Mitigating Privacy Conflicts with Computational Theory of Mind

Emre Erdogan
Utrecht University
Utrecht, Netherlands
e.erdogan1@uu.nl

Hüseyin Aydın
Utrecht University
Utrecht, Netherlands
Middle East Technical University
Ankara, Turkey
huseyin@ceng.metu.edu.tr

Frank Dignum
Umeå University
Umeå, Sweden
dignum@cs.umu.se

Rineke Verbrugge
University of Groningen
Groningen, Netherlands
l.c.verbrugge@rug.nl

Pınar Yolum
Utrecht University
Utrecht, Netherlands
p.yolum@uu.nl

ABSTRACT

Multiagent systems bring together agents that represent different users with possibly different concerns. When interacting to make decisions, conflicts occur. A well-known case is with privacy. Agents often need to manage the privacy of content that belong to multiple users, such as sharing group pictures on social media. When agents have different expectations on how the content should be shared, *multi-party privacy conflicts* can arise. How should we design agents to deal with such conflicts? We have studied an empirical user study to understand the effect of group dynamics in various multi-party privacy settings. Our findings show that as users' beliefs and knowledge about others evolve, privacy expectations shift as well. Inspired by this, we propose computational agents that mimic a human-inspired Theory of Mind (ToM) model to help their users preserve their privacy in multi-party privacy conflicts. The agents can express empathy when others are in need but can also fight for their own privacy. We evaluate our approach in multiagent simulations with varying decision-making strategies. Our results demonstrate that ToM-enabled agents improve privacy preservation for all parties, and even more when their understanding of others is dynamically updated through learning.

KEYWORDS

Multi-Party Privacy; Theory of Mind; Human-Centered AI

ACM Reference Format:
Emre Erdogan, Hüseyin Aydın, Frank Dignum, Rineke Verbrugge, and Pınar Yolum. 2025. Mitigating Privacy Conflicts with Computational Theory of Mind. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 9 pages.

1 INTRODUCTION

Understanding human values [8, 31] is an important first step in building socio-technical systems, where software agents and humans exist together. Agents in such systems need to be able to adhere to the values of the humans that they serve for, so that they make decisions or take actions that would be representative of the humans. This would enable the agents to explain the moral underpinnings of their decisions to the humans [11].

Values have been instrumental in designing and calibrating multiagent systems. Methodologies, such as value sensitive design [17, 41] have positioned values as first-class citizens in the design of socio-technical systems and identified ways that can be accounted for in the design. Similarly, the link between values and other constructs, such as norms, have been investigated thoroughly. Aydoğan et al. [4] use values as a means to negotiate the norms that will be in effect in a multiagent system. They consider a range of values from privacy to safety. Serramia et al. [32] study whether a given set of norms of a multiagent system promote a value of interest. Kayal et al. [21] use given human values to detect and resolve norm conflicts. Hadfield-Menell et al. [18] use cooperative inverse reinforcement learning to help agents learn values through interactions with a human cooperatively. An important premise here is that humans share values that the agents should adhere to work with humans. However, humans can have varying values as well. Liscio et al. [24] develop a hybrid (human and AI) methodology to identify context-specific values as opposed to general values that are overarching [31].

At the same time, it is well-known that different humans have different values and even different understanding and interpretation of the same value. Consider privacy as an important value that has received a lot of attention lately. Privacy is largely personalized (e.g., individuals have different understandings) and context-dependent (e.g., individuals change their valuations based on a given situation). When agents that represent different humans work together, they need to identify and handle the privacy conflicts that might arise. In order to address this, various multiagent decision mechanisms have been designed. On one hand, auction-based mechanisms have been designed to resolve privacy conflicts [33, 40]. In parallel, negotiation-based [16, 22, 37] and argumentation-based [23, 27] decision mechanisms have been developed to enable agents to reach privacy decisions. While such mechanisms are useful, they do not address the following questions: How should an agent behave when its privacy expectations clash with those of others in the system? What actions can be taken to mitigate potential privacy conflicts? Would taking others' perceived privacy expectations into account help in mitigating conflicts?

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

Accordingly, this paper proposes an agent design that employs *Theory of Mind* (ToM) reasoning [9, 30] — the human ability to reason about mental content of others such as their beliefs, desires, preferences, goals, etc.— to assess the values of others and incorporate them into its actions. The agent design we propose reasons about others' privacy expectations and determines whether it can account for them when a privacy conflict is about to happen. Depending on how important its privacy is for a given content, it can choose between favoring others' opinion or persisting on its own judgment. We base the ToM reasoning of our architecture to a user study [5] conducted to understand how humans deal with privacy conflicts. We show the added benefit of ToM reasoning in mitigating privacy conflicts between agents over multiagent simulations.

The rest of this paper is organized as follows. Section 2 explains how ToM enables humans to understand others' privacy expectations and adapt their behaviors in social settings with an example. Section 3 provides insights from the user study to understand privacy dynamics among humans. Section 4 uses these insights to design agents with ToM models. Section 5 evaluates these agents in terms of how well they mitigate privacy conflicts and their use of ToM. Section 6 discusses our work in relation to the literature.

2 UNDERSTANDING PRIVACY CONFLICTS VIA THEORY OF MIND

A key aspect of effective human decision-making is the ability to understand others' perspectives, and ToM reasoning plays a crucial role in this process, especially when communication is limited. In socially interactive settings, individuals use ToM to model others' expectations and recognize how their own perspective may differ from those of others. With this understanding, individuals can adapt their social behaviors in a more cooperative and socially adaptive manner. To illustrate this, consider the following example that pertains to privacy of multiple individuals.

Alice has recently joined a group that is deciding as a group whether to publicly share a group photo as in Figure 1 (e.g., on a social network). Most of the other members are hesitant to share the photo due to privacy concerns, but Alice is willing to share. Without ToM reasoning, Alice might not be aware of others' privacy expectations and simply choose to share the photo in line with her own preference. However, if she wishes to understand the privacy preferences of the group and navigate the situation accordingly, she would need to consider the group's preferences, something her ToM model could assist with. Suppose Alice assumes that her desire to share conflicts with the group's preference to keep the photo private. We can explore three possible behaviors Alice might adopt in response to this situation (as shown in Figure 1).

In the first behaviour, to respect the majority's privacy concerns, Alice decides to conform to the group's preference and refrains from sharing the photo. This choice allows her to adhere to the group's concerns, prioritizing collective privacy over her own desire to share.

In the second behaviour, Alice feels strongly about the importance of sharing the photo. While she acknowledges the group's concerns, she advocates for her perspective and chooses to share the photo in line with her original preference, even if it goes against the majority.

