Ethics in Norm-Aware Agents

Nirav Ajmeri, Hui Guo, Pradeep Murukannaiah, Munindar P. Singh

Abstract

We address the problem of designing agents that navigate social norms by selecting ethically appropriate actions. Our framework, Yumbo, incorporates multi-criteria decision making to aggregate value preferences of users and select an ethically appropriate action. We find via a simulation seeded with a survey of user values and attitudes that Yumbo agents produce ethical actions that exhibit the Rawlsian property of fairness and yield a satisfactory social experience to their users.

1 Introduction

How can we develop intelligent agents that act ethically? Studies in ethics suggest that acting ethically requires an understanding of values and value preferences of the concerned individuals. Thus, an ethical agent would evaluate alternative actions in terms of how they promote or demote various values in different contexts, resolve conflicts, and navigate social norms [Cointe *et al.*, 2016]. We refer to such an agent as a socially intelligent personal agent (SIPA).

Ethicists subsume ethics in the theory of values [Friedman *et al.*, 2008]. Values are mostly universal across human societies [Schwartz, 2012; Rokeach, 1973]. Values for Schwartz are broad motivational goals, such as stimulation, achievement, security, and benevolence. Values for Rokeach may be *terminal* (security, freedom, happiness, and recognition, refer to defined-end states) or *instrumental* (modes of behavior or means to promote terminal values). Understanding values is crucial for a SIPA to deliver an ethical experience. Dechesne *et al.* [2013] observe that these ideals could conflict since they may not be preferred equally by each individual.

Social norms describe interactions between a subject and an object in terms of what they ought to be, or as reactions to behaviors, including attempts to apply sanctions. Representing and reasoning about social norms (in context) is essential in ethical decision making. That is, an ethical SIPA acts in compliance with contextually relevant social norms (but it may choose to break some norms intentionally, e.g., when the norms conflict).

We adopt Singh's [2013] representation of social norms, and consider two norm types for simplicity: commitment and prohibition. A commitment means its subject is committed to its object to bring about a consequent if an antecedent holds, and a prohibition means its subject is forbidden by its object to bring about a consequent if an antecedent holds. For instance, *Frank* (subject), a high school student, is *committed* (norm) to *Grace* (object), his mother, that he *will keep Grace updated about his location* (consequent) when he is *away from home* (antecedent).

Da Silva Figueiredo and Da Silva [2013] apply values to identify conflicts with norms, e.g., (1) a commitment's consequent demotes a value, or (2) a prohibition's consequent promotes a value. Dechesne *et al.* [2013] study compliance of norms based on values and to decide what norms to adopt. Kayal *et al.* [2014] present a model of norms and context centered on values, which could help a SIPA identify value preferences of its users. Work on collective ethical decision frameworks Yu *et al.* [2018] considers governance based on norms and economic principles but doesn't get into the rich notion of values that motivates this paper.

If a SIPA understands its users' value preferences and reasons about the values promoted or demoted by each of its actions, it could select ethically appropriate actions that provide a satisfactory social experience to its users. Accordingly, we identify the following research question:

RQ How can a SIPA select ethical actions by reasoning about value preferences of those concerned and how its actions promote or demote various values in a specific context?

Contribution. To address this research question, we develop Yumbo, a framework that enables ethical decision-making in light of users having distinct value preferences. Yumbo adapts a multi-criteria decision-making approach [Opricovic and Tzeng, 2004] to identify a consensus action. Yumbo addresses decision making by an individual agent but in a social context.

We evaluate Yumbo via simulations of agent societies grounded in data collected from an immersive survey wherein subjects select a location sharing policy for a given context.

We find that a Yumbo SIPA acts ethically by selecting fair actions [Rawls, 1985]—that maximize the minimum (i.e., worst-case) experience for each user involved in interactions with it, and yields a better overall social experience, i.e., higher mean experience for all users.

2 Motivating Example: Location Privacy

For concreteness, we consider mobile social applications where privacy is an important value [Taylor, 2002; Such, 2017; Kökciyan and Yolum, 2017], and present an example SIPA to demonstrate our ideas. Consider Pichu, a location sharing SIPA that enables its user to stay connected with friends and family by sharing location publicly, with common friends, with companions, with specific people, or with no one. In the common friends situation, the user accompanies someone, and revealing the user's location indirectly reveals the companion's location. Pichu produces a sharing policy based on preferences of the user and of any companions and contextual attributes, such as place and activity.

