 https://dx.doi.org/10.1109/RO-MAN57019.2023.10309501

Failure Explanation in Privacy-Sensitive Contexts: An Integrated Systems Approach

Sihui Li†, Sriram Siva†, Terran Mott†, Tom Williams†, Hao Zhang‡, and Neil Dantam†

***Abstract*— In this paper, we explore how robots can properly explain failures during navigation tasks with privacy concerns. We present an integrated robotics approach to generate visual failure explanations, by combining a language-capable cognitive architecture (for recognizing intent behind commands), an object- and location-based context recognition system (for identifying the locations of people and classifying the context in which those people are situated) and an infeasibility proof-based motion planner (for explaining planning failures on the basis of contextually mediated privacy concerns). The behavior of this integrated system is validated using a series of experiments in a simulated medical environment.**

I. Introduction

Successful deployment of robots into human environments requires appropriately calibrated trust in those robotic systems. Accordingly, robots must be able to properly *explain* their plans to humans to provide the transparency necessary for such trust calibration transparency [1], [2], [3], [4], [5], [6]. This is especially important when plan failure happens and when users are non-experts, but especially when plans fail for reasons that humans cannot directly perceive [7], [8].

While robot explanations typically come in the form of natural language, this is not always the best or easiest way to communicate information, especially in the context of mapping and navigation. In this work, we thus consider how *visual* explanations might be generated in integrated robot architectures, in these sorts of contexts. Specifically, we consider how failure explanations can be represented geometrically, projected onto maps, and visualized to users.

We argue that visual failure explanations, such as those presented in this paper, may be especially useful in privacy-sensitive contexts, where the reason for planning failure is not itself physical, yet relates to readily visualizable spatial extents. Privacy concerns are especially salient for mobile robots, and in many domains where interactive robots are being proposed to be used, including sensitive medical settings like hospitals and doctor's offices where privacy is both a moral and legal issue [9], [10], [11], [12], [13]. Mobile robots may record audio and video to support speech recognition, facial recognition, object tracking, navigation, teleoperation, and so forth. Such recordings may capture sensitive health information, e.g., video recording of bandage changes, or audio recording conversations regarding STD diagnoses [14].

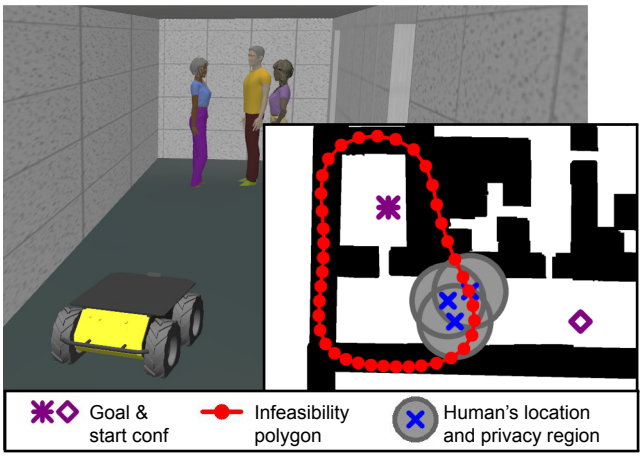

Fig. 1: **System Validation Scenario.** For the robot to reach a goal location in the context shown, it would need to enter the *physical privacy regions* of the depicted humans. We prove this by constructing an infeasibility polygon in the map. This infeasibility polygon is a visual explanation that informs planning failures to users on privacy grounds.

We argue in this paper that mobile robots can minimize these ethical and legal risks by intentionally circumventing morally fraught situations according to their own sensory capabilities, and providing visual explanations as to why this is necessary. While there are many different lenses through which privacy can be examined [15], [16], we consider how robots might protect the *physical* privacy of the humans in their environments by rejecting commands that would bring them into *"physical privacy regions"*: users' contextually scoped personal zones within which sensitive conversations could be overheard or sensitive procedures could be witnessed. These physical privacy regions are contextually scoped. That is, spatially grounded privacy concerns are mediated by environmental and interaction context: the distance within which a robot can come to passersby is likely to be a shorter distance when in a busy hallway than a private hospital room; and likely to be a farther distance when humans are obviously engaged in conversation than when they are merely passing by. Because of this change in scope, it is difficult for users to directly perceive the physical privacy regions.

In this work, we thus introduce an integrated approach enabling robots to provide visual failure explanations on privacy grounds in navigation tasks. This approach extend

† Department of Computer Science, Colorado School of Mines. Email: {li,sivasriram,terranmott,twilliams,ndantam}@mines.edu.

‡ Human-Centered Robotics Lab, University of Massachusetts Amherst. Email: hao.zhang@umass.edu.

This research was funded in part by National Science Foundation grant #IIS-1849348.

our previous demonstration of social robots using formal representations of moral and social norms to reject inappropriate commands in a context-sensitive task planner [17]. In contrast, this work addresses how robots can similarly reject commands that violate norms with respect to *motion planning*. Our integrated approach combines three key architectural subsystems: (1) the Distributed, Integrated, Affect, Reflection, Cognition (DIARC) robot architecture, which we use for language and goal-driven cognition, (2) a novel motion planning method to assess context-sensitive feasibility of users' commands on privacy grounds; and (3) a ROS-based contextual reasoning system, which we use for place recognition and object detection. The heart of this integrated approach is our novel application of recent work generating motion planning infeasibility polygons [18] on a 2D navigation map with privacy constraints. Figure 1 shows an example scenario. The infeasibility polygon separates the start and goal locations and demonstrates to users the precise reason why the navigation plan fails on privacy grounds.