Figure 1: Alice (upper-right corner) wants to share this group's photo online. Using her ToM model, she unveils that others are hesitant to share. Alice can conform to the others' preference, advocate her own, or preserve everyone's privacy.

In the third behavior, Alice prioritizes preserving everyone's privacy and therefore chooses not to share the photo. Similarly, if the situation were reversed, where the group wanted to share the photo and Alice preferred not to, she would advocate for her own privacy preference rather than conforming to the majority.

In each case, Alice must understand the privacy preferences of the other group members to make an informed sharing decision. To complicate matters, suppose Alice is not entirely sure of the others' preferences (e.g., due to lack of communication or her recent addition to the group). In this case, she might begin by projecting her own preferences onto the group through ToM reasoning, modeling how she thinks others might feel about sharing. However, Alice's ToM model could be inaccurate. Through repeated interactions and observations, Alice can learn more about the group and refine her understanding of the group's preferences, making her ToM model more accurate over time.

Just as Alice refines her understanding of others' preferences through observation and interaction, the same principles can be applied in designing computational models that mimic human decision-making. By incorporating ToM reasoning, these models can simulate how individuals navigate complex social situations where preferences are uncertain or conflicting. The gradual improvement in Alice's ToM model parallels how computational agents can be designed to learn and adapt, adjusting their behavior to match the preferences of others over time. This connection between human and agent decision-making allows us to explore how computational models can simulate real-world dynamics.

3 DYNAMICS OF PRIVACY CONFLICTS

In order to understand how humans navigate the situations like the ones outlined in the previous section, we examined a user study [5] where a game named RESOLVE is devised to i) understand whether users make different privacy choices when they are alone or in a group, and ii) if the choices differ what are the reasons that affect this change. In the beginning of the game, each player captures their own pictures mimicking a set of emojis denoting different types of

emotions. After capturing each picture they also give a score for their tendency to share that picture in Likert scale from 1 to 5 (i.e., 1: very unlikely to share, 2: unlikely to share, 3: unsure, 4: likely to share, 5: very likely to share). Next, one of the pictures of the player is merged with three other (non-participant) pictures to create grid-like group picture (see Figure 1) that mimics an online meeting setup. Then, each player participates in an auction by placing a *bid* from their available *budget* to support their preferences so that a group decision can be made on whether to share the photo or not.

After all players state their preferences and bids to support these preferences in the first round of a game, the choice with greater total amount of credits becomes the temporary group decision for the content. In order to understand whether the participants would change their behavior if they had known what others would do, the players are given a second opportunity to rethink their choices and bids and update them as they see fit. If there are changes, the outcome is recalculated and the definite group decision is determined for the content.

In the study, the game environment is configured so that the other three players are software agents with predetermined decisions and bid amounts. Each participant plays the game across 16 different scenarios, where the majority of the group either matches the participant's privacy choice (SHARE or NOT SHARE) or the participant does not receive enough support from the group for their choice. 40 participants successfully completed all 16 scenarios, resulting in 640 scenario instances where various behavioral patterns were observed. While further details can be found in the study itself, we present the main observations that we focused on and the insights gained from them.

We study how much players divert from their individual preferences in a group setting. As an example, consider a case where a participant declares the tendency to share an emoji as "likely" or "very likely", yet they bid for NOT SHARE decision in the first round of the game. We observed that a participant divert from their initial tendencies on average of only 4.3 (± 2.7) scenarios (out of 16) during the game.

Insight 1. Participants tend to make decisions that reflect their own privacy preferences.

This observation highlights how participants prioritize their own privacy preferences when making decisions. On the other hand, we also observed that some participants change their decision to SHARE or NOT SHARE in the second round of the game, after knowing others' choices. This was observed in 24 participants in 52 scenario instances which leads us to the following insight.

Insight 2. Participants sometimes make decisions that match the privacy preferences of others (whether those preferences are known or assumed).

It suggests that participants sometimes act in line with others' preferences, even when this leads to decisions that conflict with their own. This behavior can arise from their assumptions (in the first round, when they are unaware of others' decisions) or their knowledge (in the second round, after others' decisions are revealed) about others' preferences.

Another observation is that participants often revise their initial decision if the first round of the game does not result in their preferred outcome. This behavior was observed in 45 scenario instances. Furthermore, in 34 of these cases, the participant's decision was in direct conflict with the rest of the group.

Insight 3. Participants sometimes adjust their decisions to conform to the majority.

This insight becomes more apparent when a participant is left alone, i.e., when the preferences of others are in line with each other but against the participant's decision. Unlike this behavior, we also observed in 79 scenario instances, 27 participants increased their bids in the second round of the game, independent of the outcome achieved in the first round.

Insight 4. Participants sometimes increase their bids while maintaining their decisions to advocate more strongly for their own preferences.

This reflects a behavior where participants aim to reinforce their position through higher bids. If we look at the break down of the dynamic of this behavior, based on the initial decision, we observe the following: Participants that express NOT SHARE in the initial round tend to increase their bids to advocate for keeping the content private (in 48 scenario instances out of 79 in total). Those that express SHARE in the first round mostly switch to NOT SHARE when others prefer that (in 36 scenario instances out of 52 in total).

Insight 5. Participants sometimes change their decisions to conform to the majority, but only when it serves to preserve privacy.

It indicates that participants are willing to increase their bids to protect their privacy and may conform to the majority if it helps achieve this goal.

As our last major observation from the study, in 160 scenario instances, 32 different participants decreased their bids in the second round of the game. These instances include the cases where the participants kept their initial privacy decision or accommodate the group-decision.

Insight 6. Participants sometimes decrease their bids while maintaining their decisions to conserve their budget for future interactions.

This insight suggests that participants make this choice when they feel confident about the outcome and wish to save resources.

In total, we identified six insights that reflect recurring behavioral patterns in people's decision-making over group content. These patterns suggest varying degrees of awareness and consideration for others' preferences, observed both in sharing decisions and bids, which align with the core principles of ToM reasoning. Motivated by these insights, we investigate the potential benefits of implementing ToM reasoning computationally in software agents to deal with multi-party privacy conflicts.

4 AGENT DESIGN

We design computational agent models that use ToM reasoning in decision-making. Our goal is to enhance the agents' ability to simulate real-world decision-making processes as suggested by the insights from Section 3, where individuals must consider not only their own preferences but also the potential preferences of others [1, 34].

Given a set of contents $\mathcal{C}$, we represent agent $\mathbb{A}$ as a tuple such that $\mathbb{A} = \langle \mathcal{P}, \mathcal{P}', \mathcal{F}, \mathcal{G}, \mathcal{U} \rangle$, where:

- $\mathcal{P}$ is the mapping of agent's own preferences for each content in $\mathcal{C}$,
- $\mathcal{P}'$ is the mapping of other agents' (assumed) preferences for each content in $\mathcal{C}$,
- $\mathcal{F}$ is the decision function that determines which action to perform based on $\mathcal{P}$ and $\mathcal{P}'$,
- $\mathcal{G}$ is the function that estimates other agents' actions based on $\mathcal{P}'$, and
- $\mathcal{U}$ is the function that updates $\mathcal{P}'$.

