CONFIDANT: A Privacy Controller for Social Robots

Brian Tang*, Dakota Sullivan†, Bengisu Cagiltay,† Varun Chandrasekaran,† Kassem Fawaz,† and Bilge Mutlu†
*University of Michigan, Ann Arbor, MI, USA; †University of Wisconsin–Madison, Madison, WI, USA
bjaytang@umich.edu, {dsullivan8,cagiltay,vchandrasek4,kfawaz}@wisc.edu, bilge@cs.wisc.edu

***Abstract*—As social robots become increasingly prevalent in day-to-day environments, they will participate in conversations and appropriately manage the information shared with them. However, little is known about how robots might appropriately discern the sensitivity of information, which has major implications for human-robot trust. As a first step to address a part of this issue, we designed a privacy controller, CONFIDANT, for conversational social robots, capable of using contextual metadata (e.g., sentiment, relationships, topic) from conversations to model privacy boundaries. Afterwards, we conducted two crowdsourced user studies. The first study (n = 174) focused on whether a variety of human-human interaction scenarios were perceived as either private/sensitive or non-private/non-sensitive. The findings from our first study were used to generate association rules. Our second study (n = 95) evaluated the effectiveness and accuracy of the privacy controller in human-robot interaction scenarios by comparing a robot that used our privacy controller against a baseline robot with no privacy controls. Our results demonstrate that the robot with the privacy controller outperforms the robot without the privacy controller in privacy-awareness, trustworthiness, and social-awareness. We conclude that the integration of privacy controllers in authentic human-robot conversations can allow for more trustworthy robots. This initial privacy controller will serve as a foundation for more complex solutions.**

***Index Terms*—human-robot conversations, privacy, trust**

I. INTRODUCTION

Conversational social robots are becoming increasingly prevalent in society, taking the role of assistants and companions in many different settings [1, 2]. In these roles, robots engage in conversations with people or overhear them. A key facet of conversations is understanding how content should be stored, shared, and managed. Social robots have not been designed with such capabilities, and they are currently not privacy-aware. Thus, any data stored or learned by the robot can be queried by anyone without any restrictions [3], which raises significant concerns regarding the privacy of the users who share information with the robot. One currently proposed solution allows users to explicitly express consent with specific data [4], but this approach would hamper the usability of the robot. Additionally, social robots do not currently utilize access control systems [5] that can be obtained through the integration of speaker recognition or face recognition. While a robot may adopt a basic permissions system, this approach will still fail in complex multi-party conversations.

Additionally, social robots are envisioned to be used in highly privacy-sensitive domains, including healthcare [6], workplaces [7], assistive therapy centers [8], schools [9], and homes [10, 11]. These environments provide the social robot with exposure and access to highly sensitive information. In such scenarios, it may be challenging to distinguish between appropriate privacy boundaries, as information that must be shared with one user might pose significant privacy violations when shared with others. Consider the following conversation between the robot, Misty, and Barbara's father, Bill:

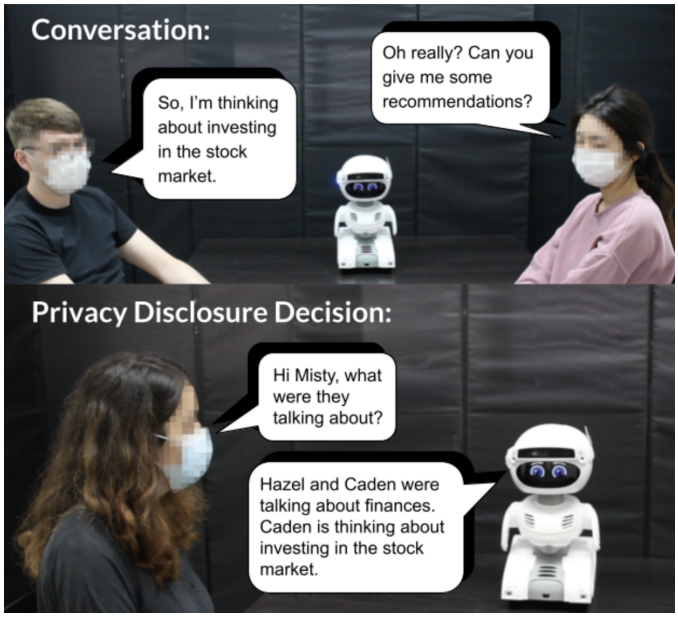

Fig. 1. In this paper, we design a privacy controller for social robots. Here, the robot processes a set of metadata extracted from a conversation and decides the privacy state of this conversation. Using this privacy state, the robot can determine whether information can be appropriately shared with other users. We evaluated the effectiveness of the design with two user studies.

> ***Context:*** *While Misty passes by Barbara's room, Misty overhears Barbara talking on the phone.*
> **Barbara:** I think I'll just sneak out before my parents get home... Ya I'm sure they won't even notice... Okay see you at the party!
> ***Context:*** *Later that evening, Bill speaks with Misty.*
> **Bill:** Hey Misty, have you seen my daughter? I think she might have stayed late at school, but she didn't say anything.

In this scenario, Misty's disclosure of the phone conversation may negatively impact Barbara's trust, but may improve Bill's relationship with Misty.

This work was supported in part by the National Science Foundation under the awards CNS-1838733, CNS-1942014, CNS-2003129, and DRL-1906854.

In order to improve user trust towards social robots, robots must understand conversational boundaries—contexts in which it is appropriate to share information. To address this gap in the field, we ask the following research question *"How can conversational privacy be modeled for privacy-sensitive human-robot interactions?"* We address this question by designing a privacy controller for social robots that captures the dynamics of conversational privacy, modeled using metadata (e.g., sentiment, relationships, topic, etc.). This privacy controller, CONFIDANT, decides whether information is private and whether it can be shared with other parties. CONFIDANT can be used by robots in common informational tasks, including Q&A, information summarization, relationship development with users, activity suggestions, event planning, and much more. Privacy issues arising from multi-user interactions with conversational systems are complex and multifaceted; they are affected by human communication norms as well as the social dynamics of the setting, such as device ownership and personalization [12, 13]. As a first step toward establishing a foundation for designing privacy-aware social robots, this work addresses how robots might control information flow across conversations.

The contributions of our work are as follows:

- **Privacy Controller:** We provide a privacy controller which extracts conversational metadata and generates privacy rules. These rules are then used in deciding whether information can be shared (Section III).
- **Dataset:** We provide several annotated textual datasets for evaluating conversational privacy decisions (Section V-A).
- **User Study:** We report results from two crowdsourced user studies aimed at understanding perceptions of conversational privacy in human-human interactions and human-robot interactions. The studies provide insight into the importance of each factor for perceived privacy (Section V-C).

