Robust Norm Emergence by Revealing and Reasoning about Context: Socially Intelligent Agents for Enhancing Privacy

Nirav Ajmeri[†], **Hui Guo**[†], **Pradeep K. Murukannaiah**[‡] and **Munindar P. Singh**[†]
[†]Department of Computer Science, North Carolina State University, Raleigh, NC 27695, USA
[‡]Department of Software Engineering, Rochester Institute of Technology, Rochester, NY 14623, USA
najmeri@ncsu.edu, hguo5@ncsu.edu, pkmvse@rit.edu, mpsingh@ncsu.edu

Abstract

Norms describe the social architecture of a society and govern the interactions of its member agents. It may be appropriate for an agent to deviate from a norm; the deviation being indicative of a specialized norm applying under a specific context. Existing approaches for norm emergence assume simplified interactions wherein deviations are negatively sanctioned. We investigate via simulation the benefits of enriched interactions where deviating agents share selected elements of their contexts. We find that as a result (1) the norms are learned better with fewer sanctions, indicating improved social cohesion; and (2) the agents are better able to satisfy their individual goals. These results are robust under societies of varying sizes and characteristics reflecting pragmatic, considerate, and selfish agents.

1 Introduction

Social *norms* provide a robust means to regulate interactions in human society. Our everyday actions tend to *comply* with social norms. For example, *ignoring a phone call during a meeting* and *remaining silent in a public library* are expected behaviors that accord with social norms. However, we often *deviate* from the applicable social norms, for instance, when *stepping out of a meeting to answer a phone call.*

The ability to deviate from norms is crucial for autonomy. We may *sanction* each other based on how we are interacting. In particular, negative sanctions in response to deviations are a means for establishing norms [Andrighetto *et al.*, 2013]. For example, when a meeting attendee's phone rings, a *scowl* on other attendees' faces hints at a norm of *keeping one's phone silent during meetings.*

Existing approaches for norms provide simplified interactions: a deviation or not, followed by a sanction or not. But real-life interactions are more complex. Whether a deviation leads to a positive or negative sanction depends on how others perceive its *context* or circumstances of occurrence. When we deviate from a norm, we may offer an apology, describing the context. One, revealing context may soften a deviation and help avert negative sanctions. Suppose, upon receiving a call during a meeting, Alice says that the call was from her sick father. As a result, the meeting attendees may excuse Alice for taking the call. A deviation may result in a positive sanction. For instance, a physician who reveals a patient's private data to save the patient's life would receive a positive sanction despite violating a norm. Even in the phone call setting, a positive sanction may ensue for deviating from a norm. For example, a user who hesitantly takes a call from his nine-month pregnant wife during a lab meeting would generally receive positive comments from coworkers. Two, context helps refine the relevant norms. For example, Alice's revelation may help refine the norm from *ignoring a phone call during a meeting* to *ignoring a phone call during a meeting, unless the call is urgent.* In essence, deviation context and any ensuing sanction help characterize the boundaries of a norm in play.

Accordingly, we propose Poros, an approach for building agents that carry out enriched interactions where deviating agents share selected elements of their contexts, and others respond appropriately. A socially intelligent personal agent (SIPA) is an agent who acts in accordance with (but may deviate from) social norms [Ajmeri *et al.*, 2017]. We imagine an artificial agent society in which SIPAs of three main types act and interact on behalf of (human) users, as a basis for empirically investigating the emergence and quality of norms.

This research applies in developing privacy-supporting SIPAs. Norms provide a basis for understanding privacy [Nissenbaum, 2011]. Regulations about information disclosure, as in healthcare, are context-dependent norms [Ajmeri *et al.*, 2016], as are social practices. Privacy involves control over when and what information to disclose [Westin, 1967]. In some construals, actions that intrude upon one's solitude or bring disapprobation are privacy violations. In essence, all privacy-relevant interactions are modulated by norms. Therefore, social intelligence in making decisions cognizant of norms while preserving social cohesion is crucial.

Our main contribution is to study two research questions in light of a specific decision by a SIPA, namely, whether to reveal its context to others when it deviates from a norm:

Q_1 Norm: Does revealing context and reasoning about revealed context promote emergence of robust social norms?

Q_2 Goal: Does acting in accordance to such robust norms result in an improved goal satisfaction?

Our results show that (1) norms that emerge in Poros are robust, implying improved social cohesion and (2) SIPAs yield higher goal satisfaction to their users when acting in Poros than when acting in a conventional setting (just sanctions).

2 Related Work

Research on normative systems has addressed the problems of conflict, compliance, and emergence of norms. We sample some of the literature from the following themes.