II. Related Work

In this section, we discuss recent work in the context of (1) failure explanation for transparent and trusted human-robot interaction; (2) privacy concerns surrounding social robots; (3) social navigation; and (4) infeasibility proofs.

A. Failure Explanation for Transparency and Trust

There has been a vast array of work on explaining robot plans in natural language to improve (justifiable) human-robot trust through increased robot transparency [1], [2], [3], [4], [5], [6], [19], [20]. Our work seeks to generate visual explanations when the plan fails. Recently, researchers have also considered communication of plan failures, especially in the context of command rejection [21], and explored factors that influence when a robot should reject directives, including not only factors like physical capacity, but also normative permissibility [22]. While many approaches to command rejection have been explicitly grounded in Deontic Norms [23] and similar formalisms [24], [25], [26], [27], other researchers have recently considered whether approaches grounded in robots' social roles and relationships may be more effective [28], [29]. Our work builds on this by ensuring that we not only detect *when* privacy violations could occur (in the form of explainable polygons) but moreover *whose* privacy would be violated (i.e, the identities of the persons whose physical privacy regions intersects the infeasibility polygon), in order to ensure compatibility with both norm-driven and relation-driven approaches.

B. Privacy Concerns Surrounding Social Robots

While social robots are predominantly designed for pro-social purposes (especially hospital robots, which seek to reduce harm and enhance life- and health-oriented capabilities, especially for vulnerable and structurally limited communities (cf. [30], [31])), they nevertheless raise key concerns that must be addressed for safe and successful deployment [32], including privacy concerns [32], [33], [9], [10], [13]. Lee et al. emphasize the relationship between these risks and issues of transparency, by showing that most users were not able to identify the types of information collected by robots [33]. Moreover, Lutz and Tamò-Larrieux demonstrate how users' privacy concerns impact the ways they choose to interact with robots. Other researchers have accordingly proposed methods for improving data privacy and safe data handling in the design of social robot systems [34], [11]. However, we argue that because users have difficulty assessing what information is actually being collected by robots, and because privacy concerns can hamper effective interaction, we suggest that robots be designed to avoid what *would be perceived* as constituting a privacy violation *even if* the way data is handled is privacy-sensitive and HIPAA-compliant [35].

C. Social Navigation

Social navigation allows robots to navigate in a way that is sensitive to human social norms and social or cultural expectations. Social navigation can increase humans' comfort while coexisting with mobile robots [36], [37], [38]. A key aspect of social navigation is endowing robots with an understanding of human proxemics. Proxemics are the set of social norms that humans use to navigate around one another, like standing in line and respecting others' personal space [39]. Robots can perceive and react to human proxemics, adjusting dynamically as people move or alter their body language [40], [41], [42], [43], [44]. Socially aware navigation allows robots to predict human motion and move in ways that are more predictable themselves, which can result in more effective navigation and interaction overall [45], [46].

Our work is closely related to social navigation in that we are interested in navigation that is sensitive to normative factors. However, our approach is unique with respect to previous work on social navigation in terms of our emphasis both on privacy considerations and on command rejection.

D. Infeasibility Proofs in Motion Planning

To enable privacy-sensitive social robot navigation, privacy concerns need to be addressed at the motion-planning level. In recent work, we have presented novel motion planning techniques grounded in plan infeasibility analysis [18], which we argue can be used to ensure privacy-sensitive social robot navigation when privacy-driven constraints could cause plan infeasibility. Infeasibility proof based motion planning differs from traditional sampling-based planning [47], [48], [49] in critical ways. Notably, while sampling-based motion planning is only *probabilistically* complete, infeasibility proof based motion planning offers stronger guarantees: it is guaranteed to terminate with either a plan or an infeasibility proof. Such a planner is important in privacy-sensitive scenarios not only for ensuring privacy-sensitive behaviors but also for the purposes of explainability since infeasibility proofs can be used to provide exact explanations as to why a command cannot be executed ethically. There have been some other approaches that generate infeasibility proofs [50], [51], [52], [53], [54]. Our approach [18] has key advantages over those approaches in terms of explainability, as our infeasibility proofs are

generated in the form of easily visualizable geometries, for example, 2D polygons for holonomic mobile robots, which facilitate analyzing and explaining plan failures.

III. Technical Approach

Our technical approach uses an integrated system with three components: (1) the DIARC Robot Architecture, which is responsible for natural language understanding and goal-driven cognition; (2) the motion planning system, responsible for privacy-based planning to achieve language-specified goals in a context-sensitive manner and generate visual failure explanations; (3) a ROS-based contextual reasoning system for context/place recognition and object detection. Figure 2 illustrates this integrated system. In this section, we detail each constituent sub-system.