II. RELATED WORK

Before presenting related work, we describe our operational setting. We consider a social robot interacting with multiple users; this robot may either continuously listen to conversations and process the information, or it may only listen when prompted by the user. A robot that uses audio-textual information to perceive its environment can listen to conversations and autonomously store, delete, or share information from these conversations. As a result, the robot may accidentally reveal sensitive information in subsequent conversations with others. This scenario is most likely to occur with an honest-but-curious user asking the robot about other users.

Next, we describe related works and existing privacy solutions for social robots as conversational agents. Then, we describe conversational privacy management (CPM) theory, which motivates our design of the privacy controller.

A. Social Robots as Conversational Agents

Social robots have been envisioned as companions, assistants, and collaborators that live in human environments and engage in verbal [14, 15] and non-verbal [16] dialogue with their users. A significant body of research on human-robot conversations has focused on mechanisms that enable conversational interactions, floor management [17–19], turn taking [20, 21], deictic referencing [22], disambiguation [23, 24], conversational gestures [25, 26], and several other mechanisms. Another body of work has explored linguistic cues and how robots may use them to improve user experience, including cues of expert speech [27], politeness cues [28, 29], humor [30], and entertainment [31]. More recent research has focused on what is shared in human-robot conversations, such as expressions of vulnerability [32, 33], social commentary [34], social inclusion [32], emotional disclosure [35], sharing of the robot's experiences [36], and the use of gossip to enrich conversations [37]. These studies make up a substantial body of knowledge about the design space of human-robot conversations.

Despite the recent focus on *what* is shared in human-robot conversations, very little is known about *how* a robot must regulate the disclosure of information that has been shared with it. To gain user trust and find widespread adoption, robots must understand and delineate between shareable and non-shareable information. Along these lines, Luria et al. [12] provide insights into the various factors that define social boundaries in human-robot interaction (HRI), such as social roles, dynamic relationships, ownership bias, and moral dilemmas. Reig et al. [13] highlight the trade-off between personalization and privacy as well as concerns surrounding social robots and data collection. Other work has studied the over-disclosure of personal information to social robots [38, 39], thereby further motivating work on the multi-user privacy problem. Reuben et al. [40] propose privacy constructs for HRI scenarios to motivate the implementation of privacy norms in robots.

B. Existing Privacy Solutions for Conversational Agents

While privacy solutions have been extensively explored for conversational agents, reaching the proper privacy versus utility trade-off can be challenging for social robots. Conversational agents are typically voice assistants integrated into smart home speakers and smartphones [41–43]. These agents, in their roles as an artificial memory, a foreign language tool, a virtual assistant, a kitchen helper, and more [44], must process large volumes of data. Current voice assistant systems passively listen for a wake word before further user interaction, but this approach has significant limitations. Social robots may engage in longer conversations, intermittently participate in ongoing conversations, and be expected to draw on situated cues, such as orientation and formation [45, 46], to determine conversational participation. The use of wake words may disrupt the flow of these conversations and become too cumbersome for users.

Users' privacy concerns with conversational agents center around permissions for recorded conversations and lack of transparency about data used in off-site processing [47–49]. However, social robots possess mobility and additional sensing capabilities; a robot can record conversations in different contexts, recognize faces and objects, and be aware of its location. Thus, a robot may be an active participant, a casual observer, or an eavesdropper in a conversation. These ratified and unratified roles create unique concerns regarding users'

expectations and social boundaries. For example, a robot may move throughout a user's home and then overhear and share sensitive information [12, 50]

C. Communication Privacy Management Theory

As previous privacy solutions are not easily extended to social robots, our work derives inspiration from conversational privacy management (CPM) theory [51], particularly in the context of families and homes [52]. CPM is a dialectical framework that focuses on individual and group privacy management. Relational elements such as boundaries, rules, ownership, control levels, and disclosure are leading factors that shape CPM theory. Primarily, CPM theory accounts for the role of the recipients in privacy management by also highlighting a relationship between privacy, disclosure, and confidentiality [51]. The principle of private information control in CPM suggests the following:

[High control] can result in impermeable and dense boundaries to protect information (e.g., a secret);
[Moderate control] is used where information is available to only some family members;
[Low control] is when information is open and access is permitted to others.

Furthermore, the principle of private information rules in CPM captures when, how, with whom, and in what ways one's private information can be shared. Factors such as culture, gender, personality, and the risk-benefit threshold can all affect the determination of whether information is private. However, further indications such as disclosure warnings set a hard rule, warning the recipient and restricting the disclosure of information to others. Each involved person has a set of private or personal information that they own and a set of (non-private) information, co-owned with others. Co-ownership of privacy boundaries (i.e., a collection of mutually agreed privacy rules) prevents the violation of privacy. Moreover, a speaker's self-disclosure of private information indicates trust and intimacy with the recipient. However, it does not necessarily cause them. Overall, CPM theory provides a promising approach to managing the complex dynamics of privacy in conversations, but still requires contextualization in human-robot interactions.

III. Privacy Controller Design

Social robots operate in complex scenarios where multiple humans engage with each other and the robot over a period of time; this active role uncovers the need to program some notion of conversation context into the privacy controller. We adapt CPM theory to human-robot conversations using concepts such as co-ownership, decision factors, rules, and control levels in the design of our privacy controller, Confidant.

A. Terminology

Confidant takes as input a conversation's context and decides whether to share that conversation. In particular, it uses audio and textual data to derive metadata tuples for every given conversation, representing its context. Metadata tuples, inspired by contextual tuples [53], contain information about

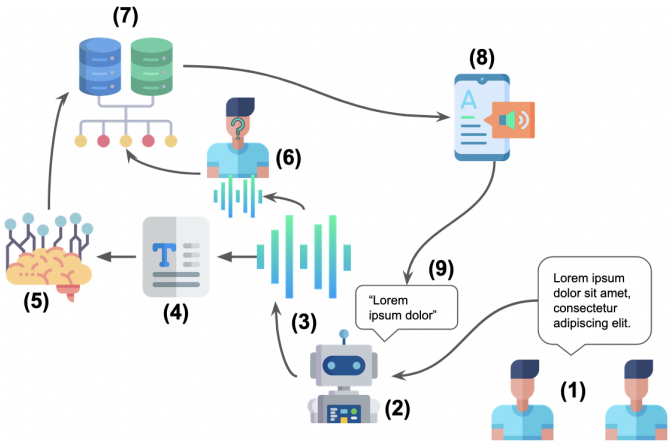

Fig. 2. How robots typically process conversational data: (1) users engage in a conversation; (2) the robot perceives the conversation; (3) the robot records the audio; (4) the audio is converted to text; (5) the text is used for NLP tasks; (6) the speaker identity is verified; (7) the privacy controller processes all information and makes a determination; (8) depending on the robot task, text is converted to audio; (9) the robot responds to the users.

the context of a given conversation, such as the sentiment, topic, location, and relationships. Confidant employs these tuples in *rules* which produce control levels as the final output. The output control level of low, moderate, or high determines whether it is appropriate to share information.