Social norms regulate agent interactions by characterizing what behavior one agent may legitimately expect from another in a particular setting [Kafalı *et al.*, 2016; Singh, 2013]. We adopt Singh's [2013] computational representation of social norms. A norm is directed from a subject (stakeholder) to an object (stakeholder), and is constructed as a conditional relationship involving an antecedent (which brings the norm into force) and a consequent (which brings the norm to satisfaction or violation). Ajmeri et al. [2017] introduce Arnor, a method to model social intelligence in personal agents. They argue that personal agents who understand the intricacies of social norms, deviations, and associated arguments can provide a privacy-preserving social experience to their users.

Works on designing *context-aware* agents emphasize modeling [Murukannaiah and Singh, 2014] and sharing [Ajmeri *et al.*, 2017]. Poros is novel in the way it helps SIPAs infer social norms by revealing deviation context and reasoning about context revealed by others. Poros examines the effect of revealing context by agents after norm deviations. Kökciyan and Yolum [2017] propose an argumentation-based approach to enable agents to reason about context and reveal information based on it. Whereas their focus is on understanding the context to make a privacy decision, we demonstrate the benefits of revealing context. Naively revealing context could violate user privacy. However, a SIPA would reveal selectively by evaluating the tradeoff between privacy lost by revealing and sanctioning faced by not revealing. (For simplicity, in our experiments, the context model is simple and the SIPAs always reveal—to demonstrate the benefit of revelation.)

The study of *norm conflicts and compliance* has drawn much interest. An agent may face conflicts between multiple applicable norms [Ajmeri *et al.*, 2016], or between norms and its own goals. Van Riemsdijk et al. [2015] develop a norm compliance framework to design socially adaptive agents in which agents identify and adopt new norms, and determine execution mechanisms to comply with those norms. Van Riemsdijk et al. argue that a personal agent needs explicit norms. Aldewereld et al. [2016] present a formalism and mechanism to comply with group norms. Ajmeri et al. [2016] present a formalism to represent normative conflicts and dominance relationships among conflicting norms. Sugawara [2011] uses reinforcement learning to resolve norm conflicts and shows how social conventions for resolving conflicts emerge. However, the efficiency and stability of the results differ across agents. These works give us insights into defining agents' decision-making processes.

Agent interactions lead to dynamic *norm emergence and evolution* [Savarimuthu *et al.*, 2009]. Boella et al. [2009] propose a normative framework to evaluate and classify normative system change. Mashayekhi et al. [2016] propose a hybrid mechanism for norm emergence and conflict resolution in sociotechnical systems. Villatoro et al. [2013] present social instruments such as "rewiring" and "observation" to assist norm emergence. Yu et al. [2013] suggest using collective, instead of pairwise, learning for norm emergence. Poros is novel in that it supports revealing and reasoning about contextual information to facilitate understanding of contextually relevant norms.

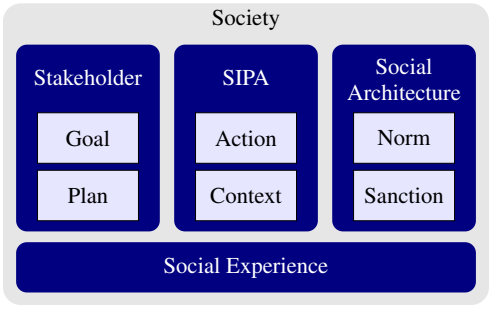

Figure 1: A society of SIPAs and stakeholders.

Sanctions are mechanisms to achieve social coherence. An agent decides whether to comply with or deviate from a norm. A sanction, negative or positive, is associated with the reaction of other agents to this decision. Previous works adopt sanctions as a way to promote norm compliance [Noussair and Tucker, 2005; Egas and Riedl, 2008]. Alechina et al. [2012] present a programming language for norm-aware agents who might deviate from norms and expect sanctions. Nardin et al. [2016] develop a sanction typology and introduce a conceptual sanctioning process model to promote governance in sociotechnical systems. Recent works explore combining norm communication with sanctions to promote cooperation [Andrighetto *et al.*, 2013]. Van Riemsdijk et al. [2015] emphasize understanding norm violations as a basis for designing socially adaptive agents. Poros differs from these works in addressing the problem of understanding a deviation by modeling the context in which a deviation occurs.

3 Interaction in a SIPA Society

A SIPA society we seek to engineer consists of stakeholders, a social architecture, and SIPAs acting on behalf of stakeholders. Figure 1 shows a conceptual model of a SIPA society.