A. Language Understanding and Goal-Driven Cognition

To decide whether to accept or reject a user's command on privacy grounds, we must first identify and manage the intent behind that command. We achieve this using the Distributed, Integrated, Affect, Reflection, Cognitive (DIARC) Robot Architecture [55]: a hybrid deliberative-reactive robot architecture with a wide array of cognitive capabilities, with special attention to language understanding and generation [56]. Our configuration of DIARC is implemented using the Agent Development Environment (ADE), a distributed multi-agent system middleware [57]. Our ADE-implemented configuration of DIARC leverages the capabilities of six key architectural components: the Speech Recognition, Parsing, Pragmatic Inference, Reference Resolution, a Spatial Consultant, the Dialogue manager, and the Goal Manager.

When a user speaks to the robot, DIARC's speech recognition component converts their speech into text, which is provided to the parser [58], [59]. The parser uses a Combinatory Categorial Grammar [60] to translate this text into a set of logical predicates encoding the surface semantics of the speaker's utterance, along with a Givenness Hierarchy [61] theoretic mapping of the variables used in those predicates to their presumed cognitive statuses, indicating whether the speaker's phrasing suggests the people, objects, and locations they referenced are in focus, activated, familiar, uniquely identifiable, or type identifiable, to facilitate the understanding of not only definite descriptions, but also anaphora, deictic pronouns, indefinite noun phrases, and so forth [62]. The predicates produced by this parser are organized into a list P whose head P_0 represents the *primary* semantics encoding the (surface-level) intent of the utterance, and whose tail $P_{1:n}$ represents the *supplemental* semantics encoding the properties ascribed in the utterance to any objects, locations, and people mentioned in the utterance.

The *primary* semantics are then sent to the pragmatics component, which uses context sensitive rules encoding, e.g., Indirect Speech Act theoretic politeness norms [63], [64], to determine the *intended* meaning of the speaker's utterance—e.g., from the semantic representation of an utterance such as "Can you go to the kitchen?" the pragmatics component would typically (depending on context) infer that the speaker wishes the listener to have a goal to go to the kitchen. The *supplemental* semantics are then sent to the reference resolution component, which uses consultants (such as the spatial consultant) to identify the objects, locations, and people described in the utterance. This component uses the DIST-POWER [65] algorithm (see also [66]), which enables reference resolution under conditions of both uncertainty (when the robot is unsure whether a property holds for a given entity) and ignorance (when the robot can determine that an object being described was previously not known to it, and thus that it must create a new mental representation for that object), and which can operate with information regarding entities distributed across different components, on different machines, using different knowledge representations. This algorithm is used as integrated into the Giveness and Relevance-Theoretic Open World Reference Resolution (GROWLER) algorithm [67], which facilitates the resolution of a wide array of linguistic forms [68], [62].

Once the intent behind an utterance and the entities referenced within that utterance are identified, a bound utterance structure is provided to the Dialogue manager, which, if the robot decides to do so, uptakes that intention; optionally uptaking the propositions behind any assertions, forming an intention to respond to any requests for information, or adopting the goals associated with any commands or requests for action. If a goal is adopted, the user's intention is sent to the Goal Manager [69], [59], which is responsible for prioritizing between and managing possibly competing goals and selecting actions in service of those goals. In this work, we equip the Goal Manager with a *Motion Action Selector* that extracts the location specified from navigation goals and sends this navigation goal to the motion planning part, through a REST API similar to that specified by Jackson et al. [17], awaiting a response that either (a) indicates motion planning success, or (b) provides the information necessary to generate an appropriate failure explanation.

B. Motion Planning

We integrate privacy regions into the motion planning problem formulation to ensure that valid plans adhere to privacy constraints. Specifically, we consider privacy regions as part of the configuration space obstacle region. Then, we produce either a feasible motion plan or a proof—and explanation—of motion planning infeasibility.

A motion planning problem [70] consists of a configuration space $\mathcal{C}$ of dimension n, a start configuration $\boldsymbol{q}_{\text{start}}$, and a goal configuration $\boldsymbol{q}_{\text{goal}}$. The configuration space $\mathcal{C}$ is the union of the disjoint obstacle region $\mathcal{C}_{\text{obs}}$ and free space $\mathcal{C}_{\text{free}}$. Both $\boldsymbol{q}_{\text{start}}$ and $\boldsymbol{q}_{\text{goal}}$ are in $\mathcal{C}_{\text{free}}$. We define a human's privacy region as an area function $\mathcal{PR}$ of the location of the human in the workspace (h_i), the privacy range of the human (r_i), and the context the human is in (c_i), where the range r_i is also determined by the context c_i. The motion planning obstacle region $\mathcal{C}_{\text{obs}}$ is the union of the space with physical