Example 1 (Olympiad). *Frank, a Pichu user, is a student in Ohio who values pleasure and social recognition. Also, he is committed (a norm) to his mother Grace that he will share his location with her when he is not at home. Sharing location promotes security but demotes privacy. Frank travels to Yale to participate in a Science Olympiad. Pichu shares publicly that Frank is at Yale participating in the Science Olympiad, and thus satisfies Frank's commitment to his mother, and promotes pleasure and social recognition for him.*

Example 2 (Times Square). *Frank visits New York and meets his uncle Harold in Times Square. Harold values privacy and prohibits (a norm) Frank from sharing his location publicly. Pichu prefers Harold's privacy over Frank's pleasure and social recognition, and shares only with Grace that Frank is at Times Square with Harold. Doing so satisfies Frank's commitment to Grace without violating Harold's prohibition.*

Note that Pichu is merely one application of Yumbo.

3 Yumbo: A Framework for Ethical Decisions

A SIPA should be aware of its users, their goals and actions to bring about the goals, which may vary with the context. A SIPA should choose and execute actions, especially when goals and social expectations conflict, based on its users' contextual preferences of the applicable social norms [Ajmeri *et al.*, 2017]. Users' preferences among values provide a basis for choosing which goals to bring about or which norms to satisfy. In Yumbo, a SIPA selects ethically appropriate actions by understanding its users' preferences across values.

A real-life society comprises humans, each of whom is the unique primary user of exactly one SIPA. A human has goals and values, is socially related to other humans, and enters into and exits from diverse contexts. A human's context is given by attributes such as its place, other humans present, and activities in which the human and others are engaged.

3.1 Representing Value Preferences

Figure 1 illustrates a Yumbo SIPA's representation and reasoning. A SIPA's *user model* describes a SIPA's users, and their goals and values. The SIPA maintains the relationships between its primary user and others. Besides the fixed primary user, a SIPA may have secondary users—humans who may be affected by the SIPA's actions.

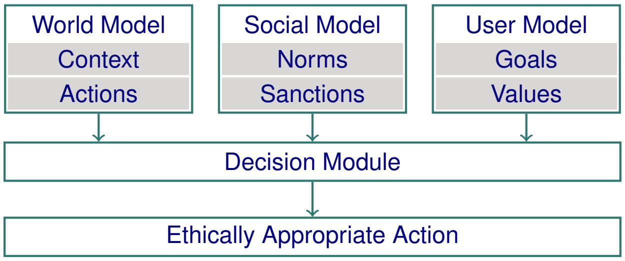

Figure 1: Representation and reasoning modules of a Yumbo SIPA.

A SIPA's *world model* describes the context in which a SIPA acts. A SIPA's *social model* specifies the norms governing a SIPA's interactions in a society and the associated sanctions. A SIPA's *decision module* is responsible for producing an *ethically appropriate action* that yields a (fair) social experience to the SIPA's users, especially in scenarios where either the norms conflict or the value preferences of users are not aligned.

In a society of Pichu SIPAs, the primary user is the one whose phone the SIPA runs on, and a secondary user is any companion of the primary user. The user goals designed in Pichu include *staying connected with friends and family*. To bring out a goal, a SIPA selects one of the following three actions: share with all, share with common friends, share only with companions. For example, when a user moves to a new place or meets new people, the SIPA may share the user's context to bring about the user's goal of *staying connected* and to promote user's value of *pleasure* or *safety*. Other values of relevance to Pichu include *privacy* and *recognition*.

The user's (and the SIPA's) context includes the user's current location in contextual terms—i.e., the place, companions, and activities. Each place is defined by attributes such as physical conditions (e.g., rainy), expected activities (e.g., hiking), social interactions (e.g., having a discussion), and temporal information (e.g., at late night). Places in Pichu include conference, hiking, restaurant, and so on. Relationships between the primary user and the companions include co-worker, family, and friend.

Each context includes a set of contextually relevant norms that govern the interaction of SIPAs in that context. For example, Frank's commitment to Grace may be relevant only when he is traveling.