The stakeholders are users, *primary* or *secondary*, depending on the context (defined later). The *primary* stakeholder of a SIPA is the user who directly interacts with it, and on whose behalf the SIPA acts and interacts. A *secondary* stakeholder is the user who may not directly interact with the SIPA, but is affected by the SIPA's actions [Ajmeri *et al.*, 2017]. Each stakeholder has goals and plans.

- A *goal* of a stakeholder describes a state the stakeholder would prefer; a stakeholder may have multiple goals.
- A *plan* of a stakeholder is a set of actions that can bring about one or more goals.

The social architecture of a society captures its structure; it comprises social norms and the sanctions that promote or ensure compliance with norms.

- A *norm* is a tuple of ⟨subject, object, antecedent, consequent, context⟩ [Singh, 2013]. Norms characterize the social architecture that promotes prosocial behavior.
- A *deviation* from a norm occurs when a stakeholder, or *deviant*, performs an action that does not comply with it.
- A *sanction* is a set of actions a stakeholder may take toward a deviant on observing a deviation. A sanction may be positive or negative [Nardin *et al.*, 2016].

A SIPA acts and interacts on behalf of a stakeholder and is aware of the social architecture of the society.

- An *action* is a step a SIPA takes to execute its stakeholder's plan, thereby bringing about the corresponding goal. An action may satisfy or violate a norm. SIPAs in a society can observe each other's actions.
- A *context* captures the circumstances under which a SIPA acts [Dey, 2001]. In our approach, the context is social and incorporates whether a norm is satisfied or violated. Context includes social relationships between stakeholders and spatiotemporal parameters relevant to describing interactions between a SIPA and its stakeholders. We adopt Murukannaiah and Singh's [2012] notion of *place* as a location such as home, library, meeting, or party understood in conceptual terms. Parameters describing a place may include physical conditions (e.g., noise level), expected activities (e.g., reading a book), social interactions (e.g., having a discussion), and temporal information (e.g., during office hours on a weekday).

The social experience a SIPA delivers reflects the extent to which the SIPA promotes its primary and secondary stakeholders' goals. It relates to how a SIPA's stakeholders perceive a norm deviation, and the sanctions they apply. Our objective is to promote each SIPA to act toward maximizing the overall social experience, despite competing interests.

We define social experience (E) as the weighted aggregation of payoffs perceived by a SIPA's stakeholders for each action executed by the SIPA. That is, for each potential action, a SIPA determines the payoffs for its primary and secondary stakeholders, and computes an aggregation as a weighted sum of the payoffs. A SIPA's aggregation method reflects its primary user's preferences and privacy attitudes. For instance, a pragmatic user's SIPA may aggregate payoffs by giving equal weight to all stakeholders, whereas a selfish user's SIPA may give a smaller weight to secondary stakeholders.

Poros Explained with an Example SIPA

We now describe Poros, a framework to build SIPAs, using Ringer, an example SIPA who answers or ignores phone calls on behalf of its primary stakeholder by ringing the phone or keeping it silent. Ringer is a privacy-enhancing technology that acts on behalf of its primary stakeholder; it determines when to allow intrusions, and when to risk being overheard in a phone call (and thus when to intrude on others' solitude).

Ringer's primary stakeholder is the *callee* with privacy goals of *being reachable by phone*, *to work uninterrupted*, and *to not disturb neighbors*. Ringer's secondary stakeholders are (1) a *caller* with the goal *to reach the callee*; and (2) a *neighbor* with a privacy goal *to not be disturbed*. Ringer observes other SIPAs' actions and potentially sanctions them based on their actions and the context as revealed by them.

Each SIPA in Poros maintains a history of interactions and the associated experience. The actual experience is determined after each interaction based on the revealed context and any resulting sanctions. The history helps a SIPA determine the action that would maximize its stakeholder's predicted social experience.

We define a SIPA's history (H) as a set of tuples $h_i = \langle c_i, g, p, N, s_i \rangle$, each of which describes an interaction i, including context c_i describing the circumstances in which goal g is brought about via plan p under a set of applicable norms N, and all resulting sanctions $\{s_i\}$. For Ringer, c_i includes the places where the stakeholders are, their social relationships, and urgency of the incoming call.

Each SIPA maintains its history locally, and scans it when selecting a plan. In a conflict situation, SIPAs look up their history to predict social experience and decide which norms or goals to prefer over which others in a given context; thus infer contextually-relevant norms.

A SIPA's behaviors include acting on behalf of its stakeholder, deciding whether to reveal its context, reasoning about the contexts revealed by others, and issuing sanctions to others. It does so based on knowledge of its context, its stakeholder's goals, associated plan, and applicable norms.