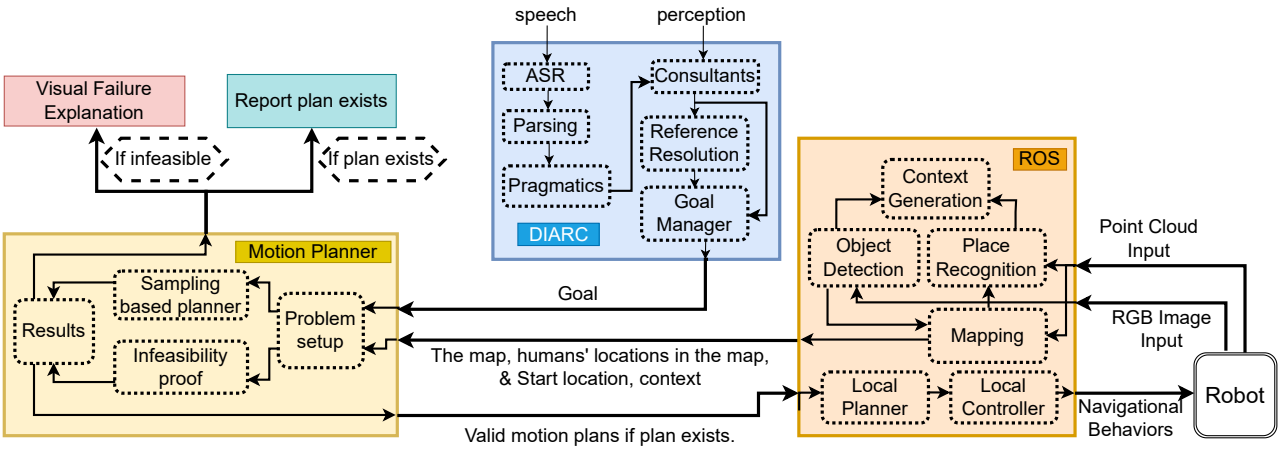

Fig. 2: Integrated System Architecture

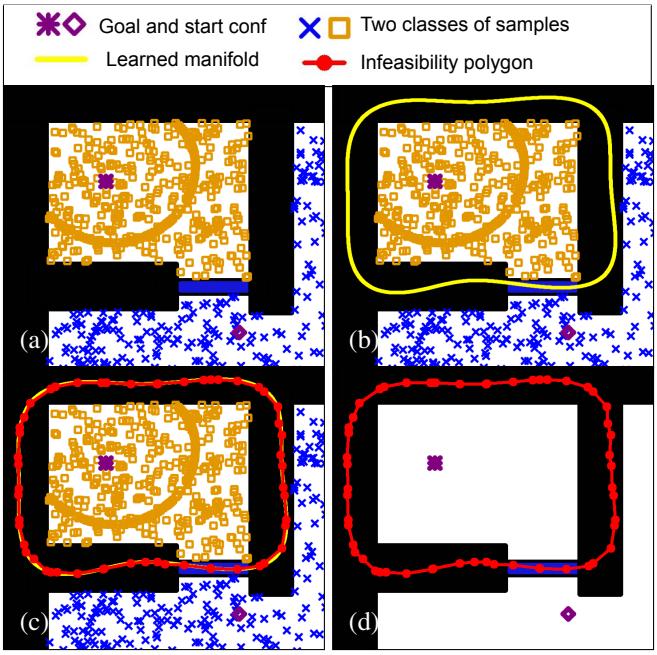

Fig. 3: Example of the motion planning infeasibility polygon construction steps from (a) to (d).

obstacles $\mathcal{C}_{\text{obs,phys}}$ and the privacy region of each human,

$$\mathcal{C}_{\text{obs}} = \mathcal{C}_{\text{obs,phys}} \bigcup_{i=1}^{n} \mathcal{PR}(h_i, r_i(c_i), c_i) \ . \tag{1}$$

For example, in Figure 1, the black regions are the physical obstacle regions caused by walls and objects in a 3D space, and the grey regions around the locations of the humans are the privacy regions. The union of these two types of obstacle regions forms the final configuration space $\mathcal{C}_{\text{obs}}$.

With this definition of $\mathcal{C}_{\text{obs}}$, we use the algorithm in [18] to find a plan if the motion planning problem is solvable or construct an infeasibility proof when no such plans exist. In contrast to most classic motion planners [71], [49], which may generate a plan when one exists but cannot show infeasibility, this algorithm with infeasibility proof construction is a good fit for our application because infeasibility proofs provide information about *why* a command must be rejected.

The motion planner runs two separate threads in parallel (Figure 2). The planning thread runs a sampling-based motion planner to find a plan, e.g., RRT-connect [71] or Probabilistic RoadMap (PRM) [49], and saves all the sampled configurations. In our system, we use PRM since its multi-directional sampling strategy can cover a complex 2D map more efficiently. Another thread attempts to construct an infeasibility proof using the sampled configurations. We explain the infeasibility proof construction steps using the 2D example in Figure 3. First, with samples from the planning thread, the algorithm groups all the $\mathcal{C}_{\text{free}}$ points connectable to the goal point as one class and all other $\mathcal{C}_{\text{free}}$ points as another class, as shown in Figure 3 (a). Then, we train a classifier with the two classes (Figure 3 (b)). Geometrically, the classifier is a manifold that separates the two classes. Next, we triangulate the manifold to construct a 2D polygon (Figure 3 (c)). If every line segment of the 2D polygon is in $\mathcal{C}_{\text{obs}}$, then the 2D polygon is an infeasibility proof. In 2D, we call it an infeasibility polygon. An infeasibility proof is a closed manifold that exists entirely in $\mathcal{C}_{\text{obs}}$ and that separates the start and the goal [18]. If an infeasibility proof exists, it means there is no collision path connecting the start and the goal. The final infeasibility polygon is shown in Figure 3 (d). With the planning thread and the infeasibility proof construction thread, our 2D motion planner returns either a valid path or an infeasibility polygon.