When a SIPA's user moves between places, or when new people (also users) join a SIPA's user, the context changes. For instance, the context changes when Harold joins Frank in Times Square from when Frank is alone in Times Square.

When the context changes, a Pichu SIPA selects an action based on the new context, and its users' value preferences.

A SIPA's value preference is represented by a set of tuples $\{(v_j, v_k, c) \mid v_j, v_k \in V, c \in C\}$ where V is a set of values and C is a set of contexts such that the SIPA prefers value v_j over value v_k in context c.

Frank's preference for values of *pleasure* and *recognition* over *privacy* during Olympiad can be represented as $\{(\textit{pleasure}, \textit{privacy}, \textit{olympiad}), (\textit{recognition}, \textit{privacy}, \textit{olympiad})\}$. We assume that, within a context, the value preferences are mutually consistent for a SIPA user and that there are no cycles. Handling cyclic preferences could be a future direction.

In a decision-making episode, a SIPA determines (1) the context it is in through the sensors the SIPA is equipped with, (2) the future state of the world for each action it can perform, (3) the value preferences of its users, and (4) the social experience its users will derive for each action it can perform. Then, based on the applicable norms in a given context and its users goals, a SIPA identifies an action to perform.

3.2 Reasoning about Value Preferences

A SIPA's users in an interaction may have inconsistent preferences. Thus, a SIPA's actions based solely on one (e.g., primary) user's preference may conflict with its other users' preferences. For instance, in Example 2, if Frank's SIPA shares publicly that Frank and Harold are in Times Square considering only Frank's preference for *pleasure* and his commitment to Grace, that action conflicts with Harold's preference for *privacy* and violates Harold's prohibition.

How can we identify which actions to perform in situations where (1) the actions prescribed by the norms conflict with the actions that promote the values preferred by a SIPA's users, or (2) the users of a SIPA have different value preferences and thus prefer different actions?

Representing preferences over values as cardinal numbers facilitates aggregating them to choose an action with the highest gain. Sotala [2016] proposes using a reward function for a human's values, which an agent can learn and maximize.

We adopt the VIKOR method [Opricovic and Tzeng, 2004], a multi-criteria decision-making (MCDM) method. VIKOR's ranking is based on closeness to the ideal solution, and provides an ethically appropriate solution that yields high social utility as against high individual utility. Whereas VIKOR relies on numeric payoffs, humans tend not to use payoff tables but (preordered) discrete preferences. We map preferences to numeric payoffs by adopting techniques such as *cumulative voting*—distributing a fixed number of points to each value preference over each available alternative action, or *cardinal voting*—giving numeric payoff (ratings) on a fixed scale to each value for all available alternative actions [Pacuit, 2017].

1. Determine the best and worst payoffs, f_x^* and f_x^- for each value x over 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 [Krause, 1973]:

   $S_y = \sum_{x=1}^{n} w_x(f_x^* - f_{xy})/(f_x^* - f_x^-)$, where w_x is the weight for value x, which is subject to a user's context and preferences over values. In particular, $S_y = 0$ when $f_x^* = f_x^-$.

3. Compute the weighted and normalized Chebyshev distance [Cantrell, 2000]:

   $R_y = \max_x[w_x(f_x^* - f_{xy})/(f_x^* - f_x^-)]$, where w_x is the weight for value 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 trades off group and individual experience.

Table 1: Computing rankings for policy alternatives using VIKOR for context *Times Square* in Example 2. Bold is best.

Alternatives	Frank's Values				Harold's Values				S_y	R_y	Q_y
	Pleasure	Privacy	Recognition	Safety	Pleasure	Privacy	Recognition	Safety			
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**
w_x	1	1	1	1	1	3	1	1			
f_x^*	10	5	10	10	5	15	5	5			
f_x^-	0	5	0	0	5	0	5	5			

5. Rank alternative actions by the values S, R, and Q, in increasing order, to produce three ranked lists of actions.

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

   - If no unique best action is identified, propose all actions y within the threshold, i.e., $|Q_y - \min Q| < h$, as compromise solutions, where h reflects the user's risk taking attitude.

Table 1 demonstrates possible numeric payoffs of the values and the calculated ranking of three alternative actions (share with all, share with common friends, and share only with Grace) that Pichu can take when Frank is with Harold in Times Square, as in Example 2. Since Harold is highly cautious about his privacy, we give a higher weight to Harold's privacy (3) and lower but equal weights to other seven criteria including Harold's other values and Frank's values. We assume $k = 0.5$ in this case, and find the alternative y_3, *share only with Grace* as the best solution.

