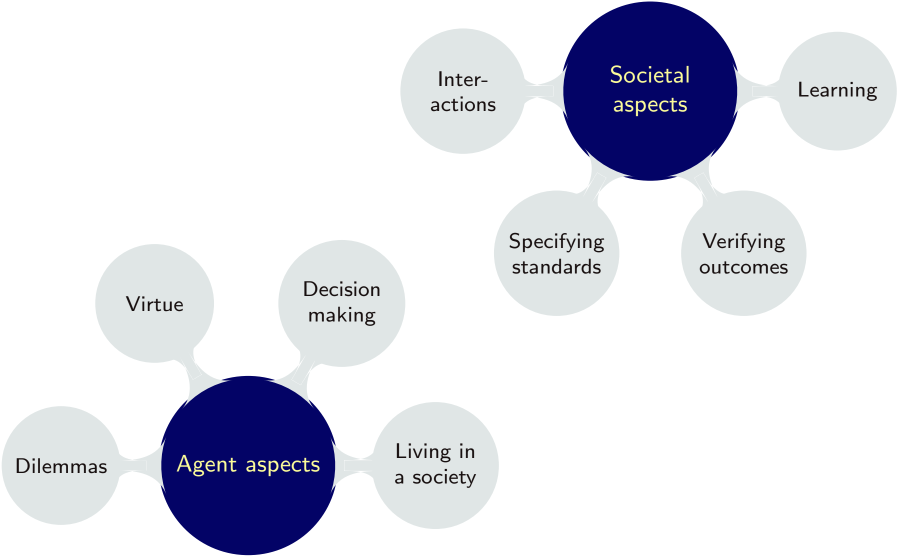
Engineering Ethical Multiagent Systems

Munindar P. Singh
singh@ncsu.edu
https://www.csc.ncsu.edu/faculty/mpsingh/
(Work with Nirav Ajmeri and Amit Chopra)
(With help from Hui Guo and Pradeep Murukannaiah)