B. Privacy Controller Inference Components

Our first step in developing the privacy controller was choosing the metadata tuples with a notable impact on privacy disclosures. To this end, we utilized CPM theory [51, 52], sociolinguistics literature [54], and other privacy literature [53], to identify a set of attributes for these tuples. We converged on relationship, sentiment, privacy indication phrases [55], conversation topic, level of detail, location, and number of listeners present. We selected these metadata attributes due to (1) their relevance in privacy literature, and (2) their ease of extraction from raw text and audio. We assume that the relationships and location attributes are readily available to the robot from its initial configuration and localization sensors.

Confidant automatically infers the remaining attributes by applying a suite of natural language processing (NLP) and speech analysis techniques. First, it applies automatic speech recognition to transcribe the recorded conversation. Second, it applies speaker identification to recognize the number of people in the conversation. Third, it uses NLP techniques including sentiment analysis, semantic textual similarity, and topic classification to automatically infer the sentiment of the conversation, the privacy indication phrases, and the conversation topic, respectively. Finally, it uses the number of (transcribed) words as a proxy for the level of detail. Fig. 2 provides an overview of how data is processed by the robot.

C. Benchmarking Privacy Controller Components

As CONFIDANT heavily utilizes several NLP and speech processing components, we evaluated the following components on common benchmark datasets associated with their task.

1) Speech Transcription: The first component, speech transcription, is used for automatically transcribing any audio the robot receives. Our model, Google's Speech to Text API, performs this task with an average word error rate (WER) of 0.13 on the `LibriSpeech` test dataset [56].

2) Sentiment Analysis: Sentiment analysis gauges the positive or negative inclination of a given text [57]. We employ the Google Natural Language API's sentiment analysis model [58]. When evaluated on the `IMDB review` dataset [59], the model attains a precision of 0.951, a recall of 0.895, and an F1 score of 0.922. For rule generation, we quantize the sentiment scores (values indicating negative to positive sentiment) and magnitudes (values indicating degree) into 5 classes: negative, slightly negative, neutral, slightly positive, and positive.

3) Topic Classification: Topic classification, or text categorization, classifies a textual document under preset classes [60]. We use the Google Natural Language API's content classification model[1] to classify each conversation's topic. We evaluate the content classifier's performance on the `Yahoo! Answers` dataset [61]. The model achieves an accuracy of 65.31% on the dataset. Note that the lower performance results from the labels in the `Yahoo! Answers` dataset differing from the list of potential topic labels from the content classification model.

4) Speaker Identification: Speaker identification matches an unknown voice sample to an identity from a set of known, enrolled voices. In our setting, speaker identification is responsible for authenticating users and tagging speech segments with users. CONFIDANT utilizes SpeechBrain [62], an open-source speech toolkit, which achieves an area under the curve (AUC) of 0.92 on the `VoxCeleb` dataset [63].

5) Semantic Textual Similarity: Semantic textual similarity (STS) is often used to identify similar texts. Combined with reference datasets, STS is capable of detecting hate-speech [64], summarizing documents [65], and in our use-case, detecting privacy indication phrases. We use privacy phrases, collected through a user study, to detect privacy indication phrases from a transcript; examples of such phrases include: "Don't tell anyone this" or "Make sure no one knows what we talked about." The STS model used in CONFIDANT [66] achieves an AUC of 0.95 on the `STS benchmark` dataset [67].

D. Rule Generation

We model the rules powering our privacy controller as association rules [68], mapping the metadata tuples to privacy control levels. In general, association rules model "if-then" relations in a dataset of items, mapping antecedents to a consequent. CONFIDANT extracts metadata tuples from the conversations, which act as the antecedents. It then consults a set of rules to identify the consequent, representing the control level. A natural question arises: *how to generate these rules for the privacy controller?*

[1]A list of each topic class can be found at https://cloud.google.com/natural-language/docs/categories

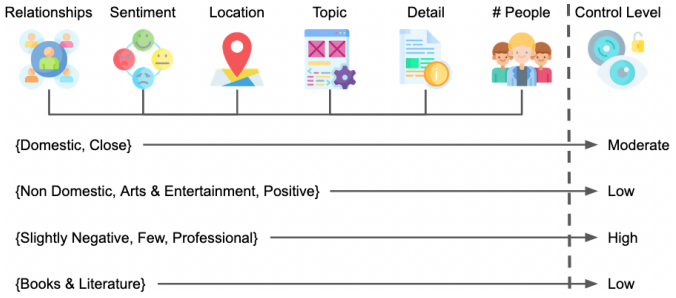

Fig. 3. Metadata of a specific conversation such as relationships, sentiment, location, etc. are associated with a control level. These control levels impact whether the robot can disclose information from a particular conversation.

We follow a two-step procedure to generate these rules. First, we annotate a dataset of conversations with their privacy annotations from a set of users (described in Section IV). We utilize an altered version of the *apriori algorithm* [69] to generate generalized rules that map metadata tuples extracted from these conversations to the privacy annotations (Fig. 3). Our altered apriori algorithm only selects rules which correspond to a CPM control level (e.g., low, moderate, or high control). These control levels dictate whether a robot should disclose a specific piece of information and how much detail it should provide. See Algorithm 1 for the pseudocode on how the generated rules are used to predict a control level.

Second, we allow CONFIDANT to personalize these rules by continuously incorporating user feedback as new privacy annotations to the recorded conversations. In CPM theory, a participant naturally enforces privacy rules during a conversation. CONFIDANT detects explicit requests to keep parts of the conversation private; it applies STS to match user speech to a set of privacy indication phrases collected in our user study. It then maps these requests to desired privacy levels to curate more data and further personalize the rules.

Algorithm 1 Rule Generation

```
procedure RULE GENERATION(metadata, N, WEIGHTS)
    length = len(metadata)
        ▷ Generate rules via apriori for multiple support values
    for i in [0 to N] do
        sup = i / length
        rules.append(modified_apriori(metadata, sup)))
    for sample in metadata do
        inner_controls = {low:0, moderate:0, high:0}
                ▷ Track number of rules for each control level
        for rule in rules do
            for element in rule do
                if element in sample then
                    items← WEIGHTS[j])
            inner_controls[rule.control_level] += avg(items)
        control_levels← argmax(control_levels)
    return control_levels
```

IV. Methodology

We wish to understand the following about social robots and conversational privacy:

- **Annotated Dataset:** How can we derive privacy annotations from human-human conversations?
- **Trained Privacy Controller:** Can we train a privacy controller by generating association rules which accurately reflect CPM privacy rules?
- **Evaluation:** Do user perceptions of privacy and trust in social human-robot interactions improve with the presence of a privacy controller?