4 Seeding Simulated Societies with Real Data

We conducted a survey of privacy attitudes and preferences to help ground our simulated society with value preferences of real users. Our 58 subjects were university students; our study was approved by our university's Institutional Review Board (IRB); we obtained informed consent from each subject.

First, the subjects completed a privacy attitude survey [Schnorf *et al.*, 2014] including their level of comfort in sharing personal information on the Internet on a Likert scale of 1 (very comfortable) to 5 (very uncomfortable), and the extent sharing personal information causes (or could cause) them negative experience, again on a Likert scale of 1 (not at all) to 5 (to a very great extent).

We form three privacy attitude buckets—*casual* to represent privacy unconcerned people, *conscientious* to represent privacy careful people who take decisions on a case-to-case basis, or *cautious* to represent privacy concerned people. We sort subjects into these buckets based on their responses. Figure 2 shows the resulting distribution of privacy attitudes.

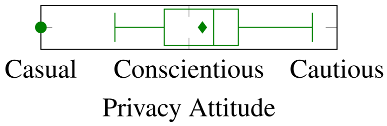

Figure 2: Distribution of privacy attitudes of subjects.

Next, the subjects completed two context-sharing surveys. In the first survey, they were given a list of contexts (Table 2), and their companions (alone, co-worker, family, friend, or crowd) in the given context, and were asked to select a sharing policy (share with all, share only with common friends, or share only with companions). In the second survey, subjects were additionally informed of the values that are promoted or demoted on sharing and on not sharing the context, and were asked to select a context-sharing policy accordingly. We use the first survey to engage and immerse the subjects in various contextual scenarios, and the second to help them make informed decisions according to the values promoted or demoted in each context.

We use the privacy attitudes of the subjects and the context-sharing policies selected by the subjects to create multiple artificial societies of users and to seed the simulation experiments described in Section 5.

Table 2: List of simulated places marked safe or 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

5 Experiments and Results

We evaluate our research question via two experiments in which we simulate societies of Pichu SIPAs who visit different places and may share their context. First, we experiment with a society of users with mixed privacy attitudes representing the subjects of our study described in Section 4. Second, we experiment with three societies with distinct dominant privacy attitudes. Our results are stable with respect to changes in the network size and the connectedness of a SIPA society.

5.1 Decision-Making Strategies

As Pichu SIPAs move between places and interact with each other, they make sharing decisions that affect their users. To evaluate our research question, we define four (Yumbo and three baseline) decision-making strategies.

S$_{\textbf{Yumbo}}$. Compute a context-sharing policy from users' value preferences using VIKOR.

S$_{\textbf{primary}}$. Produce a context-sharing policy based only on the primary user's value preferences—how location sharing works today in social networking websites.

S$_{\textbf{conservative}}$. Produce the least privacy violating, i.e., the most restrictive, context-sharing policy among the alternatives based on the users' value preferences. This strategy corresponds to selection based on the least negative consequence.

S$_{\textbf{majority}}$. Produce the most common context-sharing policy based on the users' value preferences. This strategy corresponds to majority voting [Fogués *et al.*, 2017].

5.2 Metrics

For each SIPA interaction, we compute these measures:

Mean social experience, the mean *utility* across the society based on context-sharing policy decisions. Higher is better.

Best individual experience, the maximum *utility* obtained by any user during a single interaction. Higher is better.

Worst individual experience, the minimum *utility* obtained by any user during a single interaction, which helps verify if a society supports Maximin [Rawls, 1985]. Higher is better.

Fairness, the reciprocal of the disparity between the best and the worst individual experience obtained by users during a single interaction [Rawls, 1985]. Higher is better.

Computing Experience. The utility that a SIPA obtains from a sharing policy in a certain context, whether to a primary or a secondary user, is a weighted sum of the numeric utility payoffs that the user perceives with respect to each of the values. We preset these numbers in a utility matrix such that they reflect a human-subject's preferences over the corresponding values. We assume that a user's utility is maximized when the chosen sharing policy is the user's most preferred, and the utility decreases linearly for a policy that deviates from it. Table 3 lists the preferred policies and utility numbers for each value of one human-subject in different contexts.

