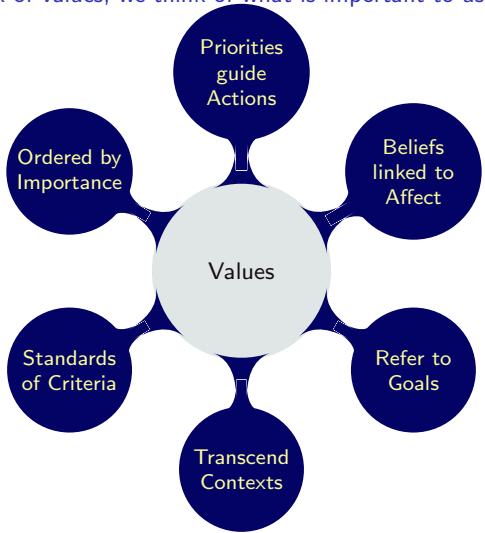
What must a SIPA represent and reason about to participate ethically in a multiagent system?
A SIPA's decision making takes into account its stakeholders, primary and secondary



[AAMAS 2020] Ajmeri et al., Elessar: Ethics in Norm-Aware Agents. Proc. AAMAS, 1–9.
[2023] Sz-Ting Tzeng. Understanding the Interplay of Social Signals and Values in Norm Emergence. Ph.D. Dissertation. NC State University

The Nature and Features of All Values

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



Incorporating Values in an Agent Model

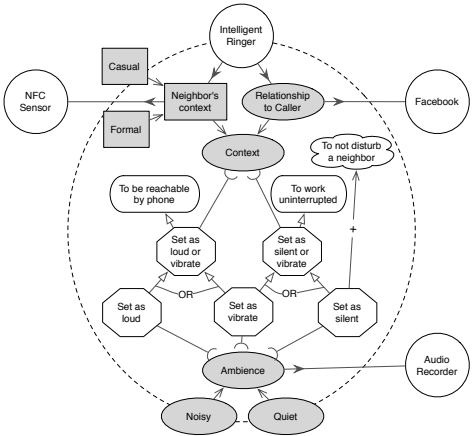
Agent models provide **technical abstractions** to represent values

- 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.,

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

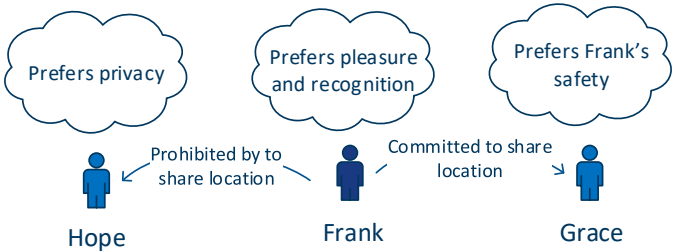
A contextual model of Intelligent Ringer



[AAMAS 2014] Murukannaiah and Singh. Xipho: Extending Tropos to engineer context-aware personal agents. Proc. AAMAS, 309–316. 2014.
[IC 2018] Ajmeri et al., Designing Ethical Personal Agents. IEEE Internet Computing, vol 22(2), 16–22. 2018.

From Personal Values to Social Norms

Consider an example of values in a location sharing app



Frank's dilemma: Which sharing policy to select?

Share with all: Pleasure for Frank ⇑

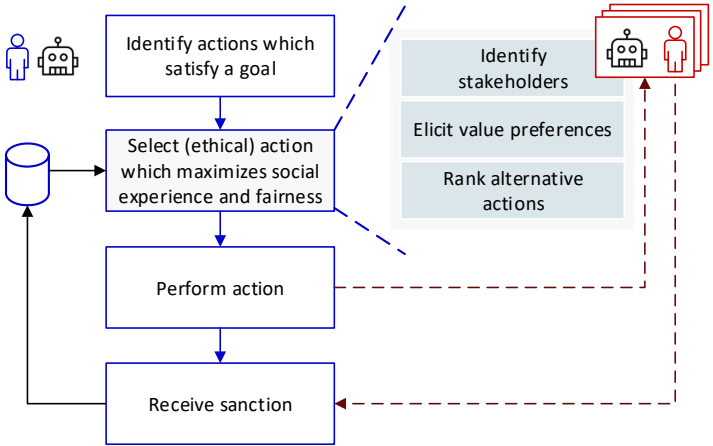
Share only with Grace: Safety for Grace ⇑

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?