The infeasibility polygon explains visually why plans do not exist, since it creates a geometrical separation in the configuration space and in the map. With the infeasibility polygon, the user can understand why planning fails, analyze this result, and decides on alternate actions. For example, if the infeasibility polygon's edges overlap with some humans' privacy regions, then the privacy regions could cause failure. Figure 1 shows the 2D configuration space of a holonomic mobile robot. The red polygon is the infeasibility polygon, which exists entirely in $\mathcal{C}_{\text{obs}}$ (the three humans' privacy regions are parts of $\mathcal{C}_{\text{obs}}$) and explains why the plan fails by separating the start and the goal. One part of future work is to collect the humans' locations h_is and privacy ranges r_is

causing failure, and send these to the language understanding module to explain the failure using natural language.

Note that the infeasibility proof construction algorithm works for kinematic motion planning problems only and does not consider dynamic constraints. Thus, our current system assumes a holonomic mobile robot without considering control uncertainties, steering functions or dynamics.

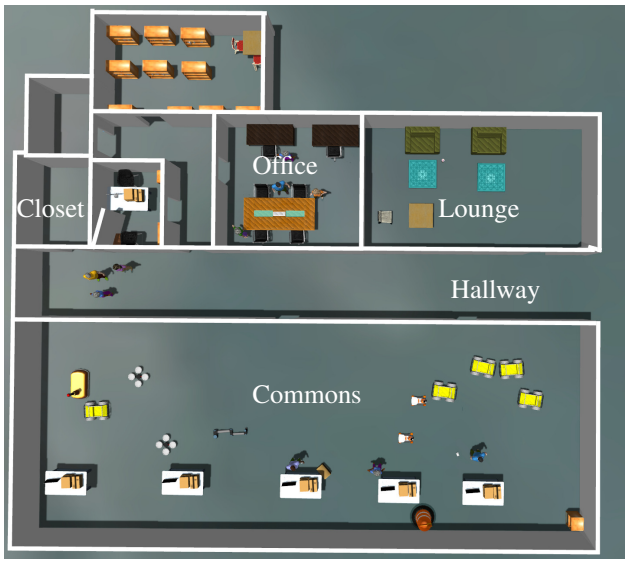

Fig. 4: Simulated Robot Environment.

C. Place Recognition and Object Detection

The capabilities described in the previous section are facilitated by a place recognition and object detection based context recognition system. Place recognition [72] seeks to identify a given location from a set of templates and is an essential capability for mobile robots. It reduces ambiguity and accumulated errors during mapping and robot localization, thereby significantly improving robot mapping and localization accuracy. Long-term Place recognition [73] addresses the key challenge that many robot navigation environments are dynamic in nature and change over time. For example, when navigating in indoor environments with changing lighting conditions, arrangement of furniture, and human movements that vary on a daily basis.

We perform long-term place recognition using *voxel-based representation learning* [74] (VBRL). The VBRL method uses 3D point clouds obtained from a LiDAR sensor to form representations of the environment and then learns from these representations to recognize previously visited locations by comparing the new point cloud scan with an existing 3D point cloud map. We use VBRL to recognize the context from a 360-degree field of view of the LiDAR sensor. This capability is helpful in indoor environments where some objects or even humans may occlude the limited file of view of a traditional RGB camera. This VBRL approach divides the input 3D point clouds into multiple voxels in a 3D space and extracts multi-modal features from each voxel. Then, VBRL uses regularized optimization to learn the importance of each feature modality within every voxel and the representativeness of voxels. The voxel importance learning is inspired by the insight that certain voxels in a 3D space are more important as they can better encode location-based context. For example, a LiDAR sensor obtains more information from points closer to it, and thus, voxels near the sensors have more information, and their importance should be learned accordingly. Mathematically, the voxel representations are achieved by using sparsity inducing regularization norms in the objective function.

Context-recognition is further improved by pairing long-term place recognition with object detection. Object detection allows the robot to find all instances of one or more given object classes irrespective of their scale, location, pose, view concerning the robot, or even illumination conditions. We detect the objects as seen by the robot's RGB camera using *You Only Look Once* (YOLO) [75], and then compare them with the location context to check if these objects belong in the context. For example, if a robot recognizes the location context as office-room through VBRL, YOLO cross-checks this information by looking for objects such as monitors, people, tables, etc., which are common in office spaces. YOLO is also used to estimate the distance of humans from the robot and ground their locations in the robot's map.

Together, these capabilities are thus leveraged to identify people and their locations, to identify the context needed to parameterize those peoples' physical privacy regions, and to generate the visual explanations for the robot's decisions.

IV. Experimental Validation

We validate our system on a Clearpath Husky ground robot simulated in a Gazebo environment, as illustrated in Figure 4. The robot is equipped with an Intel Realsense D435 camera, Ouster OS-1 64 LiDAR and a ReSpeaker Mic Array v2.0 microphone which is capable of detecting voices upto 5m away. We assume the robot's mic is omnidirectional and sound travels equally in all directions in an indoor environment, such that we can simplify the privacy regions to be circular areas centered at the location of the human. Figure 4 shows the simulated world consists of multiple rooms, the closet, the offices, the commons, and the hallway. In addition to the 3D map used for place recognition, we built a 2D robot collision map to enable motion planning using the 3D point cloud from the LiDAR sensor. Using *elevation mapping* [76], we detect the elevation of each point in the 3D point cloud with respect to the robot. We then use *traversability estimation* to evaluate the traversable regions in the map by considering the robot footprint and step height. The robot considers the region as an obstacle if (1) the robot's step height is less than the elevation of the point cloud and (2) the distance between these elevation clouds is less than the robot's footprint. Hence, we construct a 2D traversability map of free space the robot can traverse. Figure 5 shows parts of the traversability map (the collision regions are black and the free spaces are white).