- *Plan selection.* A SIPA selects a plan (and its associated actions) that would achieve its primary stakeholder's goals. In the Ringer example, it selects to ring or keep silent for an incoming phone call. If more than one plan are available, from the history (if available) it identifies the one that maximizes the social experience, or chooses a random plan from the applicable plans with a small probability α.
- *Revealing context.* When a SIPA chooses and executes a plan, it might deviate from some applicable norms. It decides which norms to prefer in the current context and whether to reveal unobserved context to other SIPAs. For instance, if Ringer decides to prefer the *family norm—always answer calls from family* over the *meeting norm—never answer calls during meetings* by ringing during a meeting for an urgent phone call from a sick family member, it reveals the unobserved context, i.e., urgency of the call and the caller's sickness to other meeting attendees. Ideally, a SIPA should selectively reveal context to others according to its stakeholder's goals and privacy attitude.
- *Sanctions.* A SIPA observes other SIPAs' actions, and sanctions them when its stakeholder is affected by their actions. On receiving the context revealed by a deviating SIPA, the SIPA of an affected stakeholder evaluates whether the observed action would be norm compliant in the revealed context. In the Ringer example, *neighbors'* and *caller's* SIPAs decide whether they would ring for an urgent phone call from a sick family member during a meeting and accordingly sanction the *callee's* SIPA.

The complete interaction, including the selected plan and executed actions, observed and revealed context, applicable norms, and sanctions, is recorded in SIPAs history. As SIPAs interact by acting and evaluating actions for norm compliance from interaction history, they understand the boundaries of applicable norms in different contexts, and thus promote emergence of robust social norms.

By place

Place	Response
Emergency (ER)	Answer
Home (H)	Answer
Library (L)	Ignore
Meeting (M)	Ignore
Party (P)	Answer

By circle and call type

Circle	Casual	Urgent
Coworker	Answer	Answer
Family	Answer	Answer
Friend	Answer	Answer
Stranger	Ignore	Answer

Table 1: Norms for answering calls based on (left) place and (right) caller's social circle and casual or urgent call types.

4 Simulation Model

We evaluate Poros via a simulated *ringer environment* built using MASON [Luke *et al.*, 2005].

4.1 The Ringer Environment

The ringer environment contains shared places (home, party, meeting, library, and emergency room). Corresponding to each place, we define social circles such as family, friends, and coworkers. Each agent belongs to a family circle, a friend circle, and a coworker circle. Agents who do not share any of these circles are considered strangers. We define the social network or place network topology in a way such that there is only one type of relationship, i.e., family, friends, coworkers, or strangers, between any pair of agents. In the ringer environment, there are (1) several homes, each corresponding to a family circle, (2) several parties, corresponding to multiple friend circles, and (3) multiple meetings, corresponding to multiple colleague circles. There is one library and one emergency room (ER). The numbers of homes, parties, and meetings follow the network setups specified in Table 6.

In the simulation, agents stay at each place for a random number of steps (averaging 60 steps) and then move. If an agent enters home, party, or meeting, it is more likely to enter the place that is associated with its own social circle than entering a place with strangers. For example, if an agent chooses to enter home, it is likelier to enter its own family's home than to enter a stranger's home. Therefore, when it is at home, an agent is usually surrounded by its family members with only a few strangers.

The agents in the ringer environment perform the following actions depending upon their roles:

- A caller initiates an urgent or a casual phone call.
- A callee answers or ignores a phone call.
- A callee shares context for answering or ignoring a call.
- A caller and neighbors respectively reason about context.
- A caller and neighbors respectively sanction a callee for answering or ignoring a phone call.

Each place and each circle has predefined norms, as defined in Table 1. For example, emergency room (ER) is conceptualized as a place where the default norm is to always answer calls, whereas the norm in a library is to ignore calls. Norms could conflict. For example, the norm to *answer an urgent phone call from a family member* conflicts with *ignore during a meeting*. We let the agents figure out contextually relevant norms in case of conflict.

Caller's Relationship	Callee's Response	Casual	Urgent
Family, Friend, or Coworker	Answer	0.50	1.00
	Ignore	0.00	–0.50
Stranger	Answer	0.00	0.50
	Ignore	0.25	–0.25

Table 2: Payoff for callee for casual or urgent call types.

Callee's Response	Casual	Urgent
Answer	0.50	1.00
Ignore	–0.50	–1.00

Table 3: Payoff for caller for casual or urgent call types.