Department of Computer Science
North Carolina State University

Ethics in Multiagent Systems

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


Societal aspects

Interactions

Learning

Specifying standards

Verifying outcomes

Agent aspects

Virtue

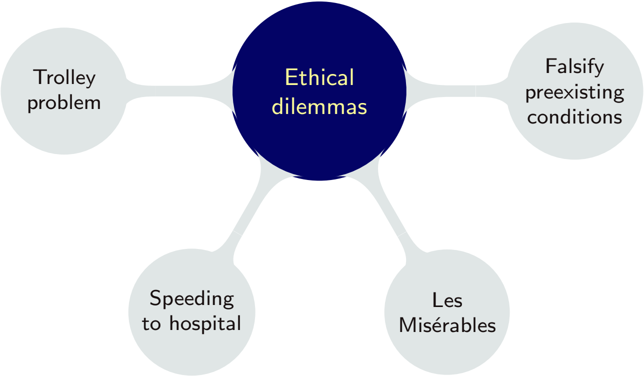
Decision making

Dilemmas

Living in a society

Ethical Dilemmas: No Good Choices

Contrast the following examples



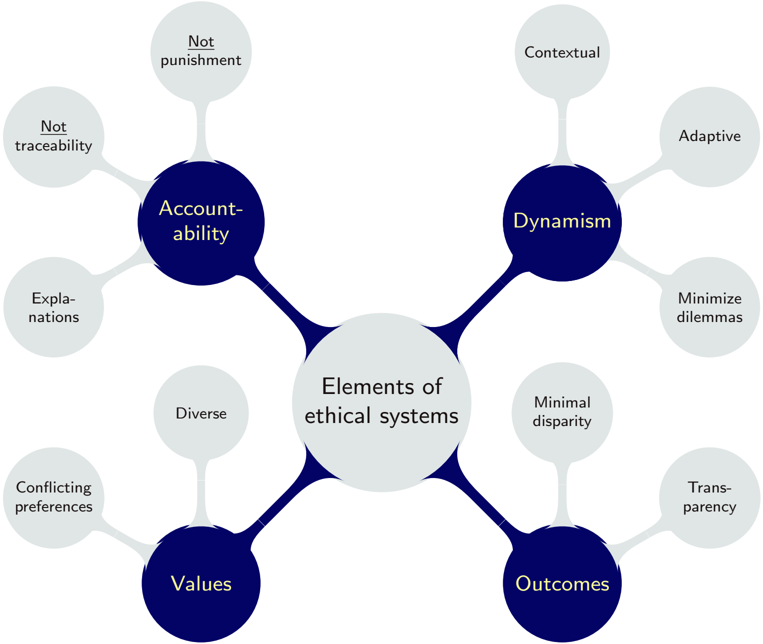
Ethical dilemmas

Trolley problem

Falsify preexisting conditions

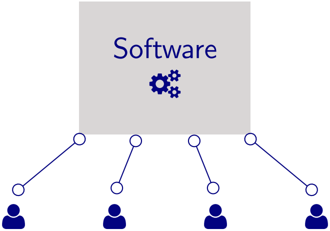
Speeding to hospital

Les Misérables


Not punishment
Not traceability
Accountability
Explanations
Contextual
Adaptive
Dynamism
Minimize dilemmas
Elements of ethical systems
Diverse
Conflicting preferences
Values
Minimal disparity
Transparency
Outcomes

Fairness of a Central Technical Entity

Today's view of fairness involves how an agent deals with people
Such as a prediction algorithm or an autonomous vehicle

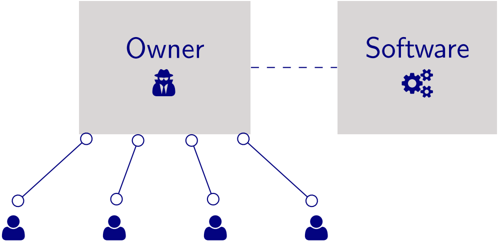


- Autonomy is defined as automation: complexity and intelligence
- Dilemmas à la trolley problems approached in an atomistic manner

Fairness of a Social Entity Equipped with Software

A social entity, assisted by software, wields power over people

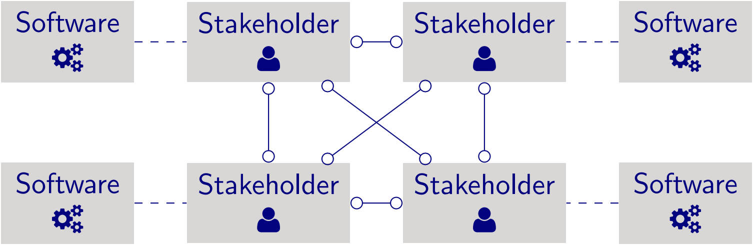
Ethical concerns focused on social entity



- Autonomy as a social construct; mirror of accountability
- Accountability rests with the social entity
- Powers and how they are exercised

Ethics in Society

Ethical considerations and accountability arise in how social entities interact



- The society itself is modeled
- Autonomy is in reference to a society
- Introduces a context to the decision making

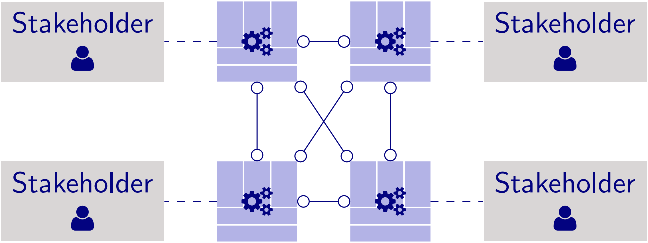
Societal Model of Ethics

Rawls: “political not metaphysical”

- Ethics is a cousin of governance
- An ethical society is one that produces ethical outcomes for its members
    - Rawls’ *difference principle*: reduce the difference in outcomes between best and worst
    - Termed *maximin* in economic terms