To simulate a real-world environment where humans' privacy regions might interfere with planning, we randomly sample a group of three or two humans' locations and orientations in three types of rooms, the offices, the hallway, and the commons. We use F-formations to create the locations

 https://dx.doi.org/10.1109/RO-MAN57019.2023.10309501

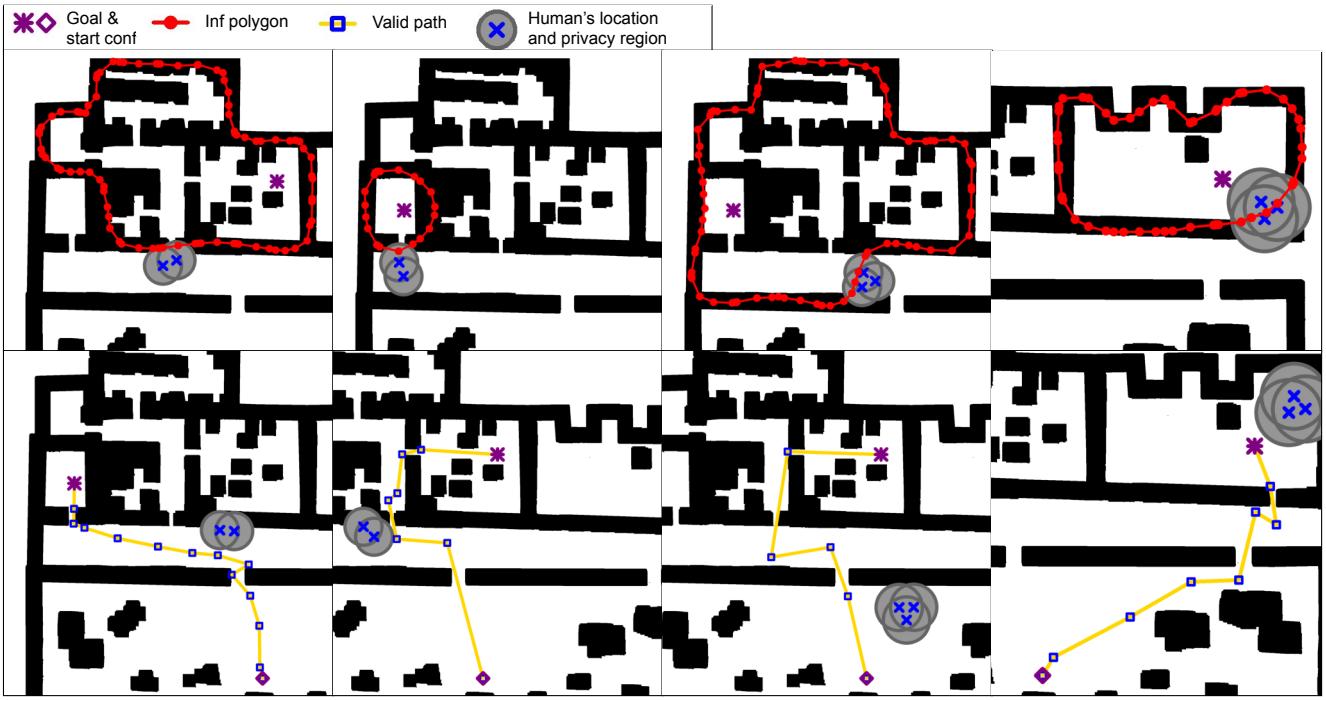

Fig. 5: We run 50 trials for each command (closet, office 1, office 2) and each privacy context (office 1, hallway, commons) with randomized human groups' location. Figures showing eight representative experimental results among the 450 trials. Top: Four plan infeasible cases caused by the humans' physical privacy regions, infeasibility polygons are constructed. Bottom: Four cases with valid paths, the paths avoid entering the humans' privacy regions.

of humans inside a group [77]. We use one type of three-person group formation and one type of two-person group formation (see Figure 5 for marked human groups' locations). To emphasize context changes, we use three different privacy region radii for the three types of location-based context. Humans in the office, the hallway, and the commons have privacy regions' radii of 1.5, 1.0, and 1.25 meters respectively. Further studies are needed to precisely determine the $\mathcal{PR}$ functions and ranges of social navigation robots in different contexts, which is a separate body of work.

In the experiments, the robot starts in the commons, and has three different goal locations associated with three commands, (1) *Go to the closet* (whose constituent referring expression can be resolved to the room at the leftmost end of the hall), (2) *Go to office number 2* (whose constituent referring expression can be resolved to the room in the middle), and (3) *Go to the office number 1* (whose constituent referring expression can be resolved to the room at the rightmost end of the hall). With no humans in the environment, the robot can travel to all three rooms. When random human groups are added, paths to the goal locations might be infeasible. We run 50 trials for each command and each location-based context.

