We address the problem of designing agents that navigate social norms by selecting ethically appropriate actions. We present ELESSAR, a framework in which agents aggregate value preferences of users and select ethically appropriate actions through multi-criteria decision making in different social contexts. Via simulations, seeded with a survey of user values and attitudes, we find that ELESSAR agents act ethically and are effective than baseline agents, in terms of (1) exhibiting the Rawlsian property of fairness, and (2) yielding a satisfactory social experience to users. Our results are stable across agent societies of different sizes and connectedness.

KEYWORDS
Ethics; values; social norms; preferences; fairness;

1 INTRODUCTION
How can we develop intelligent agents that act ethically leading toward a just society of humans and agents? Acting ethically requires an understanding of the contextually relevant social norms and value preferences of the concerned individuals [11, 21]. 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 [6]. 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 worst-case experience across the members of a society.

Values. Ethicists subsume ethics in the theory of values [13]. Values are mostly universal across human societies [27, 30]. 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. [10] 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 [32] 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 [26] or considered ethically appropriate. That is, norm compliance is a weak standard for ethicality. Justice demands the society to achieve the strong standard of fairness [26]. 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. [19] associate norms with moral values, and reason about a normative system based on the preferences over the supported values. Da Silva Figueiredo and Da Silva [9] 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. [10] study compliance of norms based on values and to decide which norms to adopt. Kayal et al. [15] 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 [37] 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 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 (MAS) approach that brings together the following crucial elements: (1) value theory [13], (2) fair society design [26], and (3) decision making [24] 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 users and other stakeholders?

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 [24] 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’ [26] 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 MOTIVATION: LOCATION PRIVACY

For concreteness, we consider mobile social applications where privacy is an important value [1, 16, 34]. 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 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 Times Square and meets his uncle Harold there. 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 above 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 its users’ goals. In scenarios where norms are in conflict or users’ goals do not align with norms, a SIPA should select actions understanding
its users’ contextual preferences over applicable social norms [2]. 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 other stakeholders—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 [23]. 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. An action may promote or demote the values preferred by the SIPA’s primary user and other stakeholders, and may satisfy or violate norms applicable in the context in which the SIPA acts. Satisfaction or violation of these norms may attract sanctions.

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 other stakeholders—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 [23]. 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.

3.2 Decision Making

In a society of Gimli app users, the primary user is the one whose phone the SIPA runs on. Companions of the primary user are other stakeholders. Gimli is designed to serve a user’s goal of staying connected with friends and family by performing one of the following 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 $\{ (v_j > v_k, c) \mid v_j, v_k \in V, c \in C \}$ 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 $\{ (\text{pleasure} > \text{privacy}, \text{olympiad}), (\text{recognition} > \text{privacy}, \text{olympiad}) \}$. 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 user and other stakeholders, and (4) the social experience its user and other stakeholders 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 user’s goals.

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

3.2 Decision Making

A SIPA’s user and other stakeholders in an interaction may have inconsistent value preferences. Thus, a SIPA’s actions based solely on primary user’s preferences may conflict with its other stakeholders’ 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 [33] models a human’s values via a reward function that an agent can learn and maximize.

We adapt VIKOR [24], 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 [25].

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 context c and the applicable norms N, let \( f_{c,a} \) be the utility of value \( v \) when action \( a \) is selected in \( c \). This utility indicates the extent to which the value is promoted. A SIPA perceives 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, \( f_{\text{max}}^v \) and \( f_{\text{min}}^v \) for each value \( v \) over alternative actions \( a \) to bring about a goal. That is, \( f_{\text{max}}^v = \max f_{c,a} \) and \( f_{\text{min}}^v = \min f_{c,a} \).

3. For each alternative action \( a \), compute the weighted and normalized Manhattan distance [17]:

\[
S_a = \begin{cases} 
    \sum_{v=1}^{n} w_v \frac{(f_{\text{max}}^v - f_{c,a})}{(f_{\text{max}}^v - f_{\text{min}}^v)}, & \text{for } f_{c,a}^v \neq f_{\text{max}}^v \\
    0, & \text{for } f_{c,a}^v = f_{\text{max}}^v
\end{cases}
\]

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 [19] goal of being able to quantitatively reason about qualitative preferences over the moral values.

4. For each alternative action \( a \), compute the weighted and normalized Chebyshev distance [4] (here \( w_v \) is the weight for value \( v \)):

\[
R_a = \begin{cases} 
    \max \left[ w_v \frac{(f_{\text{max}}^v - f_{c,a})}{(f_{\text{max}}^v - f_{\text{min}}^v)} \right], & \text{for } f_{c,a}^v \neq f_{\text{max}}^v \\
    0, & \text{for } f_{c,a}^v = f_{\text{max}}^v
\end{cases}
\]

5. For each alternative action \( a \), compute

\[
Q_a = k \frac{(S_a - S^*) + (1 - k) (R_a - R^*)}{(R^* - R^*)}
\]

Here, (i) \( S^* = \min S_a \); (ii) \( S^- = \max S_a \); (iii) \( R^* = \min R_a \); (iv) \( R^- = \max R_a \); and (v) \( k \) trades off group and individual experience.

\( S_a \) derived using Manhattan distance, represents the distance to the weighted sum of the maximum utility values. Minimizing \( S \) maximizes the total utility of the society.

\( R_a \) derived using Chebyshev distance, is the worst-case individual regret, i.e., the maximum distance between an individual’s optimal utility and actual utility [28]. Minimizing \( R \) ascertains fairness.

Conforming to Rawls’ maximin doctrine [26] 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 \( Q_a \) as the best solution if (i) it is better than the second-best action by a threshold \( h \) or (ii) it is 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 \( Q_a = \min Q_a < h \), 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 select 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 \( a_3 \), share only with Grace, is the best solution.

4 SURVEY TO SEED SIMULATED SOCIETIES

We conducted a survey of privacy attitudes following Naeini et al.’s [22] finding that users’ preferences can be accurately predicted by observing their decisions in a few scenarios. The survey data helps 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: 81% men; 19% women). First, in our survey, the respondents completed a privacy attitude survey [29] 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).

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

<table>
<thead>
<tr>
<th>Alternatives</th>
<th>Frank’s Values</th>
<th></th>
<th></th>
<th>Harold’s Values</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pleasure</td>
<td>Privacy</td>
<td>Recognition</td>
<td>Security</td>
<td>Pleasure</td>
<td>Privacy</td>
</tr>
<tr>
<td>a_1 All</td>
<td>1.00</td>
<td>0.50</td>
<td>1.00</td>
<td>0.50</td>
<td>0.50</td>
<td>0.00</td>
</tr>
<tr>
<td>a_2 Common</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>1.00</td>
<td>0.50</td>
<td>0.00</td>
</tr>
<tr>
<td>a_3 Grace</td>
<td>0.00</td>
<td>0.50</td>
<td>0.00</td>
<td>0.00</td>
<td>0.50</td>
<td>1.50</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th># of respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cautious</td>
</tr>
<tr>
<td>Conscientious</td>
</tr>
<tr>
<td>Casual</td>
</tr>
</tbody>
</table>

Figure 2: Distribution of privacy attitudes of respondents.

Table 2: Simulated places with attributes safe and sensitive.

<table>
<thead>
<tr>
<th>Place</th>
<th>Safe</th>
<th>Sensitive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attending graduation ceremony</td>
<td>–</td>
<td>No</td>
</tr>
<tr>
<td>Presenting a conference paper</td>
<td>–</td>
<td>No</td>
</tr>
<tr>
<td>Studying in library</td>
<td>Yes</td>
<td>–</td>
</tr>
<tr>
<td>Visiting airport</td>
<td>Yes</td>
<td>–</td>
</tr>
<tr>
<td>Hiking at night</td>
<td>No</td>
<td>–</td>
</tr>
<tr>
<td>Being stuck in a hurricane</td>
<td>No</td>
<td>–</td>
</tr>
<tr>
<td>Visiting a bar with fake ID</td>
<td>–</td>
<td>Yes</td>
</tr>
<tr>
<td>Visiting a drug rehab center</td>
<td>–</td>
<td>Yes</td>
</tr>
</tbody>
</table>

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.

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.

- **SElessar**. Compute a context-sharing policy from users’ value preferences using VIKOR.
- **SPrimary**. Produce a context-sharing policy based only on the primary user’s value preferences—how location sharing works today in social networking websites.
- **SConservative**. 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.
- **SMajority**. Produce the most common policy based on the users’ value preferences. This strategy corresponds to majority voting [12] 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 [26]. Higher is better.
- **Fairness**, reciprocal of the disparity between the best and worst accumulative individual experiences obtained by users in a period of time [26]. Higher is better.

**Computing Experience**. The utility that a SIPA yields from a sharing policy in a certain context, whether to a primary user or other stakeholders, is a weighted sum of the numeric utilities that the user perceives for 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.
5.3 Hypotheses

To answer our research question, we evaluate four hypotheses, each a claim 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.

- \( H_{\text{social}} \): Elessar wins on social experience.
- \( H_{\text{best}} \): Elessar wins on best individual experience.
- \( H_{\text{worst}} \): Elessar wins on worst individual experience.
- \( H_{\text{fair}} \): Elessar wins on fairness.

5.4 Experimental Setup

We adopt MASON [20], 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: Mixed Agent Society

We experiment with a society of users with mixed privacy attitudes representing the respondents of our study from Section 4. We map the SIPAs evenly to respondents. Each pair of SIPAs relates as coworkers, friends, family (with equal probability), or strangers. To improve naturalness, we select parameters for a small world [36], i.e., degree: 10, rewiring probability: 0.05, edges: 3,445, clustering coefficient: 0.56, density: 0.014, mean distance: 4.71.

To evaluate \( H_{\text{social}} \), we compare the social experience yielded by SIPAs incorporating the four decision-making strategies—\( S_{\text{ELESSAR}} \), \( S_{\text{primary}} \), \( S_{\text{conservative}} \), and \( S_{\text{majority}} \). Similarly, for \( H_{\text{best}}, H_{\text{worst}}, \) and \( H_{\text{fair}} \), 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’ \( \Delta \) [14], which is computed as the difference in the means divided by the standard deviation of the control group. We choose Glass’ \( \Delta \) to measure effect size because it is better suited when standard deviations are different between groups. Recognizing some caveats, we adopt Cohen’s [5] 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.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Social</th>
<th>Best</th>
<th>Worst</th>
<th>Fairness</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S_{\text{ELESSAR}} )</td>
<td>1.305</td>
<td>3.071</td>
<td>-0.568</td>
<td>0.279</td>
<td>~</td>
</tr>
<tr>
<td>( S_{\text{primary}} )</td>
<td>1.226</td>
<td>3.013</td>
<td>-1.138</td>
<td>0.247</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>( S_{\text{conservative}} )</td>
<td>1.065</td>
<td>3.069</td>
<td>-1.554</td>
<td>0.218</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>( S_{\text{majority}} )</td>
<td>1.276</td>
<td>3.075</td>
<td>-1.154</td>
<td>0.241</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>

Figure 3: ELESSAR vs. others: Social experience in a mixed society.

We see 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 experience, worst individual experience, and fairness (\( p < 0.01; \) Glass’ \( \Delta > 0.8 \) indicating large effect size) compared to the three baseline decision-making strategies. The null hypotheses corresponding to \( H_{\text{social}}, H_{\text{worst}}, \) and \( H_{\text{fair}} \) are rejected. These results indicate that ELESSAR yields solutions in which each companion is treated fairly—i.e., ELESSAR SIPAs act ethically.

5.6 Experiments: 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.

5.7 Threats to Validity and Mitigation

We identify and mitigate three threats. The first threat concerns simulation as an evaluation methodology. To mitigate this threat, 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 of an 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.

Figure 4: ELESSAR vs. others: Fairness in a mixed society.

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. These differences in the outcomes are statistically significant ($p < 0.01$; Glass’ $\Delta > 0.8$ indicating large effect size). 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 ($p < 0.01$; Glass’ $\Delta > 0.8$). 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 ($p < 0.01$; Glass’ $\Delta > 0.8$); thus, we reject null hypotheses related to $H_{worst}$ and $H_{fairness}$.

6 SUMMARY AND OUTLOOK

Incorporating ethics into AI is a major modern research direction. Ethics inherently involves looking beyond one’s self-interest [21]. That is, an agent must consider users in addition to its primary values such as confidentiality, disapproval, and avoiding infringements. There is an argument-based architecture for moral agents, combining norms, argumentation, and agreements to help them make a moral decision. An ethical decision, in their architecture, is one on which all users agree, not necessarily one fair to users.

Recent work has been promoting the reasoning of values in decision making to advance ethical agents. Liao et al. [18] propose an approach for making these decisions, combining norms, argumentation, and agreements to help an agent 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. [3] 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 values, which are common in social settings.

Serramia et al. [31] 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. [15] 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. [8] 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 values to make morally correct decisions, which is difficult to realize.

Ulusoy and Yolum [35] design a normative approach for making privacy decision related to content sharing. Agents in their 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. [2] develop agents who apply norms to provide privacy assistance to their users. Their notion of privacy recognizes values such as confidentiality, disapproval, and avoiding interfering into others’ space. However, Ajmeri et al.’s [2] agents seek to
Figure 6: **ELESSAR vs. other strategies:** Social experience in societies that exhibit majorities in specified privacy attitudes. **ELESSAR** yields higher social experience than baselines ($p < 0.01; \text{Glass'} \Delta > 0.8$ 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.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Attitude</th>
<th>Social</th>
<th>Best</th>
<th>Worst</th>
<th>Fairness</th>
<th>Social</th>
<th>Best</th>
<th>Worst</th>
<th>Fairness</th>
<th>Social</th>
<th>Best</th>
<th>Worst</th>
<th>Fairness</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>S_{ELESSAR}</strong></td>
<td>Cautious</td>
<td>1.253</td>
<td>2.895</td>
<td>-0.704</td>
<td>0.283</td>
<td>1.304</td>
<td>2.932</td>
<td>-0.464</td>
<td>0.302</td>
<td>1.383</td>
<td>3.120</td>
<td>-0.667</td>
<td>0.270</td>
</tr>
<tr>
<td></td>
<td>Conscientious</td>
<td>1.150</td>
<td>2.855</td>
<td>-1.066</td>
<td>0.261</td>
<td>1.217</td>
<td>2.907</td>
<td>-1.211</td>
<td>0.251</td>
<td>1.331</td>
<td>3.128</td>
<td>-1.030</td>
<td>0.244</td>
</tr>
<tr>
<td></td>
<td>Casual</td>
<td>0.929</td>
<td>2.885</td>
<td>-1.793</td>
<td>0.216</td>
<td>1.085</td>
<td>2.927</td>
<td>-1.415</td>
<td>0.232</td>
<td>1.229</td>
<td>3.128</td>
<td>-1.378</td>
<td>0.225</td>
</tr>
<tr>
<td><strong>S_{primary}</strong></td>
<td>Cautious</td>
<td>1.200</td>
<td>2.916</td>
<td>-1.270</td>
<td>0.243</td>
<td>1.277</td>
<td>2.936</td>
<td>-0.857</td>
<td>0.267</td>
<td>1.387</td>
<td>3.128</td>
<td>-0.921</td>
<td>0.250</td>
</tr>
<tr>
<td></td>
<td>Conscientious</td>
<td>1.540</td>
<td>2.916</td>
<td>-1.270</td>
<td>0.243</td>
<td>1.277</td>
<td>2.936</td>
<td>-0.857</td>
<td>0.267</td>
<td>1.387</td>
<td>3.128</td>
<td>-0.921</td>
<td>0.250</td>
</tr>
<tr>
<td></td>
<td>Casual</td>
<td>1.583</td>
<td>2.916</td>
<td>-1.270</td>
<td>0.243</td>
<td>1.277</td>
<td>2.936</td>
<td>-0.857</td>
<td>0.267</td>
<td>1.387</td>
<td>3.128</td>
<td>-0.921</td>
<td>0.250</td>
</tr>
</tbody>
</table>

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 user and other stakeholders.

Cranefield et al. [7] show how agents can learn norms based on observations of behavior and sanction in a society, somewhat similar to Ajmeri et al. [2]. 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.

**ACKNOWLEDGMENTS**

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