Table 3: Example numeric utility matrix for a user.

Place	Companion	Policy	Value			
			Pl	Pr	Re	Se
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

Pl, Pr, Re, Se = pleasure, privacy, recognition, security

5.3 Hypotheses

We evaluate the following hypotheses to answer our research question. Each hypothesis claims that Yumbo is superior to the baseline strategies with respect to the specified metric. We omit the corresponding null hypotheses for brevity.

H_{social}. Yumbo wins on mean social experience.

H_{best}. Yumbo wins on best individual experience.

$\mathbf{H_{worst}}$. Yumbo wins on worst individual experience.

$\mathbf{H_{fair}}$. Yumbo wins on fairness.

5.4 Experimental Setup

We adopt MASON [Luke *et al.*, 2005], a multiagent simulation toolkit, to develop a simulation environment containing a society of Pichu SIPAs.

We run simulations on a society of Pichu SIPAs. All parameters described below are set empirically based on our surveys. Specifically, we experiment on a society of 580 SIPAs, ten per study subject, each of which assumes the properties, preferred choices and privacy attitude of a human.

At each step, each SIPA is at one of the eight places listed in Table 2, and moves after one step to another place with equal probability. A SIPA decides a context-sharing policy based on the current place and the SIPA's users' privacy attitudes, value preferences, and decision-making strategy in Section 5.1.

For each setting, we run the simulation 2,000 steps three times and record the social experience each participating SIPA receives in each step. In the following figures, we plot the numbers in 100-step intervals for clarity.

5.5 Experiment with Mixed Agent Society

In the experiment with a mixed agent society, the SIPAs are mapped evenly to the subjects. Each pair of SIPAs relates as co-workers, friends, family (with equal probability), or strangers. Relationships are assigned at the beginning of the simulations such that they exhibit small world properties (degree: 10, rewiring probability: 0.05, edges: 3,445, clustering coefficient: 0.56, density: 0.014, average distance: 4.71) [Watts and Strogatz, 1998].

To evaluate H_{social}, we compare the *mean experience* yielded by SIPAs incorporating the four decision-making strategies—S_{Yumbo}, $S_{primary}$, $S_{conservative}$, and $S_{majority}$. Similarly, for H_{best}, H_{worst}, and H_{fair}, we compare the *best individual experience*, *worst individual experience*, and *fairness*, respectively, as yielded by these decision-making strategies.

Table 4 summarizes the results for a mixed agent society. It shows aggregated values for mean, best, and worst experience, fairness, and p-values from the two-tailed paired t-tests comparing the mean experience yielded by Yumbo and by other strategies. Figure 3 shows the mean experience plots.

Table 4: Metrics in a society with mixed privacy attitudes.

Strategy	Mean	Best	Worst	Fairness	*p*
S_{Yumbo}	**1.361**	1.715	**0.767**	**1.05**	–
$S_{primary}$	1.286	1.789	0.579	0.83	<0.01
$S_{conservative}$	1.106	1.721	0.472	0.80	<0.01
$S_{majority}$	1.339	**1.836**	0.570	0.78	<0.01

We observe that Yumbo yields better mean social experience than other decision-making strategies. Although the best individual experience obtained by Yumbo SIPA users is not the largest, they yield the highest worst individual experience and fairness. These results indicate that Yumbo yields solutions such that each companion is treated fairly, and thus Yumbo SIPAs act ethically. Thus, the null hypotheses corresponding to H_{social}, H_{best}, $H_{fairness}$ are rejected.

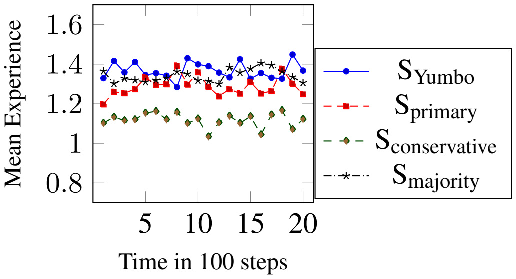

Figure 3: Yumbo vs. others: Mean experience in a mixed society.

5.6 Experiments with Majority Privacy Attitudes

To investigate the effects of societal distributions of privacy attitudes, we create three artificial societies respectively dominated by privacy casual, conscientious, and cautious users. Figure 4 shows the resulting distributions.

