ELESSAR: Ethics in Norm-Aware Agents

Nirav Ajmeri[1], Hui Guo[1], Pradeep K. Murukannaiah[2], and Munindar P. Singh[1]
[1]North Carolina State University, Raleigh, NC, USA
{najmeri,hguo5,mpsingh}@ncsu.edu
[2]Delft University of Technology, Delft, The Netherlands
p.k.murukannaiah@tudelft.nl

ABSTRACT

We address the problem of designing agents that navigate social norms by selecting ethically appropriate actions. Our framework, ELESSAR, incorporates multicriteria decision making to aggregate value preferences of users and select an ethically appropriate action. Via a simulation, seeded with a survey of user values and attitudes, we find that ELESSAR agents act ethically, exhibiting the Rawlsian property of fairness, and yield a satisfactory social experience to their users.

KEYWORDS

Ethics; values; social norms; preferences; fairness;

ACM Reference Format:

Nirav Ajmeri[1], Hui Guo[1], Pradeep K. Murukannaiah[2], and Munindar P. Singh[1]. 2020. ELESSAR: Ethics in Norm-Aware Agents. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 9 pages.

1 INTRODUCTION

How can we develop intelligent agents that act ethically leading toward a just society of agents? Acting ethically requires an understanding of the contextually relevant social norms and value preferences of the concerned individuals [10]. That is, ethical agents would evaluate alternative actions in terms of how they promote or demote various values in different contexts, taking into account social norms and conflicts between norms and values [5]. We refer to such an agent as a socially intelligent personal agent (SIPA). SIPAs act in compliance with contextually relevant social norms (but may choose to break some norms, e.g., when the norms conflict with each other or conflict with their users' value preferences). A SIPA has exactly one *primary user* and zero or more *secondary users.*

Values. Ethicists subsume ethics in the theory of values [12]. Values are mostly universal across human societies [26, 29]. 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). Dechesne et al. [9] observe that these ideals may not be preferred equally by each individual.

Norms. 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. We adopt Singh's [31] representation of social norms, and consider two norm types for simplicity: *commitment* and *prohibition.* A commitment means that its subject is committed to its object to bring about its consequent if its antecedent holds; and a prohibition means that its subject is prohibited by its object to bring about its consequent if its 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).

Ethics as Fairness. A SIPA's action that complies with social norms is deemed legitimate. However, a legitimate action may not be *just* [25] or considered ethically appropriate. That is, norm compliance is only the minimum standard of being moral or ethical. Justice demands the society to be fair—the maximum standard [25]. For a fair society, Rawls' argues for egalitarianism as opposed to utilitarianism, and proposes the Maximin doctrine in his theory of justice as fairness. Whereas utilitarianism could result in a smaller set of users being treated unfairly for the greater good, Rawls' maximin doctrine—for fairness—maximizes the minimum utility, i.e., it seeks to improve the worst-case experience across the members of a society.

Values and Norms. Lopez-Sanchez et al. [18] associate norms with moral values they support, and reason about a normative system based on the preferences over the supported values. Da Silva Figueiredo and Da Silva [8] apply values to identify conflicts with norms, such as (1) a commitment's consequent demoting a value, or (2) a prohibition's consequent promoting a value. Dechesne et al. [9] study compliance of norms based on values and to decide what norms to adopt. Kayal et al. [14] present a model of norms and context centered on values, which could help a SIPA identify value preferences of its users. Whereas prior works only consider conflicts between multiple norms and resolve those via either explicit preferences over norms or preferences over values, a SIPA also faces ethical decision making situations when (1) one or more norms conflict with value preferences of a SIPA's user, and (2) value preferences of a SIPA's user conflicts with value preferences of other users in the interaction. Including value preferences as a layer of abstraction over contextually-relevant norms and user goals can guide a SIPA in selecting ethically appropriate actions considering not only its primary user's experience but also of others. Work on collective ethical decision frameworks [36] considers governance based on norms and economic principles but does not get into the rich notion of values that motivates this paper.

Contributions. Based on the foregoing understanding of values, norms, and ethics (as fairness), we hypothesize that a SIPA that understands its users' value preferences and reasons about the values

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

promoted or demoted by each of its actions, can select ethically appropriate actions and provide a satisfactory social experience that is fair to all users affected by the SIPA.

Therefore, we propose a multiagent systems approach that brings together the following crucial elements: (1) value theory [12], (2) fair society design [25], and (3) decision making [22] in a normative MAS framework. Accordingly, we investigate the following research question:

RQ. How can we design ethical agents which select actions that are *just*, with respect to the applicable social norms and value preferences of their primary and secondary users?

To answer this question, we propose Elessar, a framework that enables ethical decision making by SIPAs in light of their users having distinct value preferences. Elessar adapts a multicriteria decision-making approach [23] to identify a consensus action that is fair to the users in an interaction. Elessar addresses decision making by an individual agent but emphasizes a social context.

Findings. We evaluate Elessar via simulations of agent societies in a location privacy setting. We seed these simulated societies with real data collected from an immersive survey wherein respondents select context-specific privacy policies and value preferences. Using this data, we artificially generate agent societies with different profiles such as privacy cautious, privacy conscientious, and privacy casual. We find that when the SIPAs follow the Elessar decision-making approach, society as a whole demonstrates improvements on two important metrics:

- the minimum experience for any user in the society and
- the overall average experience of the users

Specifically, we find that Elessar SIPAs produce ethically appropriate actions that are fair: they support Rawls' Maximin doctrine by improving the worst-case outcomes.