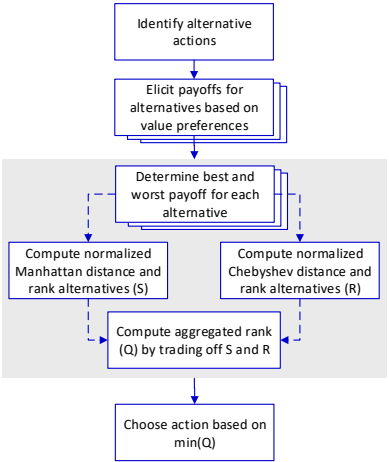
A SIPA's secondary stakeholders can change with the context



[AAMAS 2020] Ajmeri et al., Elessar: Ethics in Norm-Aware Agents. Proc. AAMAS, 1–9.

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 should Jess, Dan, and Alex Go?

Contrasting various ethical principles

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

Total Happiness

- Pancake: 22
- Pasta: 21
- Pizza: 20

Restaurant Example: Where should Jess, Dan, and Alex Go?

Contrasting various ethical principles

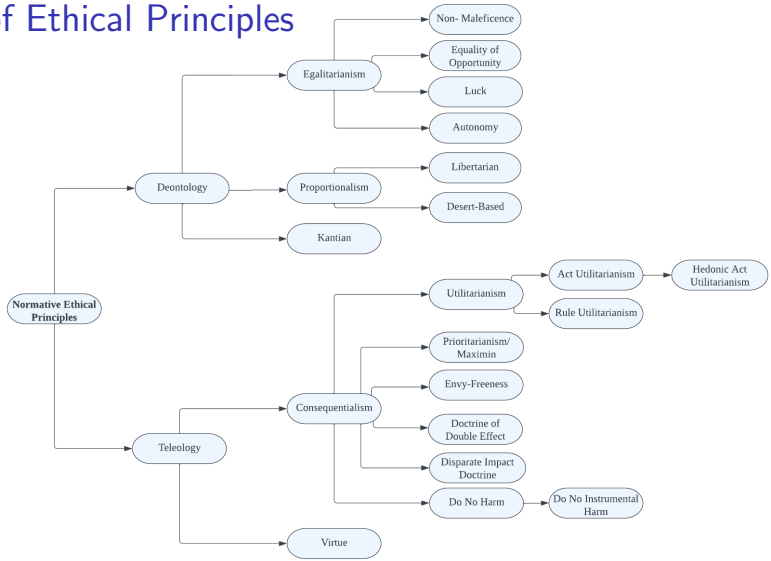
	Jess	Dan	Alex
Pancake restaurant	10	10	2
Pasta restaurant	7	7	7
Pizza restaurant	5	5	10

Total Happiness

- Pancake: 22
- Pasta: 21
- Pizza: 20

What if Jess, Dan, and Alex have to go on a lunch for three consecutive days?

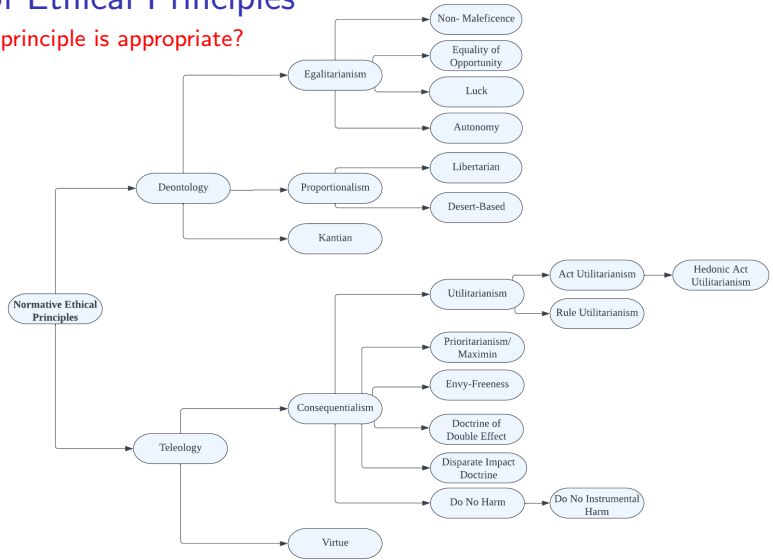
Taxonomy of Ethical Principles



Woodgate and Ajmeri. Principles for Macro Ethics of Sociotechnical Systems: Taxonomy and Future Directions. arXiv preprint arXiv:2208.12616.