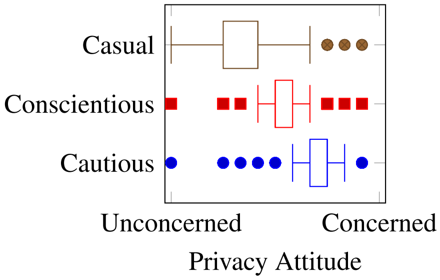

Figure 4: Privacy attitude distributions for artificial societies of cautious, conscientious, and casual users.

Table 5 summarizes the results and Figure 5 shows the mean social experience plots.

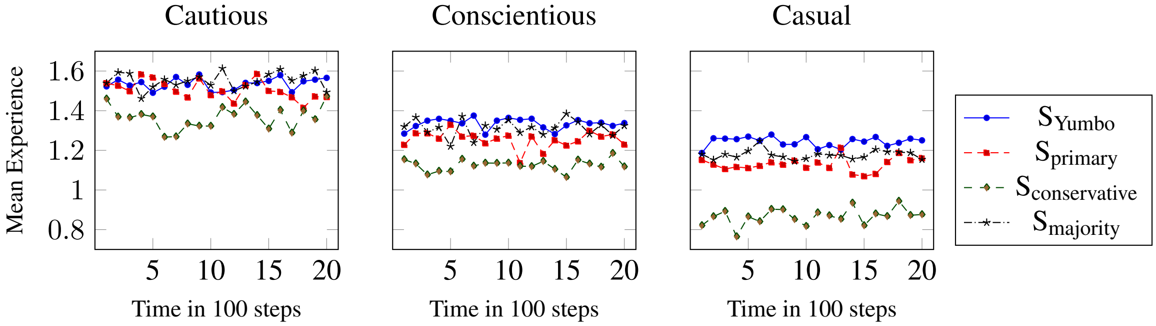

Figure 5: Yumbo vs. other strategies: Mean experience in societies that exhibit majorities in specified privacy attitudes.

Privacy Cautious Society. Yumbo yields the second-best mean experience, next only to the majority strategy. Yumbo yields the highest worst individual experience, i.e., the minimum utility that SIPA users obtain is higher compared to other decision making strategies, supporting the Maximin criterion. For fairness, Yumbo has the highest outcome. Thus, the null hypotheses related to H_{worst} and $H_{fairness}$ are rejected.

Privacy Conscientious Society. Yumbo yields the best mean experience and maximizes the worst individual experience while giving the fairest solutions. Hence, the null hypotheses related to H_{worst} and $H_{fairness}$

Table 5: Comparing mean social experience, best and worst individual experience, and fairness yielded by Yumbo SIPAs using VIKOR with other decision-making strategies in societies based on distinct majority privacy attitudes.

Strategy	Cautious				Conscientious				Casual			
	Mean	Best	Worst	Fairness	Mean	Best	Worst	Fairness	Mean	Best	Worst	Fairness
S_{Yumbo}	1.535	1.664	**1.233**	**2.27**	**1.329**	1.531	**0.867**	**1.51**	**1.242**	1.457	**0.768**	**1.45**
S_{primary}	1.506	1.766	1.082	1.46	1.253	1.592	0.679	1.10	1.129	1.466	0.584	1.13
$S_{\text{conservative}}$	1.366	1.745	1.059	1.46	1.093	1.519	0.608	1.10	0.870	1.338	0.454	1.34
S_{majority}	**1.551**	**1.858**	1.007	1.18	1.318	**1.699**	0.575	0.89	1.176	**1.534**	0.518	0.98

are rejected.

Privacy Casual Society. Yumbo yields the best mean experience while giving the fairest solutions with the highest worst individual experience; thus, the null hypotheses related to H_{worst} and H_{fairness} are rejected.

5.7 Threats to Validity and Mitigation

The first threat concerns simulation as an evaluation methodology. Simulation helps us conduct an evaluation that would be infeasible otherwise. Moreover, we ground our societies in data obtained from users.

Second, users may perceive social experience differently in reality than when completing a survey. To mitigate the threat of inaccuracies in self-reported attitudes, we employ immersive context sharing scenarios so they are prompted to think more naturally about sharing policies than otherwise.