Ethics in Society with SIPAs

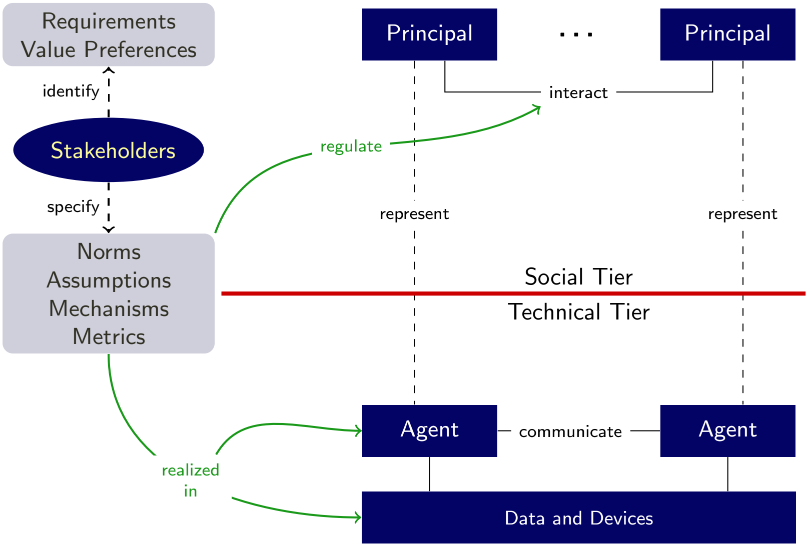
SIPA: Socially intelligent (personal) agent



- A multiagent system is a microsociety
- Each agent reflects the autonomy of its (primary) stakeholder
- How can we realize a multiagent system based on the value preferences of its stakeholders?

Sociotechnical Systems

Current AI research: atomistic, single-agent decision-making focused on ethical dilemmas
Current social sciences research: Not computational in outlook



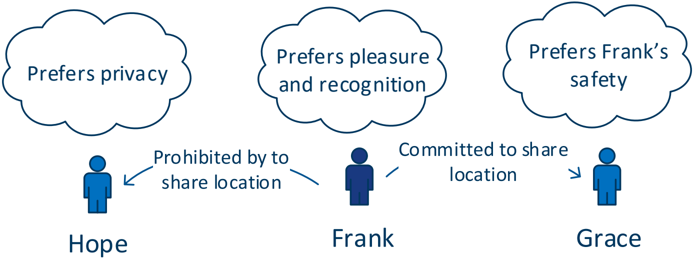
Sociotechnical Systems (STS): A Computational Norm-Based System

Context of interaction in which principals are represented by agents

- Principal: human or organization, a stakeholder who acts
- Norm: *directed* social expectation between principals
  - Types: Commitment, prohibition, authorization, power, ...
  - Standards of correctness
    - *Prima facie*, satisfaction is ethically desirable and violation undesirable
- Accountability: the power of a principal to call another to account for its actions
  - Derives from norms
  - Provides an opportunity for principals to explain their actions
    - Leading to *prima facie* judgments being reconsidered
  - Is not traceability, which is merely a supporting mechanism
  - Is not blame and sanction, which are subsequent

Example: Information Sharing

Frank: committed to his mother Grace to share his location; visits aunt Hope in NYC


Prefers privacy
Prefers pleasure and recognition
Prefers Frank's safety
Prohibited by to share location
Committed to share location
Hope
Frank
Grace

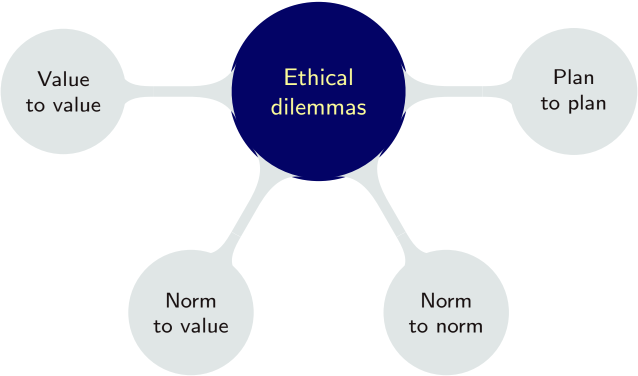
Frank's dilemma: Which sharing policy to select