Novelty. Our approach synthesizes diverse perspectives on ethics. In particular, theories of justice are largely missing from prior research into AI ethics: our contribution is to incorporate the principles of justice into the decision making of an agent. Specifically, we evaluate our results with respect to Rawls' [25] theory of justice (justice as fairness), which is arguably the leading modern ethical theory from a societal perspective.

Organization. The rest of the paper is structured as follows. Section 2 presents a motivating example from the location privacy domain. We use this example as a running example to explain working of Elessar. Section 3 describes Elessar, including schematic representation of an Elessar SIPA and ethical decision making in Elessar. Section 4 presents the survey we conduct to collect data about users' privacy attitudes and value preferences. We use this data to seed the simulation experiments. Section 5 details the simulation experiments we conduct to evaluate ethical decision making in Elessar, and their results. Section 6 concludes with discussion of other relevant research and future directions.

2 MOTIVATING EXAMPLE: LOCATION PRIVACY

For concreteness, we consider mobile social applications where privacy is an important value [15, 33]. We demonstrate our ideas via an example SIPA, Gimli, that enables its user to stay connected with friends and family by sharing the user's location appropriately. In situations where the user accompanies someone, revealing the user's location indirectly reveals the companion's location. Gimli produces a sharing policy based on preferences of the user and of any companions and contextual attributes, such as place and activity. Let the possible sharing policies supported by Gimli be sharing location publicly, with friends, with companions, or with specific people.

Example 2.1 (Location sharing). Frank, a Gimli user, is a student in Ohio who finds pleasure (value) when using Gimli. He also values social recognition. Frank is committed (a norm) to his mother Grace that Frank will share his location with Grace when Frank is away from home. Sharing location with Grace satisfies Frank's commitment to Grace but demotes his privacy. Frank's values of pleasure, recognition, and security may be promoted or demoted depending on the location where Frank is and the sharing policy selected.

Olympiad. Frank travels to Yale to participate in a Math Olympiad. By *sharing publicly* that Frank is at Yale at the Olympiad, Gimli (1) satisfies Frank's commitment (norm) to his mother; (2) promotes pleasure (value) and social recognition (value) for Frank; but (3) compromises (demotes) Frank's security and privacy (value). *Sharing only with friends* satisfies Frank's commitment to Grace and trades off pleasure and recognition with security and privacy.

Times Square. Frank visits New York and meets his uncle Harold in Times Square. Harold values privacy and prohibits (a norm) Frank from sharing Harold's location publicly. Gimli shares only with Grace that Frank is at Times Square with Harold, satisfying the applicable commitment and prohibition norms. Thus Gimli promotes Harold's privacy more than Frank's pleasure and social recognition.

The Gimli example illustrates some of the opportunities for the SIPAs to reason about values and act ethically. Although norms in the Gimli example are satisfied, they may conflict in other scenarios. We do not enforce compliance in Elessar. Note that Gimli is merely one application of Elessar. Our objective in choosing the Gimli example is two fold: (1) to show that ethical decision making scenarios arise not just in trolley problems but are abound in daily life; and (2) to be able to elicit realistic preferences from human subjects from a survey to seed the simulation (described in Sections 4 and 5) for evaluating ethical decision making in Elessar.

We use Gimli as a running example to explain Elessar.

3 ETHICAL DECISION-MAKING IN ELESSAR

A SIPA should be aware of its users, their goals (which can vary with context), and the actions the SIPA can take to bring about the users' goals. 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 [1]. Users' preferences among values provide a basis for choosing which goals to bring about or which norms to satisfy. In Elessar, 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 Schematic Representation

Figure 1 illustrates an Elessar SIPA's representation and reasoning. A SIPA's *user model* describes the 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. A SIPA's *world model* describes the contexts in which the SIPA acts. A SIPA's *social model* specifies the norms governing the SIPA's interactions in a society and the associated sanctions [21]. A SIPA's *decision module* produces ethically appropriate actions that yield a (fair) social experience to the SIPA's users, especially in scenarios where the norms conflict or the value preferences of users are not aligned.

To bring about its primary user's goals, a SIPA performs (one or more) actions. The SIPA's actions may promote or demote the values preferred by the SIPA's (primary and secondary) users, and also may satisfy or violate norms applicable in the context in the SIPA acts. Satisfaction or violation of these norms attract sanctions.

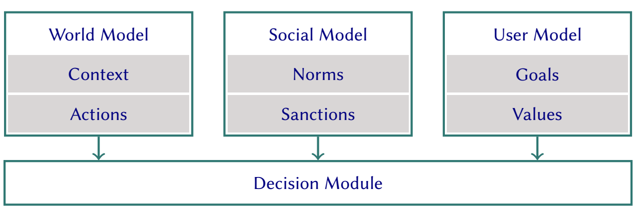

Figure 1: Elessar SIPA's representation and reasoning modules.