6 Discussion

How to incorporate ethics into AI is a major modern research direction. Ethics inherently involves looking beyond one's self interest. That is, an agent must consider users in addition to its primary users and accommodate their value preferences in its decision making. Yumbo provides a method for doing so and demonstrates the gains in fairness accruing as a result.

Dignum [2017] argues for the need for AI reasoning to take into account societal values because autonomous AI systems increasingly affect our lives. Dignum proposes several approaches to responsibility in AI design considering human values. Serramia *et al.* [2018] show how to incorporate values with norms in a heuristic decision-making framework. Kayal *et al.* [2018] propose an automatic value-based model for resolving conflicts between norms, especially social commitments, in multiagent systems. Their user study suggests that values can be used to predict, users' preferences when resolving conflicts. Yumbo goes beyond these works by providing constructs and mechanisms to develop value-driven SIPAs.

Cranefield *et al.* [2017] describe a mechanism of value-based reasoning for BDI (Belief-Desire-Intention) agents. They argue that decision making, such as the selection of norms, is influenced by the value system, and therefore do not model norms. However, without norms, agents would need a complete understanding of human values to make morally correct decisions, which is difficult to realize.

Ajmeri *et al.* [2018] develop agents who apply norms to provide privacy assistance to their users. Their notion of privacy recognizes values such as confidentiality, disapprobation, and avoiding infringing into others' space. However, Ajmeri et al.'s [2018] agents seek to maximize the social experience of their respective users. Maximizing social experience may not translate to fairness as we observed in experiments with privacy cautious society where S_{majority} yields maximum social experience but least fairness. Yumbo's focus is to balance the needs of primary and secondary users.

Barry *et al.* [2017] propose a framework that adopts an Aristotelian virtue ethics concept, especially phronesis, which describes the practical wisdom of gathering experience in a context. Barry *et al.* claim that applications with phronesis learn contextual client knowledge, and therefore make the right choices that inherently involve ethical reflection. However, their design does not address conflicts between priorities, which are common in social settings.

Cranefield *et al.* [2016] show how agents can learn norms based on observations of behavior and sanction in a society, somewhat similar to Ajmeri *et al.* [2018]. How norms emerge in societies of ethical SIPAs is an important question, relating also to the challenge below.

An obvious challenge in fielding ethical agents is that they may be exploited by unethical agents. Partly, this is an unavoidable consequence of ethics. However, it suggests the need for both additional regulatory mechanisms, both social (such as sanctioning) and psychological (such as guilt). A comprehensive study of these topics in conjunction with ethics is a major future direction.

References

Nirav Ajmeri, Pradeep K. Murukannaiah, Hui Guo, and Munindar P. Singh. Arnor: Modeling social intelligence via norms to engineer privacy-aware personal agents. In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 230–238, São Paulo, May 2017. IFAAMAS.

Nirav Ajmeri, Hui Guo, Pradeep K. Murukannaiah, and Munindar P. Singh. Robust norm emergence by revealing and reasoning about context: Socially intelligent agents for enhancing privacy. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 28–34, Stockholm, July 2018. IJCAI.

Marguerite Barry, Kevin Doherty, Jose Marcano Belisario, Josip Car, Cecily Morrison, and Gavin Doherty. mHealth for maternal mental health: Everyday wisdom in ethical design. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI)*, pages 2708–2756, Denver, 2017. ACM.

Cyrus D. Cantrell. *Modern Mathematical Methods for Physicists and Engineers*. Cambridge University Press, Cambridge, 2000.

Nicolas Cointe, Grégory Bonnet, and Olivier Boissier. Ethical judgment of agents' behaviors in multi-agent systems. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 1106–1114. International Foundation for Autonomous Agents and Multiagent Systems, 2016.

Stephen Cranefield, Felipe Meneguzzi, Nir Oren, and Bastin Tony Roy Savarimuthu. A Bayesian approach to norm identification. In *Proceedings of the 22nd European Conference on Artificial Intelligence (ECAI)*, pages 622–629, Amsterdam, August 2016. IOS Press.

Stephen Cranefield, Michael Winikoff, Virginia Dignum, and Frank Dignum. No pizza for you: Value-based plan selection in BDI agents. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 178–184, Melbourne, 2017. IJCAI.