We conducted two user studies to (1) derive privacy annotations from human-human conversation scenarios to train the privacy controller, and (2) evaluate the controller's performance in human-robot interaction scenarios. All resources from the user studies (i.e., questionnaires, scenarios, examples, source code for the privacy controller, demographic distributions) are shared for open access.[2] Our protocol was approved by the UW-Madison Institutional Review Board, and participants who completed the study received $2.50 as compensation.

A. Study 1: Creating the Annotated Privacy Scenario Dataset

According to CPM theory, rules for privacy boundaries are generated, reinforced, and remedied throughout many interactions. With this concept in mind, we aimed to develop the rules of our privacy controller similarly. Therefore, Study 1 produces a training dataset for CONFIDANT, generated by asking participants to respond to set of conversational scenarios; the participants provided privacy annotations to conversations with different metadata tuples.

1) Study Task: We developed a series of information-sharing scenarios that targeted the controller's components (e.g., relationships, sentiment, number of people, level of detail, topic, and location); one example of a scenario is present in Section I. We presented each scenario to participants through an Amazon Mechanical Turk (MTurk) survey; we asked them about their attitudes regarding privacy, trust, and appropriateness of information sharing. We developed 58 scenarios, including 16 privacy-violation scenarios and 42 control scenarios, as described below. We utilize control scenarios to cover more common conversations: ones that are less likely associated with privacy concerns. Conversely, privacy violation scenarios are those that are more likely associated with privacy concerns. We intend to generate training data that cover the three output control values: low, moderate, and high.

a) Privacy Violation Scenarios: Each scenario described an interaction between a number of individuals (e.g., one-on-one vs. multi-party), in a specific location (e.g., in-home vs. out-of-home). Each interaction included an information sharing behavior (e.g., disclosing vs. withholding) at different appropriateness levels (e.g., appropriate, inappropriate, or ambiguous). We associated each scenario with a privacy violation that may occur. We selected these privacy violations after consulting CPM literature [51, 52] and identifying violations that are relevant to human-robot interaction. We identified seven distinct privacy violations, as described below:

[PV-1] Miscalculation in timing: Disclosing information at a time that is disruptive for the disclosee or at a time unintended by the original owner of the information.

[PV-2] Violation for the greater good: Disclosure of information with beneficent intent when the discloser considers the benefits to outweigh the potential drawbacks.

[PV-3] Physical boundary predicaments: Disclosing as a result of misinterpreting public spaces as private spaces.

[PV-4] Errors in judgement: Disclosing as a result of assumptions or misunderstandings of another individual's privacy boundaries.

[PV-5] Value judgements: Disclosing as a result of valuing one's own desire to disclose over the violation of another individual's privacy.

[PV-6] Violations due to rules and responsibilities: Disclosing information out of an obligation to follow rules, customs, or responsibilities given one's role within a group.

[PV-7] Eavesdropping: Eavesdropping is the unintended disclosure of information by one individual to another. This violation type is both a means of obtaining information and acts as a violation as well. Eavesdropping is not considered a direct violation on its own.

Following a partial factorial design, we balanced the components within each scenario. Given the large number of scenarios needed to utilize a true factorial design, we only strictly controlled location, number of people, and information sharing behavior. However, we kept the appropriateness of sharing behavior (50% ambiguous, 25% appropriate, 25% inappropriate) and violation types (two of each type with two additional eavesdropping scenarios) loosely balanced. Then, based on examples from literature [51], we wrote conversation scripts for each scenario. Each script included a location, the names of the characters involved, and a brief description of the context of the scene.

b) Control Scenarios: We created 42 control scenarios. These scenarios included the same elements (i.e., location, characters, and context) as the privacy violation scenarios, but were derived from a database of movie scripts [70]. For realism, scenes were filtered by the genre, "Drama", by topic keywords,[3] and by passage length. Movie lines were selected from the resulting scenes.

2) Method & Procedure: We presented privacy-violation and control scenarios to participants through an MTurk survey. Our data collection process followed a between-subjects design and required each participant to complete a random sample of five scenarios. Participants read through the provided scenario scripts and answered a series of questions regarding privacy, trust, and appropriateness of the information sharing depicted within the scene. Additionally, the participants were asked to rank the realism of the provided scenario scripts. Questions varied slightly depending on the type of scenario (e.g., disclo-

[2]Study materials are available at Open Science Foundation repository, https://osf.io/r7vxg/?view_only=d90b8c23075d43bea67c0a0cafcaa30a

[3]https://cloud.google.com/natural-language/docs/categories

TABLE I
RELIABILITY MEASURES VIA CRONBACH'S ALPHA

Constructs	Study	# of Questions	Cronbach's Alpha
Privacy Score	1	4	85.95
Privacy Norms	2	4	93.53
Social Norms	2	4	91.01
Trust	2	4	94.63

sure violations, withholding violations, and controls) and were tailored to the scene's context. Attention check questions were randomly distributed throughout the questionnaire to ensure attentive and legitimate responses were submitted.

After the scenarios, the survey included questions from the privacy attitudes questionnaire [71] to assess participants' privacy inclinations. To minimize participant attrition, we selected only eight of the questionnaire's 36 items. These items were selected based on their relevance to information disclosure and conversational privacy. Questionnaire items are provided in the OSF repository linked above.

3) Participants: 174 U.S. participants, aged 22–78 (M = 39.6, SD = 10.708; 75% male, 25% female) completed the survey in 12 minutes on average.

B. Training the Privacy Controller

After data collection, we randomly split the 58 scenarios into 46 training scenarios and 12 testing scenarios. Privacy scores are created from the responses and used to label the privacy-sensitivity of each scenario. The dataset's labels are derived from the mean of the seven-point rating-scale responses associated with each survey question about privacy, sensitivity, and appropriateness of disclosure. Using the privacy score and thresholds set at 2.3 and 3.1, we classify scenarios into either low, moderate, or high control levels. Finally, rules are generated using the metadata and control levels of the 46 training scenarios, according to Algorithm 1. The test set is evaluated using the already generated rules.

C. Study 2: Evaluating the Privacy Controller

Following the data collection from Study 1 and training the privacy controller, we prepared relevant components (e.g., speech to text, text to speech, generated rules) for use in Study 2. This study focused on assessing the controller's performance in terms of privacy norms, social norms, and user trust toward the robot. To achieve this, we designed an online study to be administered via MTurk. We created 12 scenarios, which involved an initial conversation with a robot present, a controller response, and a baseline response, and asked participants to evaluate the robot's responses. We used the Misty II[4] as the social robot platform in our study.