Callee's Response	ER	H	L	M	P
Answer	1.00	0.67	–1.00	–1.00	–0.33
Ignore	–1.00	–0.33	1.00	1.00	0.67

Table 4: Payoff for neighbors by place (ER, H, L, M, P).

For each phone call, based on the callee's response of answering or ignoring, the caller, callee, and neighbors perceive a fixed payoff, as shown in Tables 2–4.

4.2 Agent Types

To evaluate effectiveness of Poros, we define two baseline agent types—*Fixed* and *Sanctioning*, other than Poros agents.

Fixed agents act according to the fixed set of norms listed in Table 1. If the norms conflict, the agents toss a fair coin to choose between alternative actions. If Fixed agents perceive an action as a deviation, they sanction the deviant.

Sanctioning agents infer social norms from sanctions [Andrighetto *et al.*, 2013]. These agents start with the same strategy as Fixed agents. They continue to record the interaction history. Once they have gained enough number of records in their history of sanctions, they decide their subsequent actions based on history. In our simulation, this number is empirically selected so that an agent visits each scenario at least once. As callees, when norms conflict, they select the action that provides a higher payoff, computed according to Tables 2–4. As callers and neighbors, these agents sanction callees as per fixed norms listed in Table 1.

Poros agents infer social norms by revealing and reasoning about context. They start with the same strategy as Fixed agents following norms listed in Table 1. As callees, they reveal context, i.e., reveal the caller's relationship and the call's urgency to their neighbors, and reveal their place and neighbors' relationships to the caller. As neighbors or callers, they understand the callee's revealed context and decide what action they would have performed were they in that context, and sanction accordingly. Poros agents use Table 5's payoffs.

We employ a linear regression model over interaction history to choose actions based on sanctions by stakeholders.

Callee Action	Neighbor Expects	ER	H	L	M	P
Answer	Answer	1.00	0.67	1.00	1.00	0.67
Answer	Ignore	–1.00	–0.33	–1.00	–1.00	–0.33
Ignore	Answer	–1.00	–0.33	–1.00	–1.00	–0.33
Ignore	Ignore	1.00	0.67	1.00	1.00	0.67

Table 5: Payoff for a neighbor based on how callee acts and what the neighbor expects in the context revealed by callee.

Network Type	Agents	Circles		
		Family	Coworker	Friend
Large Dense	1,000	20	20	20
Large Sparse	1,000	100	100	100
Small Dense	250	5	5	5
Small Sparse	250	25	25	25

Table 6: Characteristics of network types studied.

5 Experiments and Results

We evaluate our research questions via multiple experiments on the ringer environment in which we simulate 1,000 or 250 Fixed, Sanctioning, and Poros agents in pragmatic, considerate, and selfish agent societies. The agents in societies use different schemes to aggregate payoffs. We run each simulation for 3,000 steps and compute the following metrics.

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 according to the payoff Tables 2, 3, 4, and 5.

To answer Q_1 on norms, we consider the following hypotheses pertaining to specified agent types. For brevity, we omit the corresponding null hypotheses indicating no gain. We test significance via the two-tailed paired t-test.

$\mathbf{H}_1$ Poros yields greater *social cohesion* than Fixed.

$\mathbf{H}_2$ Poros yields greater *social cohesion* than Sanctioning.

To answer Q_2 on goals, we consider these hypotheses:

$\mathbf{H}_3$ Poros yields greater *social experience* than Fixed.

$\mathbf{H}_4$ Poros yields greater *social experience* than Sanctioning.

5.1 Experiments with Pragmatic Agent Society and Varying Network Types

We simulate Fixed, Sanctioning, and Poros agents on four network types—large or small network with dense or sparse connectivity—as Table 6 describes. The society in this experiment is pragmatic in that the agents perceive social experience as the average payoff (equally weighted) for all stakeholders in an interaction. We summarize our results next.

Fixed agents. The average social experience was found to be between 0.53 and 0.56, and the social cohesion to be about 52% for the four network types.

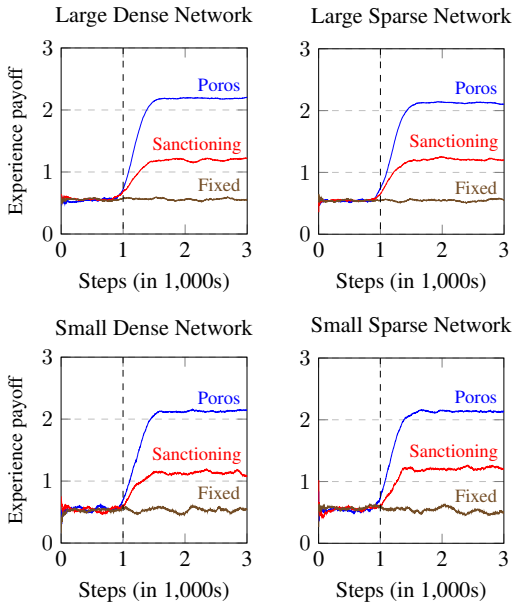

Figure 2: Social experience yielded by Poros, Sanctioning, and Fixed agents (per phone call for a window size of 200 steps) in pragmatic agent societies of different network sizes and densities.