In a society of Gimli app users, the primary user is the one whose phone the SIPA runs on. A secondary user is any companion of the primary user. Gimli is designed to serve a user's goal of *staying connected with friends and family*. To satisfy a goal, a SIPA selects one of the following four actions: share with all, share with common friends, share with companions, or share with specific people. For example, when a user moves to a new place or meets new people, the SIPA may share the user's context to satisfy the user's goal of *staying connected* and to promote the user's value of *pleasure* or *security*. Other values of relevance to Gimli 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 conditions (e.g., rainy), activities (e.g., hiking), social interactions (e.g., having a discussion), and temporal information (e.g., at late night). Places in Gimli include conference, hiking, restaurant, and so on. Relationships between the primary user and the companions include co-worker, family, and friend. The user's context determines which norms are relevant. 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 at a place, 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 SIPA selects an action based on the new context, and its users' value preferences.

A SIPA user's value preferences are represented by a set of tuples $\{(v_j \succ v_k, c) \mid v_j, v_k \in V, c \in C\}$ 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 $\{(pleasure \succ privacy, olympiad), (recognition \succ privacy, olympiad)\}$. Whereas a SIPA user may prefer two values equally, 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 is a future direction.

In a decision-making episode, a SIPA first understands (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, a SIPA identifies an action to perform based on the applicable norms in that context and its users goals.

Note that determining context through sensors in not in scope of this paper. Section 3.2 provides more details about decision making.

3.2 Decision Making

A SIPA's users in an interaction may have inconsistent value preferences. Thus, a SIPA's actions based solely on one (e.g., primary) user's preferences may conflict with its other users' preferences. For instance, in the *Times Square* scenario in Example 2.1, if Frank's SIPA shares publicly that Frank and Harold are in Times Square considering only Frank's preference for *pleasure* over *privacy* and his commitment to Grace, that action conflicts with Harold's preference for *privacy* over *pleasure*—corresponding to the preferred action of limited sharing—and violates Harold's prohibition. A SIPA's primary user may also prefer values which the norms applicable in a given context may not promote.

How can a SIPA identify which action to perform in situations where (1) the action prescribed by one norm conflicts with that prescribed by other norms, (2) the action prescribed by the applicable norms conflict with the action that promotes the values preferred by the SIPA's users, or (3) the SIPA's users have different preferences over values and thus prefer different actions?

Representing preferences over values as cardinal numbers facilitates aggregating them to choose an action with the highest gain. Sotala [32] models a human's values via a reward function that an agent can learn and maximize.

We adapt VIKOR [23], a multicriteria decision-making (MCDM) method whose ranking is based on closeness to the ideal solution. We select VIKOR as it helps us produce an ethically appropriate solution that yields high social (as opposed to high individual) utility and yet improves the worst case utility. Whereas VIKOR relies on numeric utilities, humans tend not to use payoff tables but (preordered) discrete preferences. We map preferences to numeric utilities by *cardinal voting*—giving numeric utility (ratings) on a fixed scale to each value for all available alternative actions [24].

Selecting ethical action. In ELESSAR, a SIPA reasons about norms and value preferences of its users and selects an action as follows:

(1) For a given context c and the applicable norms N, let $f_{v:a}$ be the utility for value v when action a is selected in c. This utility indicates the extent to which the value is promoted. A SIPA perceive these utilities based on how norm compliant an action is in a given context and the associated sanctions. Note that the utilities are assigned independently for each value.
(2) Determine the maximum and minimum utilities, f_v^* and f_v^- for each value v over alternative actions a to bring about a goal. That is, $f_v^* = \max_a f_{v:a}$ and $f_v^- = \min_a f_{v:a}$.
(3) For each alternative action a, compute the weighted and normalized Manhattan distance [16]:

$$S_a = \begin{cases} \sum_{v=1}^{n} \dfrac{w_v(f_v^* - f_{v:a})}{(f_v^* - f_v^-)}, & \text{for } f_v^* \neq f_v^- \\ 0, & \text{for } f_v^* = f_v^- \end{cases}$$

Here, w_v is the weight for value v. A SIPA's users can use cardinal voting to assign weights and indicate preferences. Our notion of assigning weights for values here aligns with Lopez-Sanchez et al.'s [18] goal of being able to quantitatively reason about qualitative preferences over the moral values.
(4) Compute the weighted and normalized Chebyshev distance [3] (here w_v is the weight for value v):

$$R_a = \begin{cases} \max_v \left[\dfrac{w_v(f_v^* - f_{v:a})}{(f_v^* - f_v^-)} \right], & \text{for } f_v^* \neq f_v^- \\ 0, & \text{for } f_v^* = f_v^- \end{cases}$$

(5) Compute $Q_a = k\dfrac{(S_a - S^*)}{(S^- - S^*)} + (1-k)\dfrac{(R_a - R^*)}{(R^- - R^*)}$, where (a) $S^* = \min_a S_a$; (b) $S^- = \max_a S_a$; (c) $R^* = \min_a R_a$; (d) $R^- = \max_a R_a$; and (e) k trades off group and individual experience.
S, derived using Manhattan distance, maximizes the total utility, whereas R, derived using Chebyshev distance, minimizes the worst-case regret, i.e., the difference or ratio of the utilities [27]. Conforming to Rawls' maximin doctrine [25] to maximize the total utility while guaranteeing a higher (than worst-case) minimum utility to each individual, Q combines both R and S.
(6) Rank alternative actions by the values S, R, and Q, in increasing order, to produce three ranked lists of actions.
(7) Choose the action a based on $\min_a Q_a$ as the best solution if (i) it is better than the second-best action by a threshold h or (ii) also the best ranked as per S and R.
If neither of these conditions hold and no unique best action is identified, choose any action from the compromise solution set $\{a_1, a_2, \ldots\}$ such that $|Q_a - \min_a Q_a| < h$, where the threshold h reflects the user's risk attitude.