Within the context of this paper, $\mathcal{C}$ consists of emojis that are used in the user study. Hence, $\mathcal{P}$ refers to a mapping from these emojis to corresponding values on a scale from 1 to 5 that mirror the Likert scale. Similar to $\mathcal{P}$, $\mathcal{P}'$ also refers to a mapping from a list of emojis to values on the same scale. However, while $\mathcal{P}$ represents $\mathbb{A}$'s own preferences, $\mathcal{P}'$ provides the base information that $\mathbb{A}$ uses for ToM reasoning, allowing it to make assumptions about the preferences of other agents. $\mathcal{F}$ represents the strategy that $\mathbb{A}$ uses to make a sharing decision and place a bid to support it, while $\mathcal{G}$ is used to estimate other agents' sharing decisions based on $\mathcal{P}'$. $\mathcal{U}$ updates $\mathcal{P}'$ based on new information (i.e., others' actions).

Using the insights from the user study, we have designed various software agents that mimic different participant behaviors in decision-making scenarios. The core design principle is to create agents that capture the essence of human decision-making patterns as reflected in the study, from simple self-focused strategies to more complex, prosocial behaviors [43]. The key differences between the agents lie in their decision-making strategies: some rely solely on their own preferences while others employ ToM to make assumptions about the preferences of other agents and adjust their decisions accordingly. This incorporation of ToM allows more complex agents to better simulate human behaviors. Below, we outline the characteristics and behaviors of each agent.

4.1 *ToM-0* Agent

Based on Insight 1, the *ToM-0* agent (i.e., agent without ToM) is designed to make decisions solely based on its own privacy preferences. Since *ToM-0* agent does not consider others' preferences, we represent it as a tuple *ToM-0* $= \langle \mathcal{P}, \emptyset, \mathcal{F}, \emptyset, \emptyset \rangle$. The *ToM-0* agent serves as a baseline: Its decisions are consistent with its privacy preferences, and its bids are proportional to these preferences (e.g., "5: very likely to share" indicates "SHARE with a high bid" whereas "2: not likely to share" indicates "NOT SHARE with a low bid"). To add variability, we introduce randomness in the bids, selecting them from a uniform distribution between two integers. The decision function $\mathcal{F}$ of *ToM-0* is outlined in Algorithm 1 where b_{MIN}, b_{MID}, and b_{MAX} represent the minimum, midpoint, and maximum of permissible bid amounts, respectively. Next, we move to the agents that use ToM, denoted as *ToM-1* agents.

4.2 *ToM-1.M* Agent: Majority-Conforming

The *ToM-1.M* agent inspired by Insights 2, 3 and 6, is designed to conform its decisions to the majority's preferences (*M* in *ToM-1.M* stands for "majority-conforming behavior"). We represent *ToM-1.M* agent as a tuple *ToM-1.M* $= \langle \mathcal{P}, \mathcal{P}', \mathcal{F}, \mathcal{G}, \emptyset \rangle$. Notice that it does not have $\mathcal{U}$ since *ToM-1.M* does not update $\mathcal{P}'$. Given a set of agents $\mathcal{A}$ and a mapping $\mathcal{C}' : \mathcal{A} \rightarrow \mathcal{C}$, *ToM-1.M* first calculates each agent $a \in \mathcal{A}$'s expected decision from assumed preferences $\mathcal{P}'$ for a content $c \in \mathcal{C}$ with $\mathcal{G}$, as outlined in Algorithm 2. Then, it makes a potential decision without taking the others into account by using Algorithm 1 and creates a potential bid. Finally, it makes a decision that conforms to the (expected) majority decision while placing a bid that is at most as high as the potential bid to save budget. The decision function $\mathcal{F}$ of *ToM-1.M* is outlined in Algorithm 3.

Algorithm 1: Decision Function $\mathcal{F}$ of *ToM-0* Agent

Input: $\mathcal{P}(c)$ as sharing preference for given content c
Output: sharing decision D and bid B

if $\mathcal{P}(c) = 1$ **then**
  $D \leftarrow NOT\_SHARE$
  $B \leftarrow$ a random integer in the range $(b_{MID}, b_{MAX}]$
else if $\mathcal{P}(c) = 2$ **then**
  $D \leftarrow NOT\_SHARE$
  $B \leftarrow$ a random integer in the range $(b_{MIN}, b_{MID}]$
else if $\mathcal{P}(c) = 3$ **then**
  $D \leftarrow$ randomly select NOT_SHARE or $SHARE$
  $B \leftarrow b_{MIN}$ `// if not sure, give the smallest possible bid`
else if $\mathcal{P}(c) = 4$ **then**
  $D \leftarrow SHARE$
  $B \leftarrow$ a random integer in the range $(b_{MIN}, b_{MID}]$
else
  $D \leftarrow SHARE$
  $B \leftarrow$ a random integer in the range $(b_{MID}, b_{MAX}]$
end

Algorithm 2: Estimations for Other Agents' Decisions ($\mathcal{G}$)

Input: agents $\mathcal{A}$, agents' contents $\mathcal{C}'$,agents' preferences $\mathcal{P}'$
Output: agents' (expected) decisions $\mathcal{D}'$

$\mathcal{D}' \leftarrow \emptyset$ `// initialize`
for each *agent* $a \in \mathcal{A}$ **do**
  $c \leftarrow \mathcal{C}'(a)$ `// get agent's content`
  if $\mathcal{P}'(c) = 1$ **or** $\mathcal{P}'(c) = 2$ **then**
    $\mathcal{D}'(a) \leftarrow NOT\_SHARE$
  else if $\mathcal{P}'(c) = 4$ **or** $\mathcal{P}'(c) = 5$ **then**
    $\mathcal{D}'(a) \leftarrow SHARE$
  else
    $\mathcal{D}'(a) \leftarrow$ randomly select NOT_SHARE or $SHARE$
  end
end

4.3 *ToM-1.A* Agent: Preference-Advocating

The *ToM-1.A* agent, reflecting Insights 4 and 6, advocates for its own privacy preferences by adjusting its bidding strategy (*A* in *ToM-1.A* stands for "preference-advocating behavior"). We represent *ToM-1.A* agent as a tuple *ToM-1.A* $= \langle \mathcal{P}, \mathcal{P}', \mathcal{F}, \mathcal{G}, \emptyset \rangle$. Similar to the *ToM-1.M* agent, it first calculates others' expected decisions, makes