Taxonomy of Ethical Principles

Challenge: Which principle is appropriate?

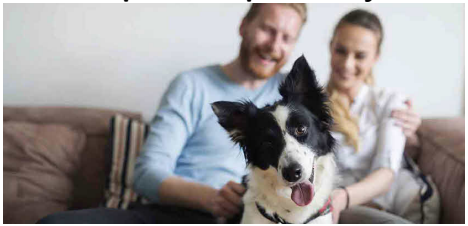


Woodgate and Ajmeri, Principles for Macro Ethics of Sociotechnical Systems: Taxonomy and Future Directions, arXiv preprint arXiv:2208.12616.

Responsibility and Norms

Most reasoning is about explicit agreements

Explicit Responsibility



https://www.vetvoice.com.au/

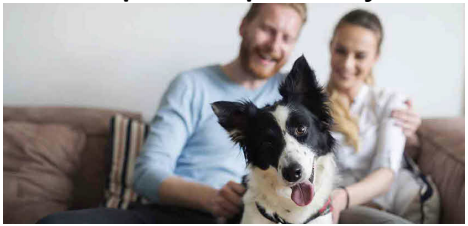
Implicit Responsibility

https://friendsofbcas.org

Responsibility and Norms

Most reasoning is about explicit agreements

Explicit Responsibility



https://www.vetvoice.com.au/

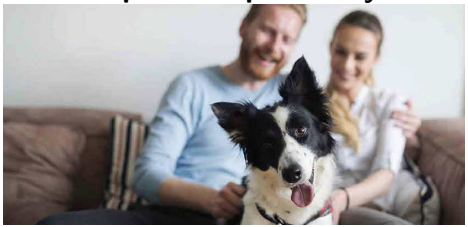
Implicit Responsibility

https://friendsofbcas.org

Responsibility and Norms

Most reasoning is about explicit agreements

Explicit Responsibility



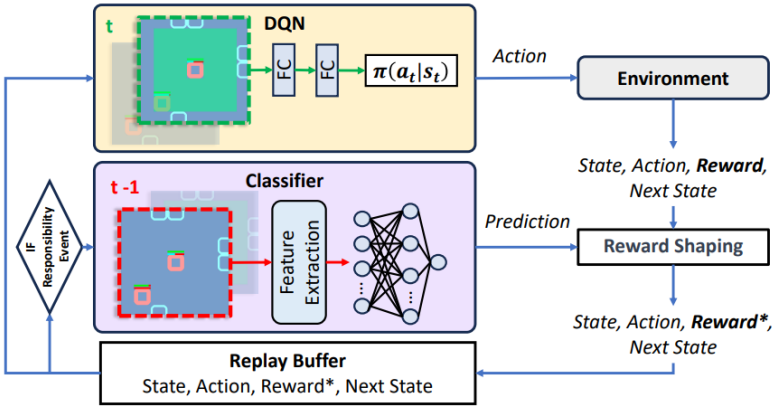
https://www.vetvoice.com.au/

Implicit Responsibility

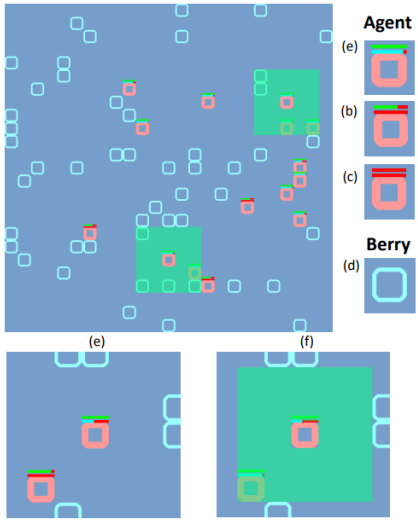
https://friendsofbcas.org

Can implicit responsibilities be learned as social norms?