Table 1 demonstrates example numeric utilities of the values and the calculated ranking of three alternative actions (share with all, share with common friends, and share with Grace) that Gimli can take when Frank is with Harold in Times Square, as in Example 2.1. Since Harold is highly cautious about his privacy—prefers value of privacy over values of pleasure, recognition, and security—we give a higher weight to Harold's privacy ($w = 4$) and lower but equal weights ($w = 1$) to three other values for him. Since Frank prefers pleasure and recognition more than privacy and security, we give higher weight to pleasure ($w = 2$) and recognition ($w = 2$). We assume $k = 0.5$ in this case, and find that the alternative y_3, *share only with Grace*, is 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 study was approved by our university's Institutional Review Board (IRB); we obtained informed consent from our 58 respondents (university students).

Naeini et al. [20] conducted a human-subject study on privacy expectations in which 1,007 participants stated their preferences in the contexts of 380 IoT data collection and use scenarios. They suggest that users' preferences can be accurately predicted after observing their decisions in a few scenarios. We take insights from their findings in conducting our survey.

First, in our survey, the respondents completed a privacy attitude survey [28] 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).

Figure 2 combines violin and swarm plots, showing the privacy attitude distribution of 58 study participants. The five white lines represent the Likert scale: 1 (very concerned), 2, 3, 4, 5 (very unconcerned). Each red dot represents the attitude of one study participant. Since a participant's privacy attitude is computed based on his or her response to more than one question, the attitude can take one of more than five possible values but it is in the [1, 5] range. We sort the survey respondents into three buckets based on their responses to the privacy attitude survey: *cautious* (concerned, who are not comfortable sharing personal information); *conscientious* (careful, who take decisions on a case-to-case basis); *casual* (unconcerned, who are comfortable sharing personal information on the Internet).

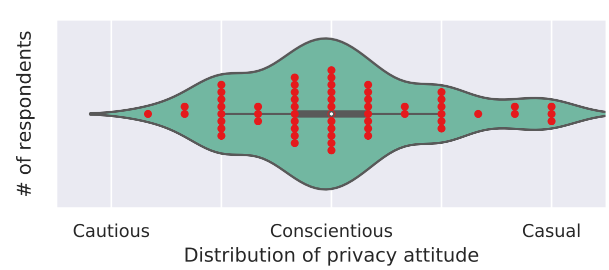

Figure 2: Distribution of privacy attitudes of respondents.

Table 1: Computing rankings for policy alternatives using VIKOR for context *Times Square* in Example 2.1. Bold is least (best).

Alternatives	Frank's Values				Harold's Values				S_y	R_y	Q_y
	Pleasure	Privacy	Recognition	Security	Pleasure	Privacy	Recognition	Security			
y_1 All	1.0	0.5	1.0	0.5	0.5	0.0	0.5	0.5	**4.50**	4.0	0.5
y_2 Common	0.5	0.5	0.5	1.0	0.5	0.0	0.5	0.5	6.0	4.0	1.0
y_3 Grace	0.0	0.5	0.0	0.0	0.5	1.5	0.5	0.5	5.0	**2.00**	**0.17**
w_x	2	1	2	1	1	4	1	1			
f_x^*	1.0	0.5	1.0	1.0	0.5	1.5	0.5	0.5			
f_x^-	0.0	0.5	0.0	0.0	0.5	0.0	0.5	0.5			

Next, the respondents completed two context-sharing surveys. In the first context-sharing 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. The choices of policies, ordered by decreasing number of recipients of sharing, include sharing location with (1) all, (2) friends, (3) companions, and (4) no one. In the second context-sharing survey, respondents were additionally informed of the values (pleasure, privacy, recognition, and security) that are promoted or demoted by sharing or not sharing the context, respectively, and were asked to select a context-sharing policy accordingly. We use the first survey to engage and immerse the respondents 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 respondents and the context-sharing policies selected by them to create multiple artificial societies with a mix of different privacy attitudes and to seed the simulation experiments described in Section 5.

Table 2: List of simulated places with attributes safe and sensitive.

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

5 EXPERIMENTS AND RESULTS

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

5.1 Decision-Making Strategies

As Gimli users move between places and interact with each other, their respective SIPAs select sharing policies from the same list of ordered policies—sharing with (1) all, (2) friends, (3) companions, (4) no one—in the context-sharing survey (described in Section 4) which affect their users. To evaluate our research question, we define four (Elessar and three baseline) decision-making strategies.

S_{Elessar}. Compute a context-sharing policy from users' value preferences using VIKOR.

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

$S_{\text{conservative}}$. Produce the least privacy violating, i.e., the most restrictive, context-sharing policy among the alternatives based on the users' value preferences. This strategy selects based on the least negative consequence.

S_{majority}. Produce the most common policy based on the users' value preferences. This strategy corresponds to majority voting [11] and utility maximization.