Karen Da Silva Figueiredo and Viviane Torres Da Silva. An algorithm to identify conflicts between norms and values. In *Proceedings of the 9th International Conference on Coordination, Organizations, Institutions, and Norms in Agent Systems (COIN)*, pages 259–274, St. Paul, MN, 2013. Springer.

Francien Dechesne, Gennaro Di Tosto, Virginia Dignum, and Frank Dignum. No smoking here: Values, norms and culture in multi-agent systems. *Artificial Intelligence and Law*, 21(1):79–107, Mar 2013.

Virginia Dignum. Responsible autonomy. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4698–4704, Melbourne, 2017. IJCAI.

Ricard López Fogués, Pradeep K. Murukannaiah, Jose M. Such, and Munindar P. Singh. Sharing policies in multiuser privacy scenarios: Incorporating context, preferences, and arguments in decision making. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(1):5:1–5:29, March 2017.

Batya Friedman, Peter H. Kahn Jr., and Alan Borning. Value sensitive design and information systems. In Kenneth Einar Himma and Herman T. Tavani, editors, *The Handbook of Information and Computer Ethics*, chapter 4, pages 69–101. John Wiley & Sons, Hoboken, New Jersey, 2008.

Alex Kayal, Willem-Paul Brinkman, Rianne Gouman, Mark A. Neerincx, and M. Birna van Riemsdijk. A value-centric model to ground norms and requirements for ePartners of children. In *Proceedings of the 9th Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems*, pages 329–345, Cham, 2014. Springer.

Alex Kayal, Willem-Paul Brinkman, Mark A. Neerincx, and M. Birna van Riemsdijk. Automatic resolution of normative conflicts in supportive technology based on user values. *ACM Transactions on Internet Technology (TOIT)*, 18(4):41:1–41:21, May 2018.

Nadin Kökciyan and Pınar Yolum. Context-based reasoning on privacy in Internet of Things. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4738–4744, Melbourne, 2017. IJCAI.

Eugene F. Krause. Taxicab geometry. *The Mathematics Teacher*, 66(8):695–706, 1973.

Sean Luke, Claudio Cioffi-Revilla, Liviu Panait, Keith Sullivan, and Gabriel Balan. MASON: A multiagent simulation environment. *Simulation: Transactions of the Society for Modeling and Simulation International*, 81(7):517–527, July 2005.

Serafim Opricovic and Gwo-Hshiung Tzeng. Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS. *European Journal of Operational Research*, 156(2):445–455, 2004.

Eric Pacuit. Voting methods. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2017 edition, 2017.

John Rawls. Justice as fairness: Political not metaphysical. *Philosophy and Public Affairs*, 14(3):223–251, 1985.

Milton Rokeach. *The Nature of Human Values*. Free Press, New York, 1973.

Sebastian Schnorf, Aaron Sedley, Martin Ortlieb, and Allison Woodruff. A comparison of six sample providers regarding online privacy benchmarks. In *In Proceedings of the SOUPS Workshop on Privacy Personas and Segmentation*, pages 1–7, Menlo Park, 2014.

Shalom H. Schwartz. An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2(1):11, 2012.

Marc Serramia, Maite López-Sánchez, Juan A. Rodríguez-Aguilar, Manel Rodríguez, Michael Wooldridge, Javier Morales, and Carlos Ansótegui. Moral values in norm decision making. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 1294–1302, Stockholm, July 2018. IFAAMAS.

Munindar P. Singh. Norms as a basis for governing sociotechnical systems. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):21:1–21:23, December 2013.

Kaj Sotala. Defining human values for value learners. In *Proceedings of the Workshops of the 30th AAAI Conference on Artificial Intelligence: AI, Ethics, and Society*, pages 113–123, Phoenix, 2016. AAAI Press.

Jose M. Such. Privacy and autonomous systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1–7, Melbourne, 2017. IJCAI.

James Stacey Taylor. Privacy and autonomy: A reappraisal. *The Southern Journal of Philosophy*, 40(4):587–604, 2002.

Duncan J. Watts and Steven H. Strogatz. Collective dynamics of ‘small-world’ networks. *Nature*, 393:440–442, June 1998.

Han Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, Victor R. Lesser, and Qiang Yang. Building Ethics into Artificial Intelligence. *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5527–5533, Stockholm, 2018.