Algorithm 3: Decision Function $\mathcal{F}$ of *ToM-1.M* Agent

Input: $\mathcal{P}(c)$ as own privacy preference for given content c, agents $\mathcal{A}$, agents' contents $\mathcal{C}'$,agents' preferences $\mathcal{P}'$
Output: sharing decision D and bid B

$\mathcal{D}' \leftarrow$ Run Algorithm 2 with $\mathcal{A}$, $\mathcal{C}'$, and $\mathcal{P}'$
$D,\ B \leftarrow$ Run Algorithm 1 with $\mathcal{P}(c)$
$m_S = \sum_{d \in \mathcal{D}'} \mathbb{1}_{d=SHARE}$ // count SHARE decisions
$m_{NS} = \sum_{d \in \mathcal{D}'} \mathbb{1}_{d=NOT\_SHARE}$ // count NOT_SHARE decisions
if $m_S < m_{NS}$ **then**
  $D \leftarrow NOT\_SHARE$ // decision update
else
  $D \leftarrow SHARE$ // decision update
end
$b_u \leftarrow \min\{B, (b_{MIN} + b_{MID})/2\}$ // upper bound
$B \leftarrow$ a random integer in the range $[b_{MIN}, b_u)$

a potential decision without considering others, and creates a potential bid. Then, it places a higher bid than the potential bid when it anticipates a majority decision that opposes its own. Conversely, it opts for a low bid when it predicts that it already conforms to the majority, aiming to conserve its budget. The decision function $\mathcal{F}$ of *ToM-1.A* is outlined in Algorithm 4.

Algorithm 4: Decision Function $\mathcal{F}$ of *ToM-1.A* Agent

Input: $\mathcal{P}(c)$ as own privacy preference for given content c, agents $\mathcal{A}$, agents' contents $\mathcal{C}'$,agents' preferences $\mathcal{P}'$
Output: sharing decision D and bid B

$\mathcal{D}' \leftarrow$ Run Algorithm 2 with $\mathcal{A}$, $\mathcal{C}'$, and $\mathcal{P}'$
$D,\ B \leftarrow$ Run Algorithm 1 with $\mathcal{P}(c)$
$m_S = \sum_{d \in \mathcal{D}'} \mathbb{1}_{d=SHARE}$ // count SHARE decisions
$m_{NS} = \sum_{d \in \mathcal{D}'} \mathbb{1}_{d=NOT\_SHARE}$ // count NOT_SHARE decisions
if ($m_S < m_{NS}$ **and** $D = SHARE$) **or** ($m_S > m_{NS}$ **and** $D = NOT\_SHARE$) **then**
  $b_l \leftarrow \max\{B, (b_{MAX} + b_{MID})/2\}$ // lower bound
  $B \leftarrow$ a random integer in the range $(b_l, b_{MAX}]$
else
  $b_u \leftarrow \min\{B, (b_{MIN} + b_{MID})/2\}$ // upper bound
  $B \leftarrow$ a random integer in the range $[b_{MIN}, b_u)$
end

4.4 *ToM-1.P* Agent: Privacy-Preserving

Based on Insights 2, 5 and 6, the *ToM-1.P* agent makes decisions to conform to the majority when this helps to preserve privacy (*P* in *ToM-1.P* stands for "privacy-preserving behavior"). We represent *ToM-1.P* agent as a tuple *ToM-1.P* $= \langle \mathcal{P}, \mathcal{P}', \mathcal{F}, \mathcal{G}, \emptyset \rangle$. This agent balances conformity to social norms with a strong emphasis on maintaining privacy. The *ToM-1.P* agent aims to protect both its own privacy and that of others, though it employs slightly different sub-strategies for each. Similar to the *ToM-1.M* agent, it first calculates others' expected decisions, makes a potential decision without considering others, and creates a potential bid. When its potential decision is "NOT SHARE" and it anticipates a majority decision "SHARE", it places a higher bid than the potential bid. Conversely, when its potential decision is "SHARE" and it anticipates a majority decision "NOT SHARE", it conforms to majority to help others protect their privacy and places a lower bid than the potential bid. Otherwise (i.e., when its potential decision matches the expected majority decision), it opts for a low bid to conserve its budget. The decision function $\mathcal{F}$ of *ToM-1.P* is outlined in Algorithm 5.

Algorithm 5: Decision Function $\mathcal{F}$ of *ToM-1.P* Agent

Input: $\mathcal{P}(c)$ as own privacy preference for given content c, agents $\mathcal{A}$, agents' contents $\mathcal{C}'$,agents' preferences $\mathcal{P}'$
Output: sharing decision D and bid B

$\mathcal{D}' \leftarrow$ Run Algorithm 2 with $\mathcal{A}$, $\mathcal{C}'$, and $\mathcal{P}'$
$D,\ B \leftarrow$ Run Algorithm 1 with $\mathcal{P}(c)$
$m_S = \sum_{d \in \mathcal{D}'} \mathbb{1}_{d=SHARE}$ // count SHARE decisions
$m_{NS} = \sum_{d \in \mathcal{D}'} \mathbb{1}_{d=NOT\_SHARE}$ // count NOT_SHARE decisions
if $m_S > m_{NS}$ **and** $D = NOT\_SHARE$ **then**
  $b_l \leftarrow \max\{B, (b_{MAX} + b_{MID})/2\}$ // lower bound
  $B \leftarrow$ a random integer in the range $(b_l, b_{MAX}]$
else
  if $m_S < m_{NS}$ **and** $D = SHARE$ **then**
    $D \leftarrow NOT\_SHARE$ // decision update
  end
  $b_u \leftarrow \min\{B, (b_{MIN} + b_{MID})/2\}$ // upper bound
  $B \leftarrow$ a random integer in the range $[b_{MIN}, b_u)$
end

4.5 Learning Theory of Mind

Every *ToM-1* agent that are described in the previous subsections initially assumes that all other agents have the same preferences as itself. This means the agent applies ToM reasoning by projecting its own preferences onto others. While this simplifies the initial reasoning process, it may not always capture others' preferences as they may be different than the *ToM-1* agent's preferences. Let's explore an example to illustrate this.

Consider four agents deciding whether to share a group photo, one of whom is the *ToM-1.A* agent whereas the others are *ToM-0* agents. The *ToM-1.A* agent may be very likely to share it (i.e., rating "5" on the Likert scale) whereas other agents may be unlikely to share it (i.e., rating "2" on the Likert scale). If the *ToM-1.A* agent predicts the other agents' decisions based solely on its own preference, it will conclude that the majority also wants to share and make a decision in favor of sharing. Now, suppose that it places a bid of 4, conserving its budget, in favor of sharing while the other agents each place bids of 5 against sharing. The final group decision, in this case, would be against sharing, with a bid difference of 11. Since the *ToM-1.A* agent is designed to advocate for its preferences, and especially more when its decision differs from the majority, this outcome, based on an inaccurate model of others' preferences, represents a miscalculation that should have been avoided. Instead, the *ToM-1.A* agent could have placed a bid of 18 if it had a more accurate model of others and would have achieved its preferred outcome.