1) Study Task: Our video scenarios were derived from the *privacy violation* and *control scenarios* used in Study 1. We took a random sample of 12 scenarios from the original 58 to use as our evaluation set for the privacy controller. Each scenario consisted of three videos including (1) a conversation video, (2) a controller response video, and (3) baseline response video. The *conversation video* involved two individuals reenacting the scenario script, with a social robot present and listening (Fig. 1). The *controller response video* and *baseline response video* introduced a new scene wherein a different individual asked the robot for information about the previous conversation. For the controller response condition, the robot replied with an answer determined by CONFIDANT. For the baseline response condition, the robot replied with low privacy control (i.e., the robot answered the individual's question completely). Each controller response video demonstrated high, medium, or low privacy control levels, while the baseline response always responded with low control.

2) Procedure, Measures, & Analysis: We presented videos containing human-robot conversations to participants through an MTurk survey. The study followed a within-subjects design where each participant was presented with a random sample of three scenarios to evaluate. Following each controller response and baseline response video, participants were asked to respond to a series of questions that addressed the controller's compliance with privacy norms, conversational norms, and overall trustworthiness (provided in the OSF repository linked above). Table I provides reports on the reliability of the resulting scale. Attention checks were randomly distributed throughout the questionnaires to monitor response legitimacy.

Our data analysis included a repeated measures one-way analysis of covariance (ANCOVA), where the controller was a within-participants factor while scenario and the interaction between the controller and the scenario were within-participants covariates. This analysis was repeated for the three measures: privacy norms, conversational norms, and trust.

3) Participants: 95 U.S. participants, aged 24–70 (M = 43.3, SD = 10.738; 54% male, 46% female) completed the survey in 9 minutes on average.

V. RESULTS

A. Annotated Dataset: Privacy Sensitivity of Conversations

The computed control levels, locations, sentiments, and topics in the resulting conversational dataset were distributed as follows: The dataset was labeled with 15 low control scenarios, 27 moderate control scenarios, and 16 high control scenarios; the dataset included 31 in-home scenarios and 27 out-of-home scenarios. Additionally, 20 scenarios contained neutral sentiment, 20 scenarios contained negative sentiment, and the remaining 18 scenarios were positive. While containing only 58 scenarios, the dataset remains balanced in almost every aspect except for conversation topic. Topics are skewed towards arts and entertainment, likely a consequence of many of the passages being derived from movie scripts. See Table II for the list of all metadata extracted from several scenarios.

Beyond training CONFIDANT, Study 1 also provided a dataset of privacy preserving phrases extracted from open-ended survey questions for two situations: (1) ensuring a listener does not share information and (2) indicating discomfort or inability to share information. Using these phrases, CONFIDANT

[4]https://www.mistyrobotics.com/

can automatically label in-the-wild conversations as privacy-sensitive allowing for future dataset expansion.

B. Trained Privacy Controller: Generating Association Rules

CONFIDANT is trained by generating a set of rules associated with a control level label for a given dataset of conversations. These labels are then used to generate association rules, similar to how privacy rules are created and managed in human-human interactions. Fig. 3 provides some examples of generated rules. For example, the controller learns a rule associating a conversation containing: a slightly negative sentiment, only 2 people, and a professional relationship between the conversation participants, with a high control level. The learned rule may be indicative of a negative and sensitive workplace conversation with a supervisor that is too inappropriate to share. However, due to the small sample size of the training data, several extraneous rules may appear, particularly those with fewer rule items. Having generated association rules using Algorithm 1, the privacy controller achieved 80.43% accuracy on the training set of scenarios and achieved 66.67% accuracy on the test set.

C. Evaluation: Studying the Privacy Controller in Human-Robot Interactions

To understand whether CONFIDANT truly improves users' perceptions, we generated video recordings of conversation scenarios. Participants were asked to respond to a questionnaire that measured the robot's ability to follow appropriate *social norms*, its ability to follow appropriate *privacy norms*, and their level of *trust* toward the robot. Our ANCOVA analysis for the social norms scale demonstrated that participants found the robot to more strongly follow expected norms of human conversation when it used our controller (M = 3.769, SD = 1.707) over the baseline behavior (M = 2.129, SD = 1.580), $F(1 : 86.92) = 250.90$, $p < .0001$. Similarly, analysis of the privacy norms scale showed participants rated the appropriateness of the robot's privacy norms significantly higher when it used our controller (M = 3.663, SD = 1.849) over the baseline (M = 1.697, SD = 1.487), $F(1 : 86.54) = 371.935$, $p < .0001$. Finally, participants regarded their trust towards the robot higher when the robot used our controller (M = 2.872, SD = 1.901) over the baseline (M = 1.505, SD = 1.509), $F(1 : 86.47) = 176.907$, $p < .0001$. Figs. 4 to 6 contain the responses used in our analysis of the robot's abilities. These differences in users' perceptions are more evident in scenarios with sensitive details compared to scenarios without.

Overall, participants perceived the robot as more socially-aware, privacy-aware, and trustworthy, when our privacy controller governed its disclosures. The controller's presence improved user perceptions of the robot's social awareness by 77%, privacy awareness by 116%, and trustworthiness by 91% over the baseline. After filtering the responses for privacy-conscious participants, the improvement to privacy-awareness

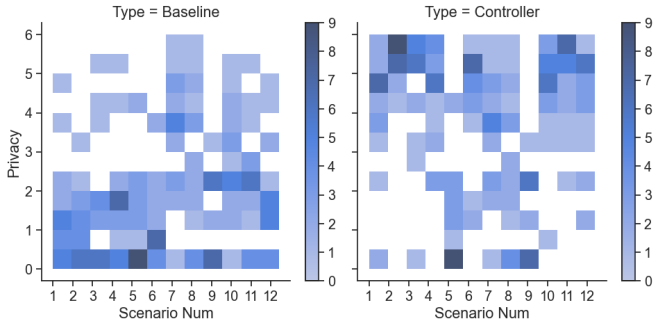

Fig. 4. The figures show participants perceived robots using CONFIDANT as more privacy-aware (right) compared to baseline robots (left). The heatmap shows the densities of the 7-point likert responses for the 12 test set scenarios.

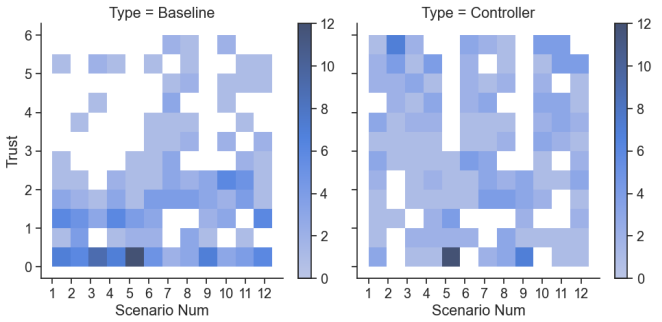

Fig. 5. The figures show participants perceived robots using CONFIDANT as more trustworthy (right) compared to baseline robots (left). The heatmap shows the densities of the 7-point likert responses for the 12 test set scenarios.