- Share with all: Pleasure for Frank ⇑
- Share only with Grace: Safety for Grace ⇑
- Share with no one: Privacy for Hope ⇑

Ethical Dilemmas in STS Terms

Dilemma: When there are no good choices

Ethical dilemma: A dilemma involving values



Ethical STS: An Objective for Governance

An STS S is ethical

at time t for value preferences V
if and only if
S's outcomes align with V at t

- Relativist: Value preferences provide frame of reference
- Omits norms—only value preferences matter
  - Norms are crucial (only) for operationalization
- Dynamic: An STS may become ethical (unethical) due to responsive (unresponsive) governance

Ethics in the Large: Values and Outcomes

Emphasizes social abstractions; deemphasizes internal decision-making

- Is an *STS* ethical?
  - Unethical systems make it difficult for principals to make ethical decisions
- Norms operationalize the ethics
  - Implement the "political" and sidestep some of the "metaphysical"
  - Reduce the complexity of individual decision making
- Accountability is conducive to innovation
  - Explanations provide a basis for reconsidering the norms

Ethics in the Large: Accountability and Adaptivity

An ethical STS presupposes good governance

An adaptive methodology undertaken by stakeholders of an STS

- Identify each stakeholder's value preferences
- Specify the norms that support those value preferences
    - Norms are operational refinements of value preferences
    - Norms make accountability concrete
- A stakeholder's SIPA
    - Adopts one or more roles
    - Carries out its part of an enactment
    - Evaluates outcomes on its (primary and secondary) stakeholders
        - Whether values are promoted in alignment with the preferences
        - Which norms are satisfied
- Iterate

Methodology and Tools for Ethical Multiagent Systems

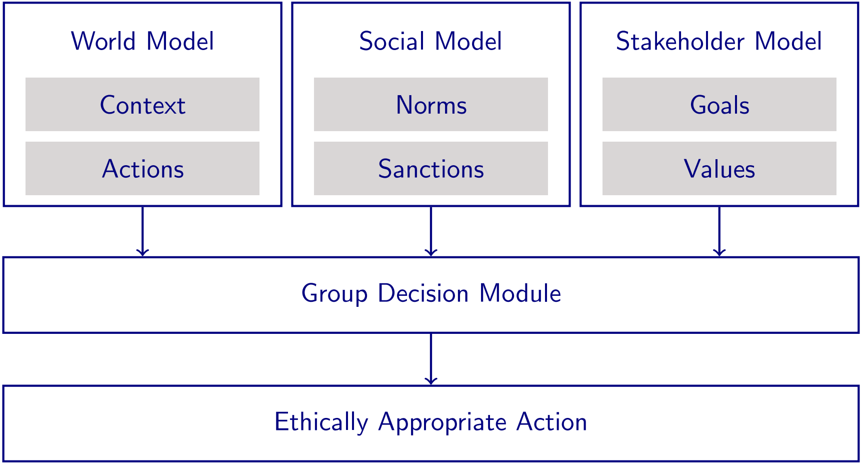
A blend of software engineering, data science, political science, philosophy, and economics

- How can we effectively elicit value preferences from stakeholders?
- How can we identify norms to operationalize those values?
- How can we support effective participation of stakeholders?
    - How may we accommodate their conflicting value preferences?
- How can we evaluate outcomes and revisit the norms to improve alignment of outcomes and value preferences?