This example shows that it is unrealistic to expect that an agent's preferences will always match those of others, which can lead to unintended outcomes. Therefore, for the agent to make better predictions, it must have the ability to update and adjust its beliefs about others' preferences. Through repeated interactions and by learning from new information, the *ToM-1* agent can continuously refine its understanding of others, enabling it to make more accurate and informed decisions over time. With this in mind, we upgrade *ToM-1.M*, *ToM-1.A*, and *ToM-1.P* agents to *ToM-1.ML*, *ToM-1.AL*, and *ToM-1.PL*, respectively, by introducing capability of learning actual preferences of others through interactions. This learning takes its place in the agent's model as follows:

$$q' = \begin{cases} (1-\alpha) \times q + \alpha, & \text{if } d_i = SHARE \\ (1-\alpha) \times q, & \text{otherwise} \end{cases} \quad (1)$$

where α is the learning rate, q and q' are the sharing probabilities of others for a given content, before and after a new interaction respectively. The decision of another agent in that interaction is denoted with d_i. If that decision is "SHARE", then the agent increases the probability as the decision is inline with the agent's current belief. In the opposite case, the belief contradicts with the actual preference, hence, the probability is updated with a lower value. With the addition of probabilities, we also updated Algorithm 2 to Algorithm 6 which calculates others' expected decisions from their sharing probabilities (instead of assumed preferences).

Algorithm 6: Estimations for Other Agents' Decisions with Probabilities ($\mathcal{G}'$)

Input: agents $\mathcal{A}$, agents' contents $\mathcal{C}'$, Q as sharing probabilities for all contents
Output: agents' (expected) decisions $\mathcal{D}'$

```
r ← a random real number in the range of [0, 1]
for each agent a ∈ A do
    c ← C'(a)                          // get agent's content
    if Q(c) ≤ r then
        D'(a) ← NOT_SHARE
    else
        D'(a) ← SHARE
    end
end
```

4.6 *Base* Agent

To replicate the scenarios from our user study, we have also designed a *Base* agent, denoted as $Base = \langle \mathcal{P}, \emptyset, \mathcal{F}, \emptyset, \emptyset \rangle$. Similar to other agents, the *Base* agent makes decisions consistent with its preferences, yet it uses a deterministic bidding approach. Complete setting for its decision and bidding strategy is given in Table 1.

Table 1: *Base* agent's decision and bidding rules based on given preferences.

Preference	Decision	Bid
1	*NOT_SHARE*	15
2	*NOT_SHARE*	10
3	Randomly select *NOT_SHARE* or *SHARE*	5
4	*SHARE*	10
5	*SHARE*	15

5 EVALUATION

In our setup with four agents, one is designated as the "focal agent" and plays a simplified version of the game described in Section 3: instead of two rounds, each bidding decision is made in a single round to better reflect real-life one-shot interactions. The focal agent competes against three other *Base* agents, replicating the conditions of the user study. The focal agent's decision-making strategy is determined by its ToM type, as outlined in Section 4.

The agents and the simulation environment are implemented with Python [13]. The simulations are carried out using a workstation that has Intel® Core™ i7 2.50 GHz processors and 16 GB RAM with 64-bit Microsoft Windows 10 operating system.

The simulation involves 16 types of emojis as contents, with each agent having its own privacy preferences for these emojis. Privacy preferences are represented on a Likert scale from 1 to 5. The privacy preferences for the focal agent are initialized with a predefined list that differs from those assigned to the other agents. This distinction allows us to explore how the focal agent responds to varying privacy behaviors within the experiment. With this setup, the learning-enabled ToM agents can observe the decisions of other agents over time and adjust their own internal model of others' preferences. This dynamic learning process enables the focal agent to refine its predictions and make more informed decisions as interactions progress. We set the learning parameter α to 0.1 to ensure that the focal agent builds a stable model of the other agents' privacy preferences incrementally.

Each simulation runs for 1000 rounds, during which agents repeatedly decide whether to share a set of four emojis representing each agent. At the start of every round, each emoji are randomly picked from 16 available types (with the total number of possible sets being 16^4). The focal agent starts with a budget of 50 units and is allocated a fixed budget of 10 units per round. If the agent's decision matches the final group decision in a round, its budget is reduced by the bid amount it placed. Additionally, a tax is deducted for each round where the agent's bid directly determines the final group decision, as suggested in the Clarke-Tax mechanism [10]. We take this tax to be 5 units. In line with the setup of user study, bids are limited between 1 and 20. Hence, the parameters b_{MIN}, b_{MID} and b_{MAX} in the algorithms are set as 1, 10 and 20 respectively.

We measure several performance indicators during the simulations, including:

- Alignment Count: The number of rounds in which the focal agent's decision is the same as the final group decision.
- Spent Budget: The total amount spent by the focal agent on bids and taxes.
- Focal Agent's Privacy Violations: Occurrences where the focal agent's privacy is violated (i.e., focal agent's ToM-less decision, which reflects its original preference, is NOT SHARE but the final group decision is SHARE).

- Other Agents' Privacy Violations: Occurrences where at least one non-focal agent's privacy is violated (i.e., at least one other agent's decision is NOT SHARE but the final group decision is SHARE).

We evaluate the performance of *ToM-0*, *ToM-1.M*, *ToM-1.A*, *ToM-1.P*, *ToM-1.ML*, *ToM-1.AL*, and *ToM-1.PL* agents through simulations. These agents are assessed based on their alignment ratios, budget management, and privacy management in the simulation setting. We conduct 10 simulations for each agent, averaging the results. Tables 2 and 3 provide an overview of the performance of all agents.

Table 2: Average alignment counts and budget expenditures for different agents across 10 simulations, with each simulation running for 1000 rounds.