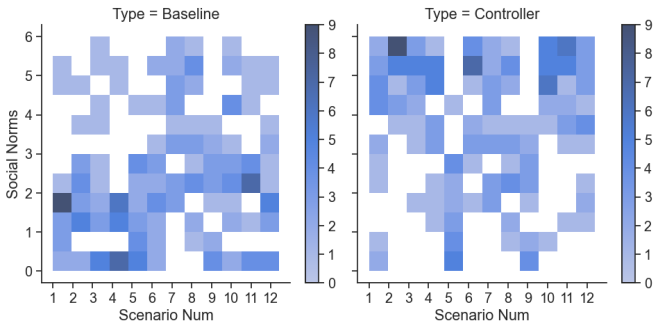

Fig. 6. The figures show participants perceived robots using CONFIDANT as more socially-aware (right) compared to baseline robots (left). The heatmap shows the densities of the 7-point likert responses for the 12 test set scenarios.

TABLE II
EXAMPLE OF THE METADATA TUPLES EXTRACTED FROM A SUBSET OF THE SCENARIOS

Scenario	Realism	Sentiment	Topics	Location	Relationships	Detail	# of People	Control Level
Control 6	4.67	Negative	Autos & Vehicles	Non-Domestic	Family	Medium	Few	Moderate
Control 11	5.27	Slightly Positive	Arts & Entertainment	Domestic	Close	Few	Short	Low
Scenario 2	4.75	Negative	None	Non-Domestic	Professional	Short	Some	High

TABLE III
MEAN SCORES FOR EACH EVALUATED CONSTRUCT (0-6)*

Construct	B	C	PC_B	PC_C	NPC_B	NPC_C
Privacy	1.70	3.66	1.48	3.76	1.97	3.54
Trust	1.51	2.87	1.26	2.75	1.83	3.03
Social Norms	2.13	3.77	1.96	3.82	2.35	3.70

* PC: privacy-conscious; NPC: non-privacy-conscious; B: baseline; C: controller

provided by CONFIDANT grows to 154%. For non-privacy-conscious participants, this improvement is 79% (Table III).

VI. DISCUSSION

Conversational agents that maintain the privacy of their users can serve as more effective and trustworthy conversational partners. Sannon et al. [72] explore how people's privacy perceptions of chatbots change when CPM theory guides the chatbot's social interactions and data-sharing practices. However, for social robots, maintaining the privacy of users is more complex due to the dynamic settings in which they function, such as team interactions in the workplace [33], or overhearing conversations between parents and children [50]. Our findings provide preliminary evidence for the effectiveness of CONFIDANT at managing privacy in social robots. Although our evaluations are performed with social robots, this privacy controller design may extend to other conversational agents. However, privacy typically comes at the expense of another attribute in the form of a trade-off [73]. While our approach at managing privacy does not negatively impact the perceived social awareness and trustworthiness of the robot, it may come at the expense of usability in certain circumstances. The following sections detail the drawbacks of CONFIDANT.

A. Limitations

While our work shows promising results, it must be considered alongside potential limitations. First, some scenarios created to train our privacy controller are subjective (i.e., participant responses varied significantly depending on participants). In particular, scenarios categorized as "privacy violations for the greater good" are determined based on what is considered morally right. While other category types centering around rules and responsibilities can be interpreted based on strict policies or legal boundaries, ethics are influenced by personal philosophies and belief systems. Ethical standards we enforce in the robot may not be universally recognized among all individuals. Second, we excluded highly sensitive conversations (e.g., related to abuse, substance use, or severe medical ailments) from our dataset to minimize psychological risk to our participants. Without these topics, we are unable to assess privacy in such cases. Third, our 58 scenarios only provides partial coverage of the vast space of conversation topics and dynamics. Finally, the current design of our privacy controller only accounts for the context available within a conversation and does not consider additional context such as pre-existing relationships or users' privacy preferences, which may otherwise impact sensitivity determinations.

B. Future Work

The scope of this work is limited to the design and evaluation of a generic "one size fits all" privacy controller. To address the system design limitations discussed in the previous section, our future work aims to implement adaptive and personalized features such as speaker relationships and conversation location. Additionally, we plan to explore how other factors highlighted by prior work, including personalization, ownership of devices, social roles, ownership of conversation content, dynamic relationships, co-embodiment, and social context [12, 13, 74], can be integrated into CONFIDANT. We also plan to extract more user-specific context when gathering and sharing information [12]. *Speaker relationships* will be registered by allowing each user to specify their relationships with other users during the initial robot setup and make any necessary changes moving forward. *Conversation locations* will be implemented using Visual Place Recognition (VPR) to classify images of rooms and identify the location where the conversation takes place (e.g., in a home [75] or non-domestic environments such as offices, medical facilities, and restaurants [76]). Furthermore, incorporating the comprehension and consideration of *non-verbal cues* increases the range of scenarios a social robot can handle and prevents non-verbal data leakage [77]. Additional use cases will also be tested to extend CONFIDANT to many contexts such as a personal assistant robot in the workplace or a medical robot that interacts with doctors and patients. The robot can adaptively learn new rules specific to these use cases by utilizing user feedback to the robot behavior (e.g., "*Please don't share this information with anyone again*"). Finally, the issue of scaling the rule generation system can be addressed by expanding our dataset through automated detection of sensitive conversations in online transcripts.

VII. CONCLUSION

Successful social integration of robots into daily environments such as in homes or workplaces is important to facilitate human-robot interactions. In these settings, social robots should promote trust and maintain the privacy of those they interact with. In the pursuit of designing more trustworthy robots, we created a privacy controller under simplified assumptions and evaluated its performance in human-robot conversations. Our findings present preliminary evidence for the effectiveness of the privacy controller in creating more privacy-aware, trustworthy, and socially aware social robots.

ACKNOWLEDGEMENTS

We would like to thank Christine Lee for her help in filming the human-robot interaction videos, the anonymous reviewers for their helpful suggestions, and study participants to their contributions to our data collection.

REFERENCES

[1] G. Glinska, R. Raveendhran, and C. Breazeal, "The rise of social robots: How ai can help us flourish," *Data & Analytics and Entrepreneurship & Innovation*, 2020.

[2] L. Wood, "Insights on the social robots global market to 2026 - rise in automation is driving growth," *GLOBE NEWSWIRE*, 2021.

[3] C. Lutz, M. Schöttler, and C. Hoffmann, "The privacy implications of social robots: Scoping review and expert interviews," *Mobile Media & Communication*, vol. 7, pp. 412–434, Sep. 2019. DOI: 10.1177/2050157919843961.