The experiments show successful performance on all three validation utterances in all 450 trials. Figure 5 shows visualizations automatically generated to explain the robots' privacy-grounded decisions. When humans' privacy regions make the planning problem infeasible, the system constructs infeasibility proof visualizations (Figure 5, top) to explain the failure. When the planning problem is feasible with humans' privacy regions, the planner generates valid paths for the robot to execute, and visualizations of those valid paths (Figure 5, bottom). Table I shows the counts and average runtime of the valid/infeasible cases for each context. Human groups' locations in the hallway create the most infeasible cases. Constructing infeasibility proofs in the infeasible cases takes slightly longer than finding a path in the valid cases. The overall average planning time is 0.15 seconds, which makes dynamic real-time applications possible in the future.

Infeasible vs Valid Runtime Result, mean±std (s)				
Location context		*Both*	*Infeasible*	*Valid*
Office	*Counts*	150	6	144
	Runtime (s)	0.25±2.04	4.71±9.93	0.06±0.04
Commons	*Counts*	150	0	150
	Runtime (s)	0.06±0.04	NAN	0.06±0.04
Hallway	*Counts*	150	19	131
	Runtime (s)	0.15±0.33	0.66±0.74	0.08±0.09
All	*Counts*	450	25	425
	Runtime (s)	0.15±1.19	1.63±4.90	0.07±0.06

TABLE I: Counts and runtime (mean±std) results of infeasible vs valid cases for each location based context.

V. Conclusion and Future Work

We have presented an integrated systems approach to enable failure explanation in privacy-sensitive robot navigation tasks. Now that we have shown that failure explanations can be visualized in the forms of maps containing obstacle regions and infeasibility polygons, future work should explore the usability and interpretability of these visualizations to users, as displayed on a physical screen [78] or using Augmented Reality techniques [79], [80], [81]. In future work, we are

also interested in combining natural language with these visualizations into multimodal explanations.

Our approach also has several limitations. First, our current approach does not apply to dynamic scenarios. This could be addressed in future work through constant re-planning. While this would address situations where previously infeasible scenes become feasible, the converse is not necessarily true: If a command is issued while a motion is feasible, but the situation changes, it may be too late for the robot to issue a rejection. In such cases, a privacy violation may yet occur. This raises interesting questions as to how such violations might be mitigated and which entities might be held accountable. For robots in medical settings, researchers have considered whether patients should be instructed on how to disable the robot's recording features to protect their own privacy. Similarly, institutions could implement predefined procedures for the treatment of accidental recordings of nonconsenting individuals, consistent with existing regulations [14].

Second, further research is needed to improve the generalizability of our approach across a wide variety of contexts. This includes exploring different methods of parameterizing privacy regions, considering how those methods are differently perceived by users, and exploring how F-formation detection [82] might be used to automatically detect when humans are in fact in conversation rather than merely standing near each other.

Third, future work should consider how the privacy risks presented by the sensing capabilities needed to enable our approach might themselves be addressed. Our approach inherently requires determination of where people are located, and could require information about those people if social roles are taken into account. While these capabilities do not inherently require *identifying* specific individuals, this could nevertheless itself present a privacy risk. Future work should thus explore how these risks can be enumerated (sensitive so social and structural context[83]), and the techniques for mitigating those risks (use of physical tokens to eliminate the need for certain visual sensor inputs; limited sensing capabilities; or data use policies) that are most compatible with stakeholder values and priorities[31].

Finally, the work presented in this paper could be combined with the task-planning oriented work presented in previous work [17]. Moreover, the method presented in this paper should be evaluated with live human subjects to better understand how people will make sense of and react to both robots' privacy-sensitive navigation behaviors (for those whose privacy would otherwise be violated) and robots' privacy-sensitive command rejections (for those whose commands would violate others' privacy).

REFERENCES

[1] N. Tintarev and R. Kutlak, "Sassy-making decisions transparent with argumentation and natural language generation," in *IUI 2014 Workshop: Interacting with Smart Objects*, 2014.

[2] T. Chakraborti, S. Sreedharan, Y. Zhang, and S. Kambhampati, "Plan explanations as model reconciliation: Moving beyond explanation as soliloquy," in *Proceedings of IJCAI*, 2017.

[3] R. Korpan and S. L. Epstein, "Toward natural explanations for a robot's navigation plans," in *HRI WS on Explainable Robotic Systems*, 2018.

[4] B. Seegebarth, F. Müller, B. Schattenberg, and S. Biundo, "Making hybrid plans more clear to human users-a formal approach for generating sound explanations," in *Proc. ICAPS*, 2012.

[5] M. Fox, D. Long, and D. Magazzeni, "Explainable planning," *IJCAI Workshop on Explainable AI*, 2017.

[6] S. Sreedharan, T. Chakraborti, and S. Kambhampati, "Balancing explicability and explanation in human-aware planning," in *AAAI Fall Symposium on AI-for-HRI*, 2017.

[7] D. Das, S. Banerjee, and S. Chernova, "Explainable ai for robot failures: Generating explanations that improve user assistance in fault recovery," in *Proc. ACM/IEEE Int'l Conf. Human-Robot Interaction*, 2021.