Agent Type	#Total Alignments	Spent Budget
ToM-0	607.2 ± 16.50	5022.0 ± 261.95
ToM-1.M	627.5 ± 13.31	**1805.7 ± 46.27**
ToM-1.A	710.2 ± 12.85	9022.4 ± 186.94
ToM-1.P	667.6 ± 7.64	5398.2 ± 290.83
ToM-1.ML	**868.4 ± 8.39**	2377.6 ± 40.50
ToM-1.AL	783.4 ± 11.89	8762.9 ± 256.91
ToM-1.PL	818.2 ± 11.12	5363.0 ± 254.53

5.1 Evaluation of Majority-Conforming Agents: *ToM-1.M* and *ToM-1.ML*

Majority-conforming agents deliberately align their decisions with others' preferences while conserving their budgets. The learning-enhanced *ToM-1.ML* agent proved highly effective, matching the final group decisions in 868.4 rounds on average compared to 627.5 for the non-learning *ToM-1.M* agent (with $p < .05$). Both agents maintained strong budget efficiency, with *ToM-1.M* spending only 1805.7 units on average and *ToM-1.ML* using 2377.6 units (Table 2). This reflects the *ToM-1.ML* agent's adaptability and strategic bid placement when learning is introduced.

Privacy preservation, however, proved to be a challenge for the *ToM-1.M* and *ToM-1.ML* agents. On average, the majority-conforming focal agent experienced 249.4 privacy violations without learning and 245.8 with learning (Table 3). Privacy violations for other agents were also high, with 375.6 and 368.1 violations, respectively. This shows that the majority-conforming strategy struggles to sufficiently prioritize privacy since it tends to reinforce majority decisions that may already exhibit privacy-violating behavior.

5.2 Evaluation of Preference-Advocating Agents: *ToM-1.A* and *ToM-1.AL*

Preference-advocating agents prioritize their own preferences, actively pushing for decisions that reflect them. They are highly competitive as *ToM-1.A* and *ToM-1.AL* have aligned with others in 710.2 and 783.4 rounds on average, respectively. However, this comes at a substantial cost, as both agents struggled to maintain a high remaining budget. Without learning, the focal agent spent an average of 9022.4 units per simulation, while the learning-enhanced version had a similarly high expenditure of 8762.9 units (Table 2).

Privacy management was slightly better for *ToM-1.A* and *ToM-1.AL* agents compared to *ToM-1.M* and *ToM-1.ML* agents. Without learning, the focal agent experienced 134.9 privacy violations on average, while learning reduced this number to 105.9 (significantly better with $p < .05$). However, violations for other agents remained steady at around 361.6 and 375.0, respectively (Table 3), which indicates that preference-advocating agents focus more on their own privacy than on protecting the privacy of others.

Table 3: Average numbers of privacy violations for the focal agent and other agents across 10 simulations, with each simulation running for 1000 rounds.

Agent Type	#Focal Agent's Privacy Violations	#Other Agents' Privacy Violations
ToM-0	190.1 ± 15.33	364.9 ± 11.09
ToM-1.M	249.4 ± 13.37	375.6 ± 12.89
ToM-1.A	134.9 ± 11.89	361.6 ± 15.81
ToM-1.P	144.0 ± 11.26	273.8 ± 12.93
ToM-1.ML	245.8 ± 13.02	368.1 ± 18.25
ToM-1.AL	**105.9 ± 8.95**	375.0 ± 18.95
ToM-1.PL	110.3 ± 9.13	**225.0 ± 15.01**

5.3 Evaluation of Privacy-Preserving Agents: *ToM-1.P* and *ToM-1.PL*

Privacy-preserving agents prioritize their own privacy as well as that of others.With learning, the *ToM-1.PL* agent achieved a higher alignment count of 818.2 compared to 667.6 for the *ToM-1.P* agent on average. Budget usage stayed nearly the same, with the learning-enhanced agent spending 5363.0 units per simulation and the non-learning version spending 5398.2 units. This suggests that the privacy-focused strategy balances costs and gains effectively.

Privacy violations were the lowest among all agents for *ToM-1.P* and *ToM-1.PL*, confirming the success of the privacy-preserving strategy. On average, The *ToM-1.P* agent experienced 144.0 violations and *ToM-1.P* experienced only 110.3. Other agents' privacy was also well-preserved, with the lowest recorded violations: 273.8 without learning and 225.0 with learning. These results highlight these agents' ability to preserve privacy, both for themselves and for others, while still achieving competitive results in the simulation.

5.4 Comparative Evaluation

We devised a metric called *Privacy-Oriented Alignment Rate (POAR)* to compare the agents. POAR is defined as the total number of alignments achieved by the agent over total number of privacy violations (i.e., the sum of the focal agent's privacy violations and other agents' privacy violations). This metric offers a general view of how efficiently each agent makes decisions that match with the group decisions while mitigating multi-party privacy conflicts. A higher value suggests that the agent aligns with the group decision more frequently with fewer privacy violations, indicating stronger privacy protection. Table 4 compares all agents based on this metric.

Starting with *ToM-0*, the agent has a relatively low POAR of 1.09 alignments per violation, reflecting its basic strategy. *ToM-1.M* has the lowest POAR at 1.00, showing that it frequently compromises

privacy in order to succeed, conforming more to group behavior than privacy concerns. On the other hand, *ToM-1.A* achieves a higher POAR of 1.43, indicating a more balanced approach to align with others while preserving privacy. *ToM-1.P*, designed to prioritize privacy, performs even better, achieving 1.60 alignments per privacy violation, emphasizing its privacy-preserving effectiveness.

For agents with learning, *ToM-1.ML* improves over *ToM-1.M* with a POAR of 1.41, reflecting better privacy management. *ToM-1.AL* has a solid score of 1.63, reflecting that learning enables more privacy-conscious decision-making. Finally, *ToM-1.PL* stands out as the best performer, with a POAR of 2.44, showing that its learning-enhanced privacy-preserving strategy is the most effective at balancing privacy protection with successful outcomes (significantly outperforms the closest agent *ToM-1.AL* with $p < .05$).

Table 4: Privacy-Oriented Alignment Rate (POAR) of different agents across 10 simulations, with each simulation running for 1000 rounds.

Agent Type	#Total Alignments	#Privacy Violations	Privacy-Oriented Alignment Rate
ToM-0	607.2 ± 16.50	555.0 ± 17.19	1.10 ± 0.06
ToM-1.M	627.5 ± 13.31	625.0 ± 20.24	1.00 ± 0.03
ToM-1.A	710.2 ± 12.85	496.5 ± 17.22	1.43 ± 0.07
ToM-1.P	667.6 ± 7.64	417.8 ± 21.45	1.60 ± 0.09
ToM-1.ML	868.4 ± 8.39	613.9 ± 27.05	1.42 ± 0.06
ToM-1.AL	783.4 ± 11.89	480.9 ± 18.84	1.63 ± 0.07
ToM-1.PL	818.2 ± 11.12	335.3 ± 21.36	**2.45 ± 0.19**

6 DISCUSSION AND FUTURE DIRECTIONS

This work investigates how ToM reasoning can address privacy conflicts in in multi-party privacy scenarios. Our proposed agents can reason about privacy preferences of others, allowing them to handle decision-making situations where individual privacy preferences clash. Our experimental setup show that ToM significantly enhances privacy preservation, especially when agents can adapt their behaviors based on assessments of others' privacy preferences.