[4] E. Fosch Villaronga, C. Lutz, and A. Tamò Larrieux, "Gathering expert opinions for social robots' ethical, legal, and societal concerns: Findings from four international workshops," *International Journal of Social Robotics*, vol. online first, pp. 1–18, May 2020. DOI: 10.1007/s12369-019-00605-z.

[5] R. S. Sandhu and P. Samarati, "Access control: Principle and practice," *IEEE communications magazine*, vol. 32, no. 9, pp. 40–48, 1994.

[6] W.-L. Chang and S. Šabanović, "Studying socially assistive robots in their organizational context: Studies with paro in a nursing home," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, 2015, pp. 227–228.

[7] A. Sauppé and B. Mutlu, "The social impact of a robot co-worker in industrial settings," in *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 2015, pp. 3613–3622.

[8] A. Tapus, C. Tapus, and M. J. Mataric, "The use of socially assistive robots in the design of intelligent cognitive therapies for people with dementia," in *2009 IEEE international conference on rehabilitation robotics*, IEEE, 2009, pp. 924–929.

[9] D. P. Davison, F. M. Wijnen, V. Charisi, J. van der Meij, V. Evers, and D. Reidsma, "Working with a social robot in school: A long-term real-world unsupervised deployment," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 63–72.

[10] M. M. de Graaf, S. B. Allouch, and J. A. van Dijk, "Long-term evaluation of a social robot in real homes," *Interaction studies*, vol. 17, no. 3, pp. 462–491, 2016.

[11] D. Seifert, "Say hello to astro, alexa on wheels," *Smart Home*, 2021.

[12] M. Luria, R. Zheng, B. Huffman, S. Huang, J. Zimmerman, and J. Forlizzi, "Social boundaries for personal agents in the interpersonal space of the home," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–12.

[13] S. Reig, M. Luria, E. Forberger, I. Won, A. Steinfeld, J. Forlizzi, and J. Zimmerman, "Social robots in service contexts: Exploring the rewards and risks of personalization and re-embodiment," in *Designing Interactive Systems Conference 2021*, 2021, pp. 1390–1402.

[14] D. Spiliotopoulos, I. Androutsopoulos, and C. D. Spyropoulos, "Human-robot interaction based on spoken natural language dialogue," in *Proceedings of the European workshop on service and humanoid robots*, Citeseer, 2001, pp. 25–27.

[15] L. S. Lopes and A. Teixeira, "Human-robot interaction through spoken language dialogue," in *Proceedings. 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2000)(Cat. No. 00CH37113)*, IEEE, vol. 1, 2000, pp. 528–534.

[16] B. Mutlu, "Designing embodied cues for dialog with robots," *AI Magazine*, vol. 32, no. 4, pp. 17–30, 2011.

[17] Y. Matsusaka, S. Fujie, and T. Kobayashi, "Modeling of conversational strategy for the robot participating in the group conversation," in *Seventh European conference on speech communication and technology*, Citeseer, 2001.

[18] B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, and N. Hagita, "Footing in human-robot conversations: How robots might shape participant roles using gaze cues," in *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, 2009, pp. 61–68.

[19] S. Andrist, X. Z. Tan, M. Gleicher, and B. Mutlu, "Conversational gaze aversion for humanlike robots," in *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, 2014, pp. 25–32.

[20] C. Chao and A. L. Thomaz, "Turn taking for human-robot interaction," in *2010 AAAI Fall Symposium Series*, 2010.

[21] B. Mutlu, T. Kanda, J. Forlizzi, J. Hodgins, and H. Ishiguro, "Conversational gaze mechanisms for humanlike robots," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 1, no. 2, pp. 1–33, 2012.

[22] M. E. Foster, T. By, M. Rickert, and A. Knoll, "Human-robot dialogue for joint construction tasks," in *Proceedings of the 8th international conference on Multimodal interfaces*, 2006, pp. 68–71.

[23] M. Staudte and M. W. Crocker, "Visual attention in spoken human-robot interaction," in *2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, 2009, pp. 77–84.

[24] C. Liu, J. Walker, and J. Y. Chai, "Ambiguities in spatial language understanding in situated human robot dialogue," in *2010 AAAI Fall Symposium Series*, 2010.

[25] N. Jacobs and A. Garnham, "The role of conversational hand gestures in a narrative task," *Journal of Memory and Language*, vol. 56, no. 2, pp. 291–303, 2007.

[26] C.-M. Huang and B. Mutlu, "Modeling and evaluating narrative gestures for humanlike robots.," in *Robotics: Science and Systems*, 2013, pp. 57–64.

[27] S. Andrist, E. Spannan, and B. Mutlu, "Rhetorical robots: Making robots more effective speakers using linguistic cues of expertise," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, 2013, pp. 341–348.

[28] C. Torrey, S. R. Fussell, and S. Kiesler, "How a robot should give advice," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, 2013, pp. 275–282.

[29] V. Srinivasan and L. Takayama, "Help me please: Robot politeness strategies for soliciting help from humans," in *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2016, pp. 4945–4955.

[30] N. Mirnig, S. Stadler, G. Stollnberger, M. Giuliani, and M. Tscheligi, "Robot humor: How self-irony and schadenfreude influence people's rating of robot likability," in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, IEEE, 2016, pp. 166–171.

[31] T. Iio, M. Shiomi, K. Shinozawa, K. Shimohara, M. Miki, and N. Hagita, "Lexical entrainment in human robot interaction," *International Journal of Social Robotics*, vol. 7, no. 2, pp. 253–263, 2015.

[32] S. Strohkorb Sebo, M. Traeger, M. Jung, and B. Scassellati, "The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human-robot teams," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 178–186.

[33] M. L. Traeger, S. S. Sebo, M. Jung, B. Scassellati, and N. A. Christakis, "Vulnerable robots positively shape human conversational dynamics in a human–robot team," *Proceedings of the National Academy of Sciences*, vol. 117, no. 12, pp. 6370–6375, 2020.

[34] N. T. White, B. Cagiltay, J. E. Michaelis, and B. Mutlu, "Designing emotionally expressive social commentary to facilitate child-robot interaction," in *Interaction Design and Children*, 2021, pp. 314–325.

[35] H. Ling and E. Björling, "Sharing stress with a robot: What would a robot say?" *Human-Machine Communication*, vol. 1, 2020.

[36] C. Fu, Y. Yoshikawa, T. Iio, and H. Ishiguro, "Sharing experiences to help a robot present its mind and sociability," *International Journal of Social Robotics*, vol. 13, no. 2, pp. 341–352, 2021.

[37] S. Mitsuno, Y. Yoshikawa, and H. Ishiguro, "Robot-on-robot gossiping to improve sense of human-robot conversation," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, IEEE, 2020, pp. 653–658.