Sanctioning agents. As expected, at around step 1,000 we see Sanctioning agents offer a rise in social experience over Fixed agents. The rise is gradual as the agents start to infer from history. For the first 1,000 steps, the average social experience is the same as Fixed agents. It later stabilizes between 1.11 and 1.21 for all four networks. The social cohesion values were between 61.2% and 63.7%.

Poros agents. At around step 1,000, as agents acquire confidence, we see a significant increase in social experience offered by Poros agents. It stabilizes between 2.14 and 2.19 for the different networks. Social cohesion was found be significantly higher between 82.0% and 83.2%. For the first 1,000 steps, Poros agents yield the same average social experience as Fixed and Sanctioning agents.

Social cohesion and experience offered by Poros agents are significantly greater than those offered by Fixed and Sanctioning agents; thus the null hypotheses corresponding to H_1, H_2, H_3, and H_4 are rejected. Figure 2 shows the social experience plots indicating the results are consistent across the four network types. Table 7 summarizes the findings of the experiment with pragmatic agents. It shows stabilized values for social experience and social cohesion, and p-values from the two-tailed paired t-tests.

5.2 Experiment with Considerate Agent Society

We experiment with a considerate agent society where agents give a larger weight to their neighbors' payoffs than to their own payoffs when computing social experience and deciding the actions to perform when norms conflict. These agents continue to sanction based on their history.

Figure 3 shows the social experience for considerate Sanctioning and Poros agents in a Small-Dense network. The average social experience drops for Sanctioning and Poros agents

	Agent Type	Experience	Cohesion	p
Large Dense	Fixed	0.56	52.7%	< 0.01
	Sanctioning	1.21	63.5%	< 0.01
	Poros	2.19	83.2%	–
Large Sparse	Fixed	0.55	52.5%	< 0.01
	Sanctioning	1.21	63.5%	< 0.01
	Poros	2.19	83.2%	–
Small Dense	Fixed	0.53	52.1%	< 0.01
	Sanctioning	1.11	61.2%	< 0.01
	Poros	2.14	82.0%	–
Small Sparse	Fixed	0.54	52.5%	< 0.01
	Sanctioning	1.22	63.7%	< 0.01
	Poros	2.14	82.1%	–

Table 7: Effectiveness of Poros in a pragmatic society.

	Agent Type	Experience	Cohesion	p
Considerate	Sanctioning	–0.33	41.3%	< 0.01
	Poros	–0.14	48.4%	–
Selfish	Sanctioning	1.22	63.5%	< 0.01
	Poros	2.13	82.0%	–

Table 8: Effectiveness of Poros in considerate and selfish societies.

after they have gained enough confidence. We attribute this decline to the fact that these agents value the neighbors' experience more than their own, and thus ignore calls they should have answered. Poros agents offer higher social cohesion and experience than Sanctioning agents because the secondary stakeholders give smaller negative sanctions when they reason about context. The results for the other three network types are similar. Table 8 summarizes these results.

5.3 Experiment with Selfish Agent Society

In a selfish agent society, agents give a very large weight to their own payoffs when computing social experience. Agents here may not always negatively sanction others who disturb them. As in other societies, agents in a selfish society sanction a deviant based on their history.

Figure 3 shows the social experience plot for selfish Sanctioning and Poros agents in a Small-Dense network. The plots resemble those in the experiment with pragmatic agents, but with slightly lower stabilized values. Here, agents tend to answer all calls, which benefits both caller and callee most of the time. We observe similar results for the other three networks. Table 8 summarizes these results.