[8] S. Sreedharan, S. Srivastava, D. Smith, and S. Kambhampati, "Why can't you do that hal? explaining unsolvability of planning tasks," in *International Joint Conference on Artificial Intelligence*, 2019.

[9] C. Lutz, M. Schöttler, and C. P. Hoffmann, "The privacy implications of social robots: Scoping review and expert interviews," *Mobile Media & Communication*, vol. 7, no. 3, pp. 412–434, 2019.

[10] C. Lutz and A. Tamó-Larrieux, "The robot privacy paradox: Understanding how privacy concerns shape intentions to use social robots," *Human-Machine Communication Journal (HMC)*, 2020.

[11] E. Sedenberg, J. Chuang, and D. Mulligan, "Designing commercial therapeutic robots for privacy preserving systems and ethical research practices within the home," *Int'l Journal of Social Robotics*, 2016.

[12] 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," *Int'l Jour. Social Robotics*, 2020.

[13] C. Lutz and A. Tamò-Larrieux, "Do privacy concerns about social robots affect use intentions? evidence from an experimental vignette study," *Frontiers in Robotics and AI*, vol. 8, p. 63, 2021.

[14] S. Jeong, B. O'Connell, L. Anderson, and S. Graca, "Deploying social robots in pediatric hospitals: What needs to be considered?" in *HRI '15 Workshop on The Emerging Policy and Ethics of HRI*, 2015.

[15] L. Brandeis and S. Warren, "The right to privacy," *Harvard law review*, vol. 4, no. 5, pp. 193–220, 1890.

[16] B.-J. Koops, B. C. Newell, T. Timan, I. Skorvanek, T. Chokrevski, and M. Galic, "A typology of privacy," *U. Pa. J. Int'l L.*, vol. 38, 2016.

[17] R. B. Jackson, S. Li, S. B. Banisetty, S. Siva, H. Zhang, N. T. Dantam, and T. Williams, "An integrated approach to context-sensitive moral cognition in robot cognitive architectures," in *International Conference on Intelligent Robots and Systems*. IEEE/RSJ, 2021.

[18] S. Li and N. T. Dantam, "Learning proofs of motion planning infeasibility," in *RSS*, 2021.

[19] S. Sohrabi, J. A. Baier, and S. A. McIlraith, "Preferred explanations: Theory and generation via planning." in *AAAI*, 2011.

[20] M. Lomas, R. Chevalier, E. V. Cross, R. C. Garrett, J. Hoare, and M. Kopack, "Explaining robot actions," in *ACM/IEEE international conference on Human-Robot Interaction*, 2012.

[21] G. Briggs, T. Williams, R. B. Jackson, and M. Scheutz, "Why and how robots should say 'no'," *Int'l Journal of Social Robotics*, 2021.

[22] G. Briggs and M. Scheutz, ""Sorry, I Can't Do That": Developing mechanisms to appropriately reject directives in human-robot interactions," in *AAAI Fall Symposia*, 2015.

[23] S. Bringsjord, K. Arkoudas, and P. Bello, "Toward a general logicist methodology for engineering ethically correct robots," *IEEE Intelligent Systems*, vol. 21, no. 4, pp. 38–44, 2006.

[24] L. M. Pereira and A. Saptawijaya, "Modelling morality with prospective logic," *Int'l Jour. Reasoning-based Intelligent Systems*, 2009.

[25] T. Ågotnes, W. Van Der Hoek, J. A. Rodríguez-Aguilar, C. Sierra, and M. J. Wooldridge, "On the logic of normative systems." in *IJCAI*, vol. 7, 2007, pp. 1175–1180.

[26] L. Dennis, M. Fisher, M. Slavkovik, and M. Webster, "Formal verification of ethical choices in autonomous systems," *Robotics and Autonomous Systems*, vol. 77, pp. 1–14, 2016.

[27] R. C. Arkin, "Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture," in *Proc. ACM/IEEE Int'l Conf. Human robot interaction*, 2008.

[28] T. Williams, Q. Zhu, R. Wen, and E. J. de Visser, "The confucian matador: three defenses against the mechanical bull," in *Comp.ACM/IEEE Int'l Conf. on Human-Robot Interaction (alt.HRI)*, 2020, pp. 25–33.

[29] Q. Zhu, T. Williams, B. Jackson, and R. Wen, "Blame-laden moral rebukes and the morally competent robot: A confucian ethical perspective," *Science and Engineering Ethics*, vol. 26, no. 5, 2020.

[30] M. C. Nussbaum, "Creating capabilities: The human development approach and its implementation," *Hypatia*, vol. 24, no. 3, 2009.

[31] J. A. Leydens and J. C. Lucena, *Engineering justice: Transforming engineering education and practice*. John Wiley & Sons, 2017.

[32] R. Calo, "Robots and privacy. robot ethics: The ethical and social implications of robotics," 2010.

[33] M. K. Lee, K. P. Tang, J. Forlizzi, and S. Kiesler, "Understanding users perception of privacy in human-robot interaction," in *ACM/IEEE Int'l Conf. on Human-Robot Interaction (HRI)*. IEEE, 2011.

[34] A. Chatzimichali, R. Harrison, and D. Chrysostomou, "Toward privacy-sensitive human–robot interaction: Privacy terms and human–data interaction in the personal robot era," *Paladyn, Journal of