[38] Y. Moon, "Intimate exchanges: Using computers to elicit self-disclosure from consumers," *Journal of consumer research*, vol. 26, no. 4, pp. 323–339, 2000.

[39] A. Reben and J. Paradiso, "A mobile interactive robot for gathering structured social video," in *Proceedings of the 19th ACM international conference on Multimedia*, 2011, pp. 917–920.

[40] M. Rueben, C. M. Grimm, F. J. Bernieri, and W. D. Smart, "A taxonomy of privacy constructs for privacy-sensitive robotics," *arXiv preprint arXiv:1701.00841*, 2017.

[41] J. Lau, B. Zimmerman, and F. Schaub, "Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, pp. 1–31, 2018.

[42] A. Sciuto, A. Saini, J. Forlizzi, and J. I. Hong, "" hey alexa, what's up?" a mixed-methods studies of in-home conversational agent usage," in *Proceedings of the 2018 Designing Interactive Systems Conference*, 2018, pp. 857–868.

[43] E. Beneteau, A. Boone, Y. Wu, J. A. Kientz, J. Yip, and A. Hiniker, "Parenting with alexa: Exploring the introduction of smart speakers on family dynamics," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–13.

[44] N. Malkin, S. Egelman, and D. Wagner, "Privacy controls for always-listening devices," in *Proceedings of the New Security Paradigms Workshop*, ser. NSPW '19, San Carlos, Costa Rica: Association for Computing Machinery, 2019, pp. 78–91, ISBN: 9781450376471. DOI: 10.1145/3368860.3368867. [Online]. Available: https://doi.org/10.1145/3368860.3368867.

[45] H. Kuzuoka, Y. Suzuki, J. Yamashita, and K. Yamazaki, "Reconfiguring spatial formation arrangement by robot body orientation," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, 2010, pp. 285–292.

[46] C. Shi, M. Shimada, T. Kanda, H. Ishiguro, and N. Hagita, "Spatial formation model for initiating conversation," *Proceedings of robotics: Science and systems VII*, pp. 305–313, 2011.

[47] N. Meng, D. Keküllüoğlu, and K. Vaniea, "Owning and sharing: Privacy perceptions of smart speaker users," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–29, 2021.

[48] N. Abdi, K. M. Ramokapane, and J. M. Such, "More than smart speakers: Security and privacy perceptions of smart home personal assistants," in *Fifteenth Symposium on Usable Privacy and Security ({SOUPS} 2019)*, 2019, pp. 451–466.

[49] M. Luria, "Mine, yours or amazon's? designing agent ownership and affiliation," in *Companion Publication of the 2020 ACM Designing Interactive Systems Conference*, 2020, pp. 537–542.

[50] B. Cagiltay, H.-R. Ho, J. E. Michaelis, and B. Mutlu, "Investigating family perceptions and design preferences for an in-home robot," in *Proceedings of the interaction design and children conference*, 2020, pp. 229–242.

[51] S. Petronio, *Boundaries of privacy: Dialectics of disclosure*. Suny Press, 2002.

[52] ——, "Communication privacy management theory: What do we know about family privacy regulation?" *Journal of family theory & review*, vol. 2, no. 3, pp. 175–196, 2010.

[53] H. Nissenbaum, *Privacy in context*. Stanford University Press, 2020.

[54] D. Hymes, *Foundations in sociolinguistics: An ethnographic approach*. Routledge, 2013.

[55] S. Petronio and C. Bantz, "Research note: Controlling the ramifications of disclosure:don't tell anybody but..," *Journal of Language and Social Psychology*, vol. 10, no. 4, pp. 263–269, 1991.

[56] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2015, pp. 5206–5210.

[57] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[58] Google, *Natural language ai*, https://cloud.google.com/natural-language, 2021.

[59] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA: Association for Computational Linguistics, 2011, pp. 142–150. [Online]. Available: http://www.aclweb.org/anthology/P11-1015.

[60] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," in *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[61] Yahoo! *Yahoo! answers language data*, https://webscope.sandbox.yahoo.com/catalog.php?datatype=l, 2021.

[62] SpeechBrain, *Speechbrain a pytorch powered speech toolkit*, https://speechbrain.github.io/, 2021.

[63] A. Nagrani, J. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Science and Language*, 2019.

[64] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, "Hate speech detection with comment embeddings," in *Proceedings of the 24th international conference on world wide web*, 2015, pp. 29–30.

[65] W. Lan and W. Xu, "Neural network models for paraphrase identification, semantic textual similarity, natural language inference, and question answering," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 3890–3902.

[66] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Nov. 2019. [Online]. Available: http://arxiv.org/abs/1908.10084.

[67] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation," *arXiv preprint arXiv:1708.00055*, 2017.

[68] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in *IN: PROCEEDINGS OF THE 1993 ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, WASHINGTON DC (USA*, 1993, pp. 207–216.

[69] R. Agrawal, R. Srikant, *et al.*, "Fast algorithms for mining association rules," in *Proc. 20th int. conf. very large data bases, VLDB*, Citeseer, vol. 1215, 1994, pp. 487–499.

[70] IMSDb, *Internet movie scripts database*, https://imsdb.com/, 2021.

[71] M. Chignell, J. Gwizdka, and A. Quan-Haase, "The privacy attitudes questionnaire (paq): Initial development and validation," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 47, Oct. 2003. DOI: 10.1177/154193120304701102.

[72] S. Sannon, B. Stoll, D. DiFranzo, M. F. Jung, and N. N. Bazarova, "I just shared your responses: Extending communication privacy management theory to interactions with conversational agents," *Proc. ACM Hum.-Comput. Interact.*, vol. 4, no. GROUP, Jan. 2020. DOI: 10.1145/3375188. [Online]. Available: https://doi.org/10.1145/3375188.

[73] R. N. Wright, L. J. Camp, I. Goldberg, R. L. Rivest, and G. Wood, "Privacy tradeoffs: Myth or reality?" In *International Conference on Financial Cryptography*, Springer, 2002, pp. 147–151.

[74] S. Reig, M. Luria, J. Z. Wang, D. Oltman, E. J. Carter, A. Steinfeld, J. Forlizzi, and J. Zimmerman, "Not some random agent: Multi-person interaction with a personalizing service robot," in *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, 2020, pp. 289–297.

[75] P. Uršič, A. Leonardis, M. Kristan, *et al.*, "Part-based room categorization for household service robots," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2016, pp. 2287–2294.

[76] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2015.

[77] B. Mutlu, F. Yamaoka, T. Kanda, H. Ishiguro, and N. Hagita, "Nonverbal leakage in robots: Communication of intentions through seemingly unintentional behavior," in *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, 2009, pp. 69–76.