Identifying Responsibility Events



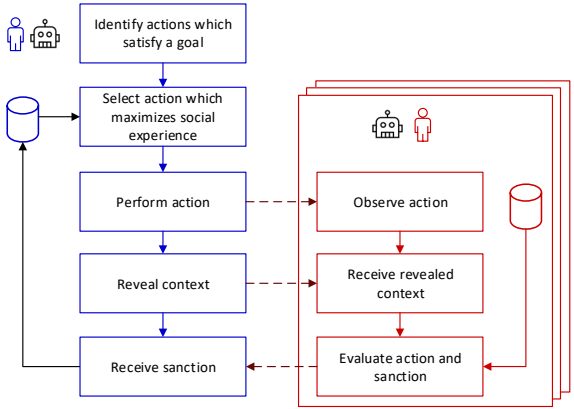
Ongoing Work: The Foraging Environment



Agents (a) use energy to move around to collect berries (d), which confers some reward.
When energy is zero (b), health starts to deplete, and the agent cannot move.
When health reaches zero (c), the agent dies and stops observing and acting.
Berries (d) spawn randomly in the environment and decay after some time.
Collected berries can be eaten to restore health and energy; or gifted to another agent (f).
Agent enters a responsibility event (e) if A is close to an agent B, with zero energy, and A has a berry they can gift to B.

Explaining an Action using Values and Norms

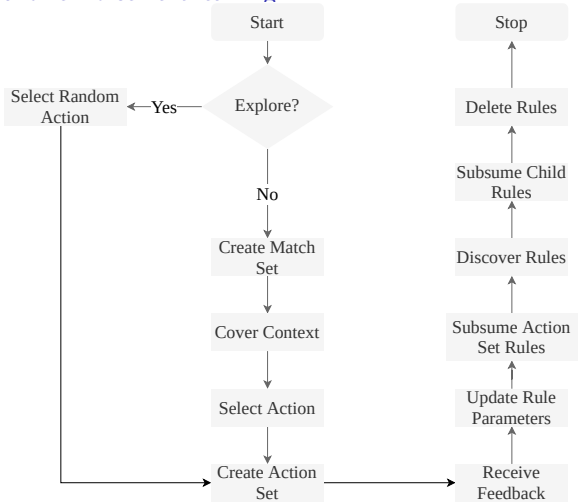
Deviating SIPAs explain their deviations by sharing elements of their contexts



[IJCAI 2022] Agrawal, Ajmeri, and Singh. Socially Intelligent Genetic Agents for the Emergence of Explicit Norms. Proc. IJCAI, 10–16.
[IJCAI 2018] Ajmeri et al., Robust Norm Emergence by Revealing and Reasoning about Context: Socially Intelligent Agents for Enhancing Privacy. Proc. IJCAI, 28–34.

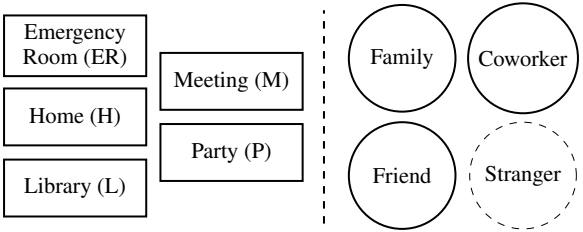
Generating Explanations

SIPAs use genetic algorithm and reinforcement learning



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

Evaluation: The Ringer Environment



Agent societies:

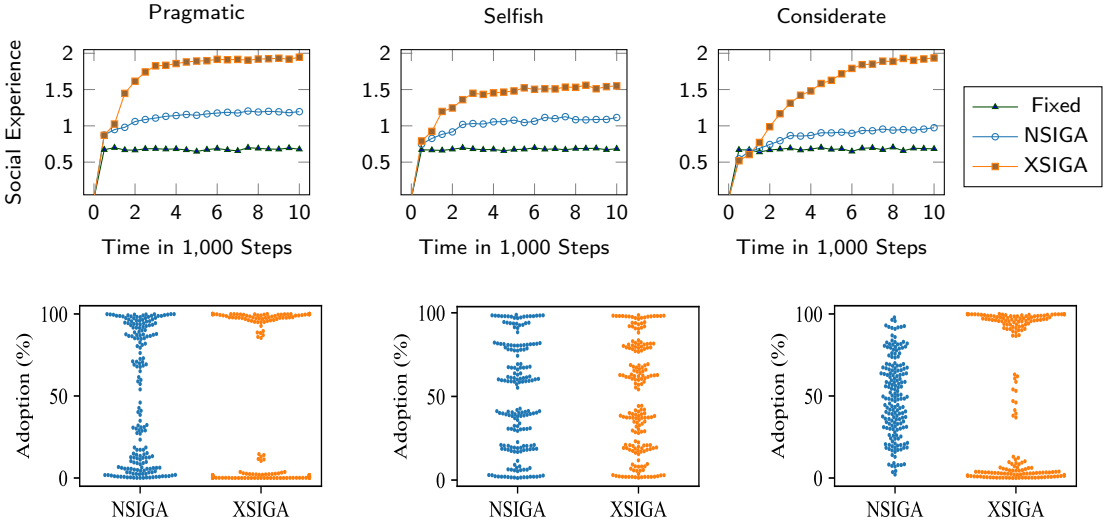
- Pragmatic
- Considerate
- Selfish

Learning strategies:

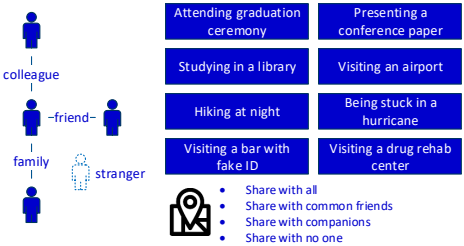
- Fixed
- Sanctioning
- Sharing Context (SIPA) or Explanations (XSIGA)

Results: Social Experience and Norm Adoption

SIPAs yield higher social experience and norm adoption than baselines



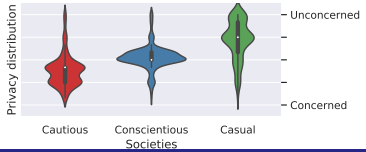
Evaluation: The Context-Sharing Environment



Simulated societies:

- Mixed
- Cautious
- Conscientious
- Casual

Privacy attitude:



Decision-making strategies:

$S_{Elessar}$: Policy based on VIKOR

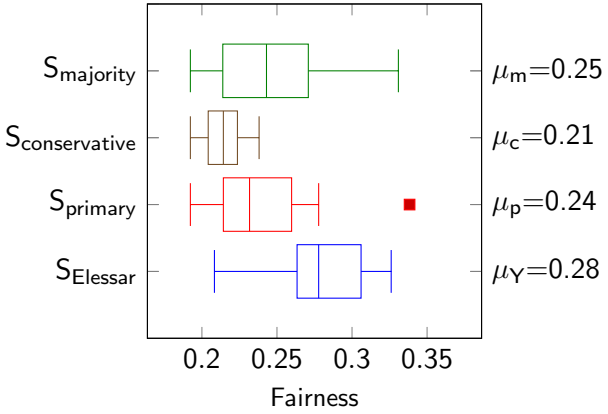
$S_{primary}$: Primary user's preference

$S_{conservative}$: Least privacy-violating

$S_{majority}$: Most common

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
- Ethics and fairness considerations apply in mundane settings—anywhere we have *multi-user multi-agent* scenario
- Socially intelligent agents could help stakeholders navigate social norms of the society and support selecting ethically-appropriate actions

Opportunities and Directions

- How to promote justice and fairness at the community level through individual decision making?
- How can an agent infer and elicit its user's values and value preferences *unintrusively*?
- How can we support decision making by an agent that takes into account the value preferences (of principals and STS)?
  - aggregate value preferences considering the current social context?
  - incorporate ideas on emotions, guilt, consent, and prosociality?
  - formally verify that agent's decisions align with STS specification and stakeholder requirements?
- How can an agent explain its decisions to its users and to other agents and yet *preserve privacy*?
- Realistic data for realistic simulation?

Thank You

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

Acknowledgement

- Munindar P. Singh — NC State University
- Pradeep K. Murukannaiah — TU Delft
- Sz-Ting Tzeng – NC State University
- Hui Guo – Quora
- ...
- Dan Collins, Jessica Woodgate, Alex Davies
- Conor Houghton, Paul Marshall, and Telmo Silva Filho at Bristol