5.4 Threats to Validity

We identified and mitigated two threats. The first concerns a differences in how users perceive experience. In reality, not all users perceive social experience the same way, and thus aggregating with only one scheme introduces the threat of difference in perceiving social experience. To mitigate this threat, we conduct experiments with three agent societies with different experience aggregation schemes. The second

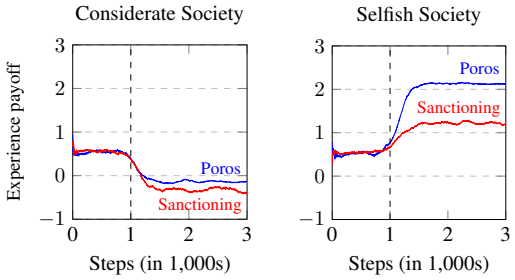

Figure 3: Social experience (averaged over a window size of 200 steps) yielded by Poros and Sanctioning agents in considerate and selfish agent societies simulated in a Small-Dense network.

threat concerns scalability. Since we simulate agent actions and interactions, a threat is whether our results scale to a large number of agents. To mitigate this threat, we evaluate Poros considering varying network sizes and types.

However, some threats remain. In particular, first, our results are based on simulation. Testing a SIPA's adaptability with end-users across contexts is challenging, as is reliably eliciting user attitudes and preferences.

Second, Poros agents always reveal context, which may pose a privacy threat. Ideally SIPAs should reveal context selectively. We leave this reasoning for future studies.

6 Conclusion and Future Directions

In Poros, SIPAs reveal and reason about context to understand the boundary of applicable norms and infer contextually relevant social norms. We find that Poros agents deliver significantly higher (1) social cohesion and (2) social experience than other agents. These findings are stable under changes to network size and characteristics of agents.

Being sensitive to norms, Poros SIPAs can naturally address challenges in engineering software tools for privacy. A SIPA would need data about its user's sharing preferences, privacy attitudes, and values and ethics [Ajmeri *et al.*, 2018] to make effective recommendations. A SIPA can learn its user's preferences and attitudes, but it would be helpful to bootstrap a SIPA via crowdsourced data about diverse user classes [Fogués *et al.*, 2017a; 2017b]. To better support privacy-respecting SIPAs, Poros could incorporate characteristics suggested by Such [2017] and adopt argumentation as in Kökciyan and Yolum's [2017] work when deciding the subset of context to reveal.

Other future directions are incorporating affect in relation to norms [Ferreira *et al.*, 2013] and supporting white lies to promote privacy (and social cohesion). For example, Bob may say his son is in hospital, instead of drug rehab. It would be instructive to study how such deception modulates effects on norms and goals.

Acknowledgments

We thank the US Department of Defense for support through the Science of Security Lablet and the Laboratory for Analytic Sciences at NC State University.

References

[Ajmeri *et al.*, 2016] Nirav Ajmeri, Jiaming Jiang, Rada Chirkova, Jon Doyle, and Munindar P. Singh. Coco: Runtime reasoning about conflicting commitments. *Proc. IJCAI*, pp. 17–23, New York, 2016.

[Ajmeri *et al.*, 2017] Nirav Ajmeri, Pradeep K. Murukannaiah, Hui Guo, and Munnindar P. Singh. Arnor: Modeling social intelligence via norms to engineer privacy-aware personal agents. *Proc. AAMAS*, pp. 230–238, 2017.

[Ajmeri *et al.*, 2018] Nirav Ajmeri, Hui Guo, Pradeep K. Murukannaiah, and Munindar P. Singh. Designing Ethical Personal Agents. *IEEE Internet Computing*, 22(2):16–22, 2018.

[Aldewereld *et al.*, 2016] Huib Aldewereld, Virginia Dignum, and Wamberto W. Vasconcelos. Group norms for multi-agent organisations. *ACM TAAS*, 11(2):15:1–15:31, 2016.

[Alechina *et al.*, 2012] Natasha Alechina, Mehdi Dastani, and Brian Logan. Programming norm-aware agents. In *Proc. AAMAS*, pp. 1057–1064, Valencia, 2012.

[Andrighetto *et al.*, 2013] Giulia Andrighetto, Jordi Brandts, Rosaria Conte, Jordi Sabater-Mir, Hector Solaz, and Daniel Villatoro. Punish and voice *PLoS ONE*, 8(6):1–8, 2013.

[Boella *et al.*, 2009] Guido Boella, Gabriella Pigozzi, and Leendert van der Torre. Normative framework for normative system change. In *Proc. AAMAS*, pp. 169–176, 2009.

[Dey, 2001] Anind K. Dey. Understanding and using context. *Personal and Ubiquitous Computing*, 5(1):4–7, 2001.

[Egas and Riedl, 2008] Martijn Egas and Arno Riedl. The economics of altruistic punishment and the maintenance of cooperation. *Proc. Royal Society of London B: Biological Sciences*, 275(1637):871–878, 2008.