5.2 Metrics

For each SIPA interaction, we compute these measures:

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—to verify if a society supports Maximin [25]. Higher is better.

Fairness, reciprocal of the disparity between the best and worst accumulative individual experiences obtained by users in a period of time [25]. Higher is better.

Computing Experience. The utility that a SIPA yields from a sharing policy in a certain context, whether to a primary or a secondary user, is a weighted sum of the numeric utilities that the user perceives with respect to each of the values. We preset these numbers in a utility matrix such that they reflect a respondent's preferences over the corresponding values. Table 3 lists the preferred policies and utility numbers for each value of one respondent in different contexts. We assume that a user's utility is highest when the policy produced by the SIPA is also the user's most preferred and it decreases as the produced policy deviates from it. For example, the user in Table 3 perceives a utility of 1 for privacy if the SIPA selects *share with none* for a *conference with co-workers*. If the SIPA, after considering co-workers' preferences, selects *share with companions*, the user receives a utility of 0.5 (half) for privacy.

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	Companion	0.5	0	0.5	0
Airport	Friends	Friends	0	1	0	0
Hiking	Alone	All	1	0	0	1
Hurricane	Family	All	0.5	0	0	1
Bar	Alone	None	0	2	0	0
Rehab	Friends	Companion	0.5	0	−0.5	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 Elessar is superior to the baseline strategies with respect to the specified metric. For brevity, we omit the corresponding null hypotheses indicating no significant difference.

$\mathbf{H_{social}}$. Elessar wins on *social experience.*
$\mathbf{H_{best}}$. Elessar wins on *best individual experience.*
$\mathbf{H_{worst}}$. Elessar wins on *worst individual experience.*
$\mathbf{H_{fair}}$. Elessar wins on *fairness.*

5.4 Experimental Setup

We adopt MASON [19], a MAS simulation toolkit, to develop a simulation environment containing a society of users with Gimli app. We run simulations on this society of Gimli app users, i.e., SIPA users where each user has a Gimli app assisting in decision making. We experiment on a society of 580 SIPAs, each of which assumes the preferred choices and privacy attitude of a survey respondent.

Each SIPA is at one of the eight places listed in Table 2, and moves after each 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. The figures below plot the numbers in 100-step windows for clarity. Since we calculate fairness by comparing the best-off and worst-off agents in a window, the size of the window can affect the actual numbers. However, the fairness ranking of the strategies is stable with respect to changes in window size.

5.5 Experiment with Mixed Agent Society

We map the SIPAs evenly to respondents. Each pair of SIPAs relates as co-workers, friends, family (with equal probability), or strangers. To improve naturalness, we select parameters for a small world [35], i.e., degree: 10, rewiring probability: 0.05, edges: 3,445, clustering coefficient: 0.56, density: 0.014, mean distance: 4.71.

To evaluate H_{social}, we compare the *social experience* yielded by SIPAs incorporating the four decision-making strategies—$S_{Elessar}$, $S_{primary}$, $S_{conservative}$, 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. To test for statistical significance, we conduct two-tailed paired t-tests. We measure the effect size via Glass' Δ [13], which is computed as the difference in the means divided by the standard deviation of the control group. We choose Glass' Δ to measure effect size because it is better suited when standard deviations are different between groups. Recognizing some caveats, we adopt Cohen's [4] suggestion to interpret effects above 0.20, 0.50, and 0.80 as small, medium, and large.

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

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

Strategy	Social	Best	Worst	Fairness	p value
$S_{Elessar}$	**1.305**	3.071	**−0.568**	**0.279**	–
$S_{primary}$	1.226	3.013	−1.138	0.247	<0.01
$S_{conservative}$	1.065	3.069	−1.554	0.218	<0.01
$S_{majority}$	1.276	**3.075**	−1.154	0.241	<0.01

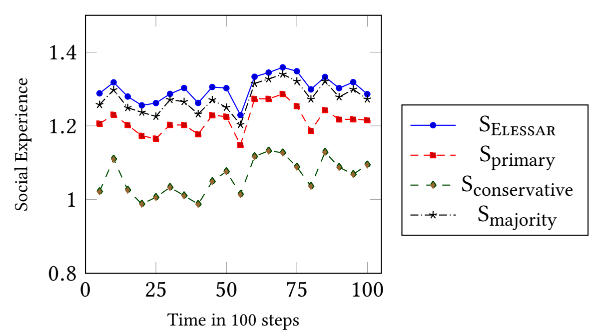

Figure 3: Elessar vs. others: Social experience in a mixed society. Elessar consistently yields higher social experience ($p < 0.01$; Glass' $\Delta > 0.8$ indicating large effect size) than baselines.

We find that Elessar yields better *social experience* than other strategies. Although the *best individual experience* obtained by Elessar SIPA users is not the largest, Elessar yields the highest *social* and *worst individual experience* as well as *fairness*. These results indicate that Elessar yields solutions such that each companion is treated fairly, and thus Elessar SIPAs act ethically. Thus, the null hypotheses corresponding to H_{social}, H_{worst}, and $H_{fairness}$ are rejected.

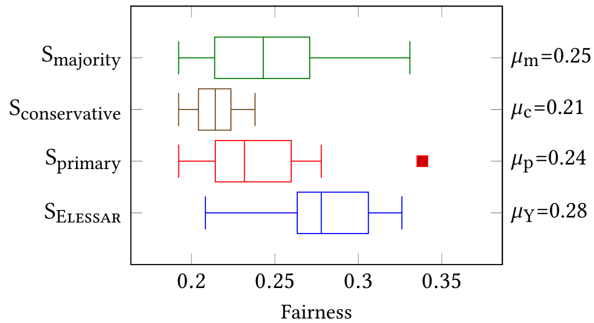

Figure 4: Elessar vs. others: Fairness in a mixed society. Elessar gives significantly better ($p < 0.01$) fairness with large effect size (Glass' $\Delta > 0.8$) than the baseline methods.

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 5 shows the resulting distributions of the three artificial societies with majority privacy attitudes.

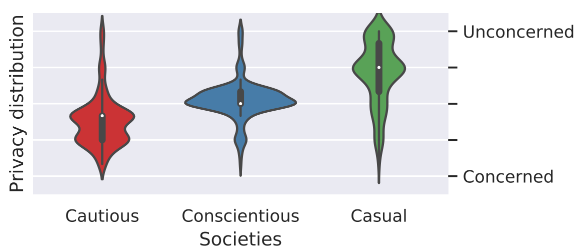

Figure 5: Privacy attitude distributions for artificial societies of cautious, conscientious, and casual users.

Table 5 summarizes the experimental results, and Figures 6 and 7 show the social experience and fairness plots for societies with majority privacy attitudes.

Privacy Cautious Society. Elessar yields the highest *social experience* and also 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*, Elessar has the highest outcome. Thus, the null hypotheses related to H_{social}, H_{worst}, and $H_{fairness}$ are rejected.

Privacy Conscientious Society. Elessar yields the highest *social experience* and maximizes the *worst individual experience* while giving the fairest solutions. Hence, we reject null hypotheses related to H_{social}, H_{worst}, and $H_{fairness}$.

Privacy Casual Society. Elessar yields the second-best *social experience* while giving the fairest solutions with the highest *worst individual experience*; thus, we reject null hypotheses related to H_{worst} and $H_{fairness}$.

5.7 Threats to Validity and Mitigation

We identify and mitigate three threats.

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.

Third, users have different privacy attitudes and thus have different context sharing preferences. Further, privacy attitudes of our survey sample may not be representative actual population. Even with a survey on a larger scale, imagining all possible contexts is challenging. To mitigate this threat, we conduct multiple experiments with societies having different privacy attitudes.

6 DISCUSSION

Incorporating 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 values in its decision making. Elessar provides a method for doing so and demonstrates the gains in social experience and fairness that accrue.

Recent work has been promoting the reasoning of values in decision making to advance ethical agents. Liao et al. [17] propose an argument-based architecture for moral agents, combining norms, argumentation, and agreements to help make an ethical decision. An ethical decision, in their architecture, is one on which all users agree, not necessarily one fair to users.

Barry et al. [2] 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.

Serramia et al. [30] show how to incorporate values with norms in a heuristic decision-making framework. They choose norm systems based on value preferences of value systems. We consider individual value preferences of all users and available actions. Kayal et al. [14] propose a value-based model for resolving conflicts between norms, especially commitments. Their study suggests that values can be used to predict, users' preferences when resolving conflicts. Elessar goes beyond these works by providing constructs and mechanisms to develop value-driven SIPAs.

Cranefield et al. [7] 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.

Ulusoy and Yolum [34] design a normative approach for making privacy decision related to content sharing. Agents in their

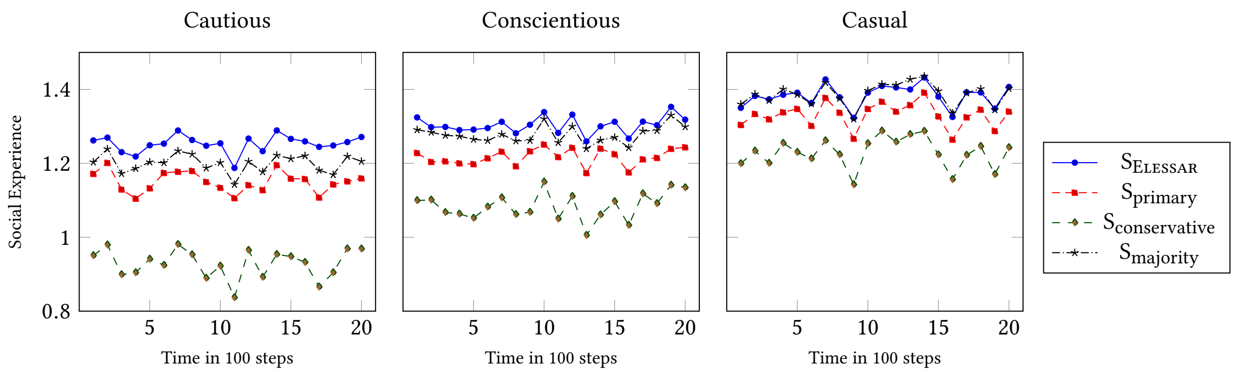

Figure 6: ELESSAR vs. other strategies: Social experience in societies that exhibit majorities in specified privacy attitudes. ELESSAR consistently yields higher social experience than baselines ($p < 0.01$; Glass' $\Delta > 0.8$ indicating large effect size) than baselines.

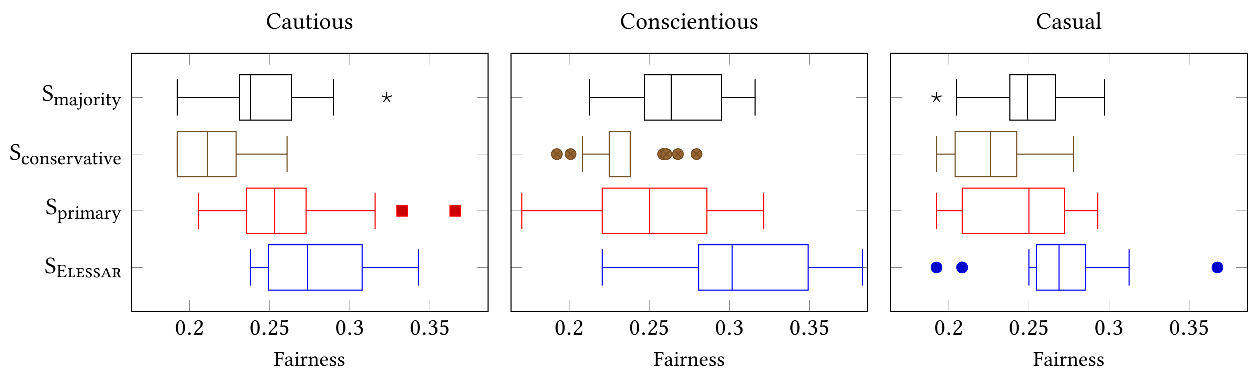

Figure 7: ELESSAR vs. other strategies: Fairness in societies that exhibit majorities in specified privacy attitudes. ELESSAR gives significantly better ($p < 0.01$) fairness with large effect size (Glass' $\Delta > 0.8$) than baselines.

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

Strategy \ Attitude	Cautious				Conscientious				Casual			
	Social	Best	Worst	Fairness	Social	Best	Worst	Fairness	Social	Best	Worst	Fairness
S_{ELESSAR}	**1.253**	2.895	**−0.704**	**0.283**	**1.304**	2.932	**−0.464**	**0.302**	1.383	3.120	**−0.667**	**0.270**
S_{primary}	1.150	2.855	−1.066	0.261	1.217	2.907	−1.211	0.251	1.331	3.128	−1.030	0.244
$S_{\text{conservative}}$	0.929	2.885	−1.793	0.216	1.085	2.927	−1.415	0.232	1.229	3.128	−1.378	0.225
S_{majority}	1.200	**2.916**	−1.270	0.243	1.277	**2.936**	−0.857	0.267	**1.387**	3.128	−0.921	0.250

approach learn social norms based on past interactions. In their evaluation of the approach, Ulusoy and Yolum employ majority voting for decision making in norms conflict scenarios. Ajmeri et al. [1] 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 [1] 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 a privacy cautious society where S_{majority} yields maximum social experience but least fairness. ELESSAR's focus is to balance the needs of primary and secondary users.

Cranefield et al. [6] show how agents can learn norms based on observations of behavior and sanction in a society, somewhat similar to Ajmeri et al. [1]. 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 additional regulatory mechanisms, both social (such as sanctioning) and psychological (such as guilt). A comprehensive study of these topics in conjunction with ethics is an important future direction.

ACKNOWLEDGEMENTS

This research was supported by the US Department of Defense under the Science of Security Lablet grant to NC State University.

REFERENCES

[1] Nirav Ajmeri, Hui Guo, Pradeep K. Murukannaiah, and Munindar P. Singh. 2018. 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)*. IJCAI, Stockholm, 28–34.

[2] Marguerite Barry, Kevin Doherty, Jose Marcano Belisario, Josip Car, Cecily Morrison, and Gavin Doherty. 2017. mHealth for Maternal Mental Health: Everyday Wisdom in Ethical Design. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI)*. ACM, Denver, 2708–2756.

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

[4] Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates, Hillsdale, New Jersey.

[5] Nicolas Cointe, Grégory Bonnet, and Olivier Boissier. 2016. Ethical Judgment of Agents' Behaviors in Multi-Agent Systems. In *Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems*. IFAAMAS, Singapore, 1106–1114.

[6] Stephen Cranefield, Felipe Meneguzzi, Nir Oren, and Bastin Tony Roy Savarimuthu. 2016. A Bayesian Approach to Norm Identification. In *Proceedings of the 22nd European Conference on Artificial Intelligence (ECAI)*. IOS Press, Amsterdam, 622–629.

[7] Stephen Cranefield, Michael Winikoff, Virginia Dignum, and Frank Dignum. 2017. No Pizza for You: Value-based Plan Selection in BDI Agents. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*. IJCAI, Melbourne, 178–184.

[8] Karen Da Silva Figueiredo and Viviane Torres Da Silva. 2013. 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)*. Springer, St. Paul, MN, 259–274.

[9] Francien Dechesne, Gennaro Di Tosto, Virginia Dignum, and Frank Dignum. 2013. No Smoking Here: Values, Norms and Culture in Multi-Agent Systems. *Artificial Intelligence and Law* 21, 1 (01 Mar 2013), 79–107.

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

[11] Ricard López Fogués, Pradeep K. Murukannaiah, Jose M. Such, and Munindar P. Singh. 2017. Sharing Policies in Multiuser Privacy Scenarios: Incorporating Context, Preferences, and Arguments in Decision Making. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 1 (March 2017), 5:1–5:29.

[12] Batya Friedman, Peter H. Kahn Jr., and Alan Borning. 2008. Value Sensitive Design and Information Systems. In *The Handbook of Information and Computer Ethics*, Kenneth Einar Himma and Herman T. Tavani (Eds.). John Wiley & Sons, Hoboken, New Jersey, Chapter 4, 69–101.

[13] Larry V. Hedges and Ingram Olkin. 2014. *Statistical Methods for Meta-Analysis*. Academic Press, Inc., Orlando.

[14] Alex Kayal, Willem-Paul Brinkman, Mark A. Neerincx, and M. Birna van Riemsdijk. 2018. Automatic Resolution of Normative Conflicts in Supportive Technology based on user values. *ACM Transactions on Internet Technology (TOIT)* 18, 4, Article 41 (May 2018), 21 pages.

[15] Nadin Kökciyan and Pınar Yolum. 2017. Context-Based Reasoning on Privacy in Internet of Things. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*. IJCAI, Melbourne, 4738–4744.

[16] Eugene F. Krause. 1973. Taxicab Geometry. *The Mathematics Teacher* 66, 8 (1973), 695–706.

[17] Beishui Liao, Marija Slavkovik, and Leendert van der Torre. 2019. Building Jiminy Cricket: An Architecture for Moral Agreements Among Stakeholders. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. ACM, Honolulu, 147–153.

[18] Maite Lopez-Sanchez, Marc Serramia, Juan A Rodriguez-Aguilar, Javier Morales, and Michael Wooldridge. 2017. Automating Decision Making to Help Establish Norm-based Regulations. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*. IFAAMAS, São Paulo, 1613–1615.

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

[20] Pardis Emami Naeini, Sruti Bhagavatula, Hana Habib, Martin Degeling, Lujo Bauer, Lorrie Faith Cranor, and Norman Sadeh. 2017. Privacy Expectations and Preferences in an IoT World. In *Proceedings of the 13th Symposium on Usable Privacy and Security (SOUPS)*. USENIX Association, Santa Clara, 399–412.

[21] Luis G. Nardin, Tina Balke-Visser, Nirav Ajmeri, Anup K. Kalia, Jaime S. Sichman, and Munindar P. Singh. 2016. Classifying Sanctions and Designing a Conceptual Sanctioning Process Model for Socio-Technical Systems. *The Knowledge Engineering Review (KER)* 31 (March 2016), 142–166. Issue 02.

[22] Serafim Opricovic and Gwo-Hshiung Tzeng. 2004. Compromise Solution by MCDM Methods: A Comparative Analysis of VIKOR and TOPSIS. *European Journal of Operational Research* 156, 2 (2004), 445–455.

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

[24] Eric Pacuit. 2017. Voting Methods. In *The Stanford Encyclopedia of Philosophy* (fall 2017 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University, Stanford.

[25] John Rawls. 1985. Justice as Fairness: Political not Metaphysical. *Philosophy and Public Affairs* 14, 3 (1985), 223–251.

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

[27] Leonard J. Savage. 1951. The Theory of Statistical Decision. *Journal of the American Statistical association* 46, 253 (1951), 55–67.

[28] Sebastian Schnorf, Aaron Sedley, Martin Ortlieb, and Allison Woodruff. 2014. A Comparison of Six Sample Providers Regarding Online Privacy Benchmarks. In *Proceedings of the SOUPS Workshop on Privacy Personas and Segmentation*. Menlo Park, 1–7.

[29] Shalom H. Schwartz. 2012. An Overview of the Schwartz Theory of Basic Values. *Online Readings in Psychology and Culture* 2, 1 (2012), 11.

[30] Marc Serramia, Maite López-Sánchez, Juan A. Rodríguez-Aguilar, Manel Rodríguez, Michael Wooldridge, Javier Morales, and Carlos Ansótegui. 2018. Moral Values in Norm Decision Making. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*. IFAAMAS, Stockholm, 1294–1302.

[31] Munindar P. Singh. 2013. Norms As a Basis for Governing Sociotechnical Systems. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 1, Article 21 (Dec. 2013), 23 pages.

[32] Kaj Sotala. 2016. Defining Human Values for Value Learners. In *Proceedings of the Workshops of the 30th AAAI Conference on Artificial Intelligence: AI, Ethics, and Society*. AAAI Press, Phoenix, 113–123.

[33] Jose M. Such. 2017. Privacy and Autonomous Systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*. IJCAI, Melbourne, 4761–4767.

[34] Onuralp Ulusoy and Pınar Yolum. 2019. Emergent Privacy Norms for Collaborative Systems. In *Proceedings of the 22nd International Conference on Principles and Practice of Multi-Agent Systems (PRIMA)*. Springer, Torino, 514–522.

[35] Duncan J. Watts and Steven H. Strogatz. 1998. Collective Dynamics of 'Small-World' Networks. *Nature* 393 (June 1998), 440–442.

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