Architecture of a SIPA

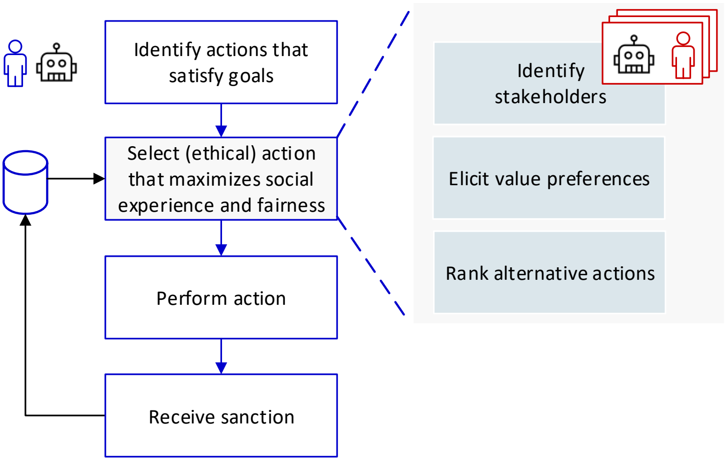
What must a SIPA represent and reason about to participate ethically in a multiagent system?

A SIPA's decision making takes into account its stakeholders, primary and secondary



Interaction in Elessar

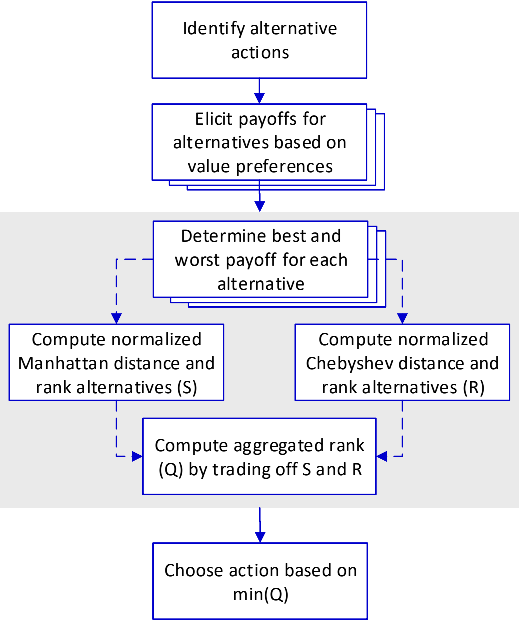
A SIPA's secondary stakeholders can change with the context



Identify actions that satisfy goals
Select (ethical) action that maximizes social experience and fairness
Perform action
Receive sanction
Identify stakeholders
Elicit value preferences
Rank alternative actions

Choosing an Ethical Action

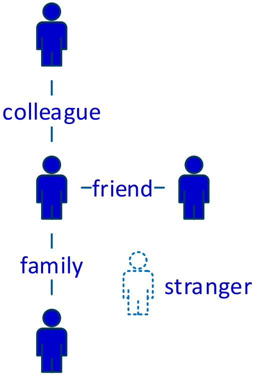
Elessar SIPAs adapt VIKOR to trade off group and individual experience



Identify alternative actions
Elicit payoffs for alternatives based on value preferences
Determine best and worst payoff for each alternative
Compute normalized Manhattan distance and rank alternatives (S)
Compute normalized Chebyshev distance and rank alternatives (R)
Compute aggregated rank (Q) by trading off S and R
Choose action based on min(Q)

Setting: Information Sharing

Places, companion, and sharing policies



Safe | ¬Sensitive
Attending graduation ceremony

Safe | ¬Sensitive
Presenting a conference paper

Safe | ¬Sensitive
Studying in a library

Safe | ¬Sensitive
Visiting an airport

¬Safe | ¬Sensitive
Hiking at night

¬Safe | ¬Sensitive
Being stuck in a hurricane

Safe? | Sensitive
Visiting a bar with fake ID

Safe? | Sensitive
Visiting a drug rehab center

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

Evaluation: Crowdsourcing Study

Schnorff et al.'s privacy attitude survey: Level of comfort in sharing personal information

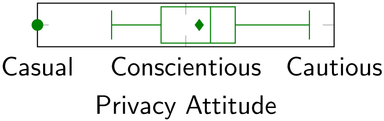
Level of comfort in setting context sharing policy