Multi-party conflicts have been studied in privacy context for some time. Thomas *et al.* [38] define *multi-party privacy conflicts* as the disagreement of users over the privacy of the data that pertains to all of them. Squicciarini *et al.* [33] propose to use an auction-based mechanism to resolve such conflicts, where each user bids for an amount that reflects how much they would like to see their preferences as the group-decision. Such and Criado [35, 36] suggest a software where the auctions are conducted by the agents representing the users, reducing the need for human intervention. Ulusoy and Yolum [39] propose, PANO, that uses Clarke-Tax mechanism in multi-party privacy decisions with certain limitations to avoid manipulations in the auctions. In their follow up study, PANOLA [40] extends this mechanism to design agents that can learn to bid effectively for different types of users. Ajmeri et al. [2] propose privacy-aware personal agents that are able to incorporate social norms unlike traditional agent-oriented software engineering methods. Mosca et al. [26] address the issue with a utility and value-driven approach which aims to meet explainability, role-agnosticism and adaptability requirements at the same time. In their follow-up study [27], they provide empirical proofs to show such agents outperform the existing methods in terms of balancing the individual utility and value adherence along with users' satisfaction for the explainability of the system. None of these aforementioned studies investigates the problem by introducing agents that are equipped with ToM.

Many computational ToM models have been developed recently to assess their effectiveness in various contexts [6, 14, 15, 19, 20, 28, 29, 42, 44]. We discuss two specific studies in relation to our work.

Montes et al. [25] develop a BDI agent architecture that can also do ToM reasoning using abductive reasoning. They demonstrate that their agents can interact with a variety of agents, including ones that do not have a ToM. Their experimental evaluation show that agents with a ToM reasoning perform significantly better than those that do not have one in the game of Hanabi [7]. While they evaluate their agent in a competitive setting, we study how an agent's ability to have ToM reasoning helps a society of agents as a whole for dealing with multi-party conflicts.

De Weerd et al. [12] investigate how higher-order ToM reasoning benefits agents in unpredictable negotiation environments, using the Colored Trails setting. The study reveals that higher-order ToM agents outperform lower-order agents, particularly in dynamic, less predictable environments, where agents without ToM struggle to predict the behavior of others. In their setting, agents negotiate resources in a mixed-motive environment involving both cooperation and competition, similar to our setting, where our agents, inspired by insights from a user study, must balance self-interest with others' preferences for effective overall privacy preservation. In both cases, ToM allows agents to better predict others' actions.

One limitation of our study is the non-focal agents' basic bidding behavior. The base agent uses a simple strategy with a deterministic bidding approach that replicates the scenarios from our user study well; however, they may not fully reflect the complexity of real-world decision-making where agents may employ more sophisticated, adaptive approaches. Moreover, our focal agents rely on a single ToM model to represent all other agents' preferences. This resembles representing privacy preferences as norms of the society as a whole [3]. While this simplifies the implementation, it may reduce the accuracy of the model, as each other agent can have diverse set of privacy expectations. Using separate ToM models for each agent and possibly integrating a multi-dimensional model [45] would enable a more fine-grained representation of others' preferences. For future work, we will focus on developing more sophisticated bidding behaviors and implementing separate ToM models for each agent to better reflect the complexity and diversity of real-world decision-making.

ACKNOWLEDGMENTS

This research is financially supported by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research (grant number 024.004.022). Hüseyin Aydın is supported by the Scientific and Technological Research Council of Turkey, through BIDEB 2219 International Postdoctoral Research Scholarship Program.

REFERENCES

[1] Nirav Ajmeri, Hui Guo, Pradeep K Murukannaiah, and Munindar P Singh. 2020. Elessar: Ethics in Norm-Aware Agents. In *Proceedings of the 19th International Conference on Autonomous Agents and Multi-Agent Systems*. 16–24.

[2] Nirav Ajmeri, Pradeep K Murukannaiah, Hui Guo, and Munindar P Singh. 2017. Arnor: Modeling social intelligence via norms to engineer privacy-aware personal agents. In *Proceedings of the 16th International Conference on Autonomous Agents and Multi-Agent Systems*. 230–238.

[3] Marc Serramia Amoros, William Seymour, Natalia Criado, and Michael Luck. 2023. Predicting Privacy Preferences for Smart Devices as Norms. In *The 22nd International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS).

[4] Reyhan Aydoğan, Özgür Kafali, Furkan Arslan, Catholijn M Jonker, and Munindar P Singh. 2021. Nova: Value-based negotiation of norms. *ACM Transactions on Intelligent Systems and Technology (TIST)* 12, 4 (2021), 1–29.

[5] Hüseyin Aydın, Onuralp Ulusoy, Ilaria Liccardi, and Pınar Yolum. 2025. Analyzing Privacy Dynamics within Groups using Gamified Auctions. arXiv:2502.10788 [cs.HC] https://arxiv.org/abs/2502.10788

[6] Chris L Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B Tenenbaum. 2017. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour* 1, 4 (2017), 1–10.

[7] Nolan Bard, Jakob N Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, et al. 2020. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence* 280 (2020), 103216.

[8] Anat Bardi and Shalom H Schwartz. 2003. Values and behavior: Strength and structure of relations. *Personality and Social Psychology Bulletin* 29, 10 (2003), 1207–1220.

[9] Peter Carruthers and Peter K. Smith (Eds.). 1996. *Theories of Theories of Mind*. Cambridge University Press.

[10] Edward H Clarke. 1971. Multipart pricing of public goods. *Public Choice* (1971), 17–33.

[11] Stephen Cranefield, Michael Winikoff, Virginia Dignum, and Frank Dignum. 2017. No Pizza for You: Value-based Plan Selection in BDI Agents. In *Twenty-Sixth International Joint Conference on Artificial Intelligence*. 178–184.

[12] Harmen De Weerd, Rineke Verbrugge, and Bart Verheij. 2022. Higher-order theory of mind is especially useful in unpredictable negotiations. *Autonomous Agents and Multi-Agent Systems* 36, 2 (2022), 30.

[13] Emre Erdogan and Hüseyin Aydın. 2024. Mitigating Privacy Conflicts via Computational Theory of Mind - Experiment Page. https://git.science.uu.nl/e.erdogan1/tom-project-privacy. Accessed on: 2025-02-03.

[14] Emre Erdogan, Frank Dignum, and Rineke Verbrugge. 2024. Effective Maintenance of Computational Theory of Mind for Human-AI Collaboration. In *HHAI 2024: Hybrid Human AI Systems for the Social Good*. IOS Press, 114–123.