[Ferreira *et al.*, 2013] Nuno Ferreira, Samuel Mascarenhas, Ana Paiva, Gennaro Di Tosto, Frank Dignum, John McBreen, Nick Degens, Gert Jan Hofstede, Giulia Andrighetto, and Rosaria Conte. An agent model for the appraisal of normative events based in in-group and outgroup relations. *Proc. AAAI*, pp. 1220–1226, 2013.

[Fogués *et al.*, 2017a] Ricard L. Fogués, Pradeep K. Murukannaiah, Jose M. Such, and Munindar P. Singh. Sharing policies in multiparty scenarios. *ACM TOCHI*, 24(1), article 5, 2017.

[Fogués *et al.*, 2017b] Ricard L. Fogués, Pradeep K. Murukannaiah, Jose M. Such, and Munindar P. Singh. SoSharP: Recommending sharing policies in multiuser privacy scenarios. *IEEE Internet Computing*, 21(6):28–36, 2017.

[Kafalı *et al.*, 2016] Özgur Kafalı, Nirav Ajmeri, and Munindar P. Singh. Revani: Revision and verification of normative specifications for privacy. *IEEE Intelligent Systems*, 31(5):8–15, 2016.

[Kökciyan and Yolum, 2017] Nadin Kökciyan and Pınar Yolum. Context-based reasoning on privacy in Internet of Things. In *Proc. IJCAI*, pp. 4738–4744, Melbourne, 2017.

[Luke *et al.*, 2005] Sean Luke, Claudio Cioffi-Revilla, Liviu Panait, Keith Sullivan, and Gabriel Balan. MASON: A multiagent simulation environment. *Simulation*, 81(7):517–527, 2005.

[Mashayekhi *et al.*, 2016] Mehdi Mashayekhi, Hongying Du, George F. List, and Munindar P. Singh. Silk: A simulation study of regulating open normative multiagent systems. In *Proc. IJCAI*, pp. 373–379, New York, 2016.

[Murukannaiah and Singh, 2012] Pradeep K. Murukannaiah and Munindar P. Singh. Platys Social: Relating shared places and private social circles. *IEEE Internet Computing*, 16(3):53–59, 2012.

[Murukannaiah and Singh, 2014] Pradeep K. Murukannaiah and Munindar P. Singh. Xipho: Extending Tropos to engineer context-aware personal agents. In *Proc. AAMAS*, pp. 309–316, Paris, 2014.

[Nardin *et al.*, 2016] Luis G. Nardin, Tina Balke-Visser, Nirav Ajmeri, Anup K. Kalia, Jaime S. Sichman, and Munindar P. Singh. Classifying sanctions and designing a conceptual sanctioning process model for socio-technical systems. *Knowledge Engineering Review*, 31:142–166, 2016.

[Nissenbaum, 2011] Helen Nissenbaum. A contextual approach to privacy online. *Dædalus*, 140(4):32–48, 2011.

[Noussair and Tucker, 2005] Charles Noussair and Steven Tucker. Combining monetary and social sanctions to promote cooperation. *Economic Inquiry*, 43(3):649–660, 2005.

[Savarimuthu *et al.*, 2009] Bastin Tony R. Savarimuthu, Stephen Cranefield, Martin K. Purvis, and Maryam A. Purvis. Norm emergence in agent societies formed by dynamically changing networks. *Web Intelligence and Agent Systems*, 7(3):223–232, 2009.

[Singh, 2013] Munindar P. Singh. Norms as a basis for governing sociotechnical systems. *ACM TIST*, 5(1), article 21, 2013.

[Such, 2017] Jose M. Such. Privacy and autonomous systems. *Proc. IJCAI*, pp. 4761–4767, Melbourne, 2017.

[Sugawara, 2011] Toshiharu Sugawara. Emergence and stability of social conventions in conflict situations. *Proc. IJCAI*, pp. 371–378, Barcelona, 2011.

[van Riemsdijk *et al.*, 2015] M. Birna van Riemsdijk, Louise Dennis, Michael Fisher, and Koen V. Hindriks. A semantic framework for socially adaptive agents: Towards strong norm compliance. *Proc. AAMAS*, pp. 423–432, 2015.

[Villatoro *et al.*, 2013] Daniel Villatoro, Jordi Sabater-Mir, and Sandeep Sen. Robust convention emergence in social networks through self-reinforcing structures dissolution. *ACM TAAS*, 8(1), article 2, 2013.

[Westin, 1967] Alan F. Westin. Privacy and Freedom. Atheneum. 1967.

[Yu *et al.*, 2013] Chao Yu, Minjie Zhang, Fenghui Ren, and Xudong Luo. Emergence of social norms through collective learning in networked agent societies. *Proc. AAMAS*, pp. 475–482, Saint Paul, 2013.