- Context includes place, activity, and social relationship with companions
- Places provided by us but not their safety and sensitivity ratings

Priming Based only on context to prime the users

Survey Based on context and value preferences (pleasure, privacy, recognition, safety)

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



Example Numeric Utility Matrix for a Stakeholder

Captures value preferences, one per row

Describes the payoff resulting from applying the sharing policy in the specified place with the specified companion

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

Multi-Criteria Decision Making

Example 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	0.75
y_2 Common	5	5	5	10	5	0	5	5	4.0	3.0	1.00
y_3 Grace	0	5	0	0	5	15	5	5	**3.0**	**1.0**	**0.00**
Weight, w_x	1	1	1	1	1	3	1	1			
Max payoff, f_x^*	10	5	10	10	5	15	5	5			
Min payoff, f_x^-	0	5	0	0	5	0	5	5			

Here,

- S_y is the Manhattan distance normalized to the maximum
- R_y is the Chebyshev distance normalized to the maximum
- Q_y is the average of the two, normalized to $[0, 1]$

Measures of Ethicality

For each interaction, . . .

Best individual experience is the maximum utility obtained across the SIPA's stakeholders during a single interaction

Worst individual experience is the minimum utility obtained across the SIPA's stakeholders during a single interaction

Social experience is the utility obtained by a society as a whole divided by the number of stakeholders

Fairness is the reciprocal of the difference between the best and worst individual experience

Evaluation: Simulation

Study unit: A context-sharing SIPA

Decision-making strategies:

S_{Elessar}: Policy based on VIKOR

S_{primary}: Policy based on primary stakeholder's preferences

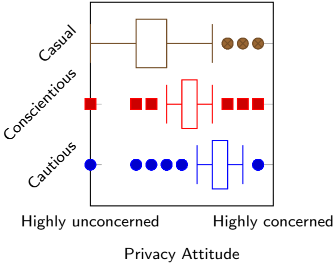
$S_{\text{conservative}}$: Least privacy-violating sharing policy

S_{majority}: Most common sharing policy

Simulated societies

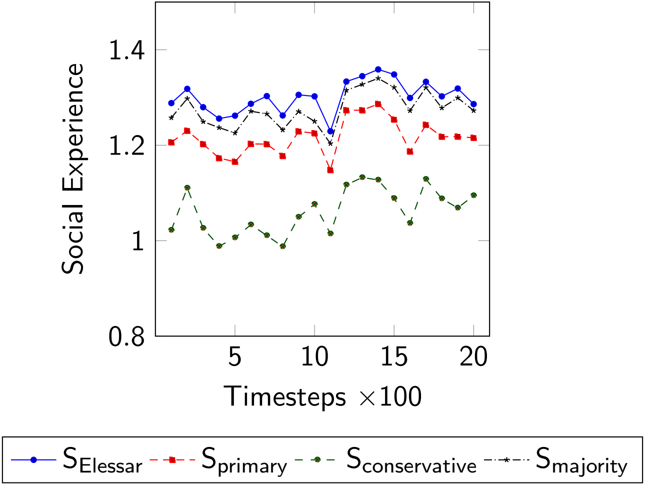
- Mixed
- Cautious
- Conscientious
- Casual

Privacy attitude distribution of societies



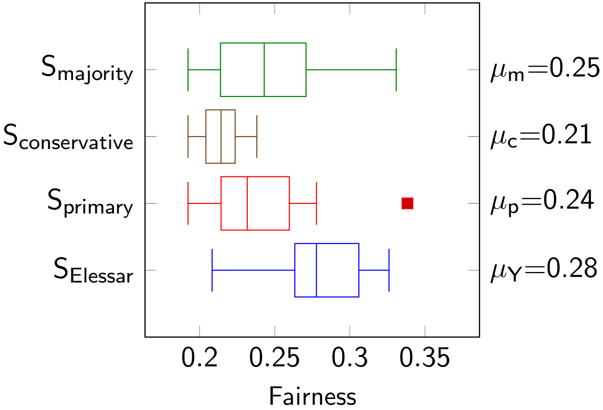
Experiment: Society of Mixed Privacy Attitudes

Result: Elessar yields higher social experience ($p < 0.01$; Glass' $\Delta > 0.8$ indicating large effect size)



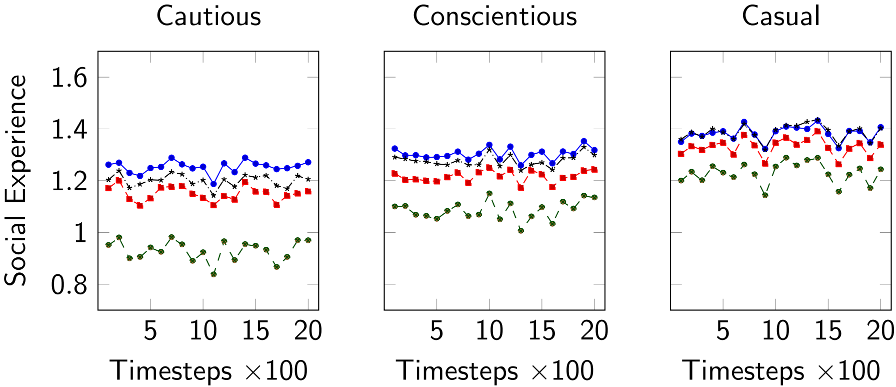
Fairness: Experiment with Mixed Privacy Attitudes

Results: Fairness in a mixed society. Elessary gives significantly better ($p < 0.01$) fairness with large effect size (Glass' $\Delta > 0.8$) than the baselines



Experiments: Three Societies of Majority Privacy Attitudes

Result: Elessar yields superior social experience than the other decision-making strategies across three types of societies (shown) without hurting the other metrics (not shown)



Comparing Metrics for a Society of Mixed Privacy Attitudes

Strategy	Social	Best	Worst	Fairness
S_{Elessar}	**1.36**	1.72	**0.77**	**1.05**
S_{primary}	1.29	1.79	0.58	0.83
$S_{\text{conservative}}$	1.11	1.72	0.47	0.80
S_{majority}	1.34	**1.84**	0.57	0.78

Bold indicates the winner

Comparing Metrics for a Societies with Majority Privacy Attitudes

Strategy	Cautious				Conscientious				Casual			
	S.	B.	W.	F.	S.	B.	W.	F.	S.	B.	W.	F.
S_{Elessar}	1.54	1.66	**1.23**	**2.27**	**1.33**	1.53	**0.87**	**1.51**	**1.24**	1.46	**0.77**	**1.45**
$S_{\text{pri.}}$	1.51	1.77	1.08	1.46	1.25	1.59	0.68	1.10	1.13	1.47	0.58	1.13
$S_{\text{cons.}}$	1.37	1.75	1.06	1.46	1.09	1.52	0.61	1.10	0.87	1.34	0.45	1.34
$S_{\text{maj.}}$	**1.55**	**1.86**	1.01	1.18	1.32	**1.70**	0.58	0.89	1.18	**1.53**	0.52	0.98

Bold indicates the winner

Conclusions

- Ethics inherently involves looking beyond one's narrow interest
- Ethical considerations apply in mundane settings—anywhere agents of multiple stakeholders interact
- A multiagent understanding of ethics can provide a foundation for a science of security and privacy

Elements of Ethics: From Agents to Systems

	Agent Level	System Level
Scope	Individual	Individual in society
Autonomy	Intelligence and complexity	Decision making in social relationships
Transparency	About data and algorithms	About norms and incentives
Bases of Trust	Construction and traceability	Norms and accountability
Fairness	Preset criteria: Statistics	Reasoning about others' outcomes
Focus	Dilemmas for individuals	System properties

Thanks!

- Science of Security Lablet
- Laboratory of Analytic Sciences

http://www.csc.ncsu.edu/faculty/mpsingh/
https://research.csc.ncsu.edu/mas/