[15] Emre Erdogan, Frank Dignum, Rineke Verbrugge, and Pinar Yolum. 2025. TOMA: Computational Theory of Mind with Abstractions for Hybrid Intelligence. *Journal of Artificial Intelligence Research* 82 (2025), 285–311.

[16] Dorota Filipczuk, Tim Baarslag, Enrico H Gerding, and MC Schraefel. 2022. Automated privacy negotiations with preference uncertainty. *Autonomous Agents and Multi-Agent Systems* 36, 2 (2022), 49.

[17] Batya Friedman, Peter H Kahn, Alan Borning, and Alina Huldtgren. 2013. Value sensitive design and information systems. *Early engagement and new technologies: Opening up the laboratory* (2013), 55–95.

[18] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. 2016. Cooperative inverse reinforcement learning. *Advances in neural information processing systems* 29 (2016).

[19] Lasse Dissing Hansen and Thomas Bolander. 2020. Implementing theory of mind on a robot using dynamic epistemic logic. In *Twenty-Ninth International Joint Conference on Artificial Intelligence*. 1615–1621.

[20] Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua B Tenenbaum, and Tianmin Shu. 2024. MMToM-QA: Multimodal Theory of Mind Question Answering. arXiv:2401.08743 [cs.AI]

[21] 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* 18, 4, Article 41 (2018), 21 pages.

[22] Dilara Kekulluoglu, Nadin Kökciyan, and Pınar Yolum. 2018. Preserving Privacy as Social Responsibility in Online Social Networks. *ACM Transactions on Internet Technology* 18, 4, Article 42 (2018), 22 pages.

[23] Nadin Kökciyan, Nefise Yaglikci, and Pınar Yolum. 2017. An argumentation approach for resolving privacy disputes in online social networks. *ACM Transactions on Internet Technology* 17, 3, Article 27 (2017), 22 pages.

[24] Enrico Liscio, Michiel van der Meer, Luciano C Siebert, Catholijn M Jonker, and Pradeep K Murukannaiah. 2022. What values should an agent align with? An empirical comparison of general and context-specific values. *Autonomous Agents and Multi-Agent Systems* 36, 1 (2022), 23.

[25] Nieves Montes, Michael Luck, Nardine Osman, Odinaldo Rodrigues, and Carles Sierra. 2023. Combining theory of mind and abductive reasoning in agent-oriented programming. *Autonomous Agents and Multi-Agent Systems* 37, 2 (2023), 36.

[26] Francesca Mosca, Jose Such, and Peter McBurney. 2020. Towards a value-driven explainable agent for collective privacy. In *Proceedings of the 19th International Conference on Autonomous Agents and Multi-Agent Systems*.

[27] Francesca Mosca and Jose M Such. 2021. ELVIRA: An Explainable Agent for Value and Utility-Driven Multiuser Privacy. In *Proceedings of the 20th International Conference on Autonomous Agents and Multi-Agent Systems*. 916–924.

[28] Dung Nguyen, Phuoc Nguyen, Hung Le, Kien Do, Svetha Venkatesh, and Truyen Tran. 2022. Learning Theory of Mind via Dynamic Traits Attribution. In *Proceedings of the 21st International Conference on Autonomous Agents and Multi-Agent Systems*. 954–962.

[29] Nancirose Piazza and Vahid Behzadan. 2023. A Theory of Mind Approach as Test-Time Mitigation Against Emergent Adversarial Communication. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multi-Agent Systems*. 2842–2844.

[30] David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 1, 4 (1978), 515–526.

[31] Shalom H Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, et al. 2012. Refining the theory of basic individual values. *Journal of Personality and Social Psychology* 103, 4 (2012), 663.

[32] Marc Serramia, Manel Rodriguez-Soto, Maite Lopez-Sanchez, Juan A Rodriguez-Aguilar, Filippo Bistaffa, Paula Boddington, Michael Wooldridge, and Carlos Ansotegui. 2023. Encoding ethics to compute value-aligned norms. *Minds and Machines* 33, 4 (2023), 761–790.

[33] Anna Cinzia Squicciarini, Mohamed Shehab, and Federica Paci. 2009. Collective privacy management in social networks. In *Proceedings of the 18th International Conference on World Wide Web*. 521–530.

[34] Sebastian Stein and Vahid Yazdanpanah. 2023. Citizen-Centric Multiagent Systems. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multi-Agent Systems*. 1802–1807.

[35] Jose M Such and Natalia Criado. 2016. Resolving multi-party privacy conflicts in social media. *IEEE Transactions on Knowledge and Data Engineering* 28, 7 (2016), 1851–1863.

[36] Jose M Such and Natalia Criado. 2018. Multiparty privacy in social media. *Commun. ACM* 61, 8 (2018), 74–81.

[37] Jose M Such and Michael Rovatsos. 2016. Privacy policy negotiation in social media. *ACM Transactions on Autonomous and Adaptive Systems* 11, 1 (2016), 1–29.

[38] Kurt Thomas, Chris Grier, and David M Nicol. 2010. unfriendly: Multi-party privacy risks in social networks. In *Privacy Enhancing Technologies: 10th International Symposium, PETS 2010, Berlin, Germany, July 21-23, 2010. Proceedings 10*. Springer, 236–252.

[39] Onuralp Ulusoy and Pınar Yolum. 2018. PANO: Privacy Auctioning for Online Social Networks. In *Proceedings of the 17th International Conference on Autonomous Agents and Multi-Agent Systems* (Stockholm, Sweden). 2103–2105.

[40] Onuralp Ulusoy and Pınar Yolum. 2021. PANOLA: A Personal Assistant for Supporting Users in Preserving Privacy. *ACM Transactions on Internet Technology* 22, 1, Article 27 (2021), 32 pages.

[41] Ibo Van de Poel. 2013. Translating values into design requirements. *Philosophy and engineering: Reflections on practice, principles and process* (2013), 253–266.

[42] Alan F. T. Winfield. 2018. Experiments in Artificial Theory of Mind: From Safety to Story-Telling. *Frontiers in Robotics and AI* 5 (2018), 75.

[43] Jessica M Woodgate and Nirav Ajmeri. 2022. Macro ethics for governing equitable sociotechnical systems. In *Proceedings of the 21st International Conference on Autonomous Agents and Multi-Agent Systems*. 1824–1828.

[44] Haochen Wu, Pedro Sequeira, and David V Pynadath. 2023. Multiagent Inverse Reinforcement Learning via Theory of Mind Reasoning. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multi-Agent Systems*. 708–716.

[45] Yaqing Yang, Tony W Li, and Haojian Jin. 2024. On the Feasibility of Predicting Users' Privacy Concerns using Contextual Labels and Personal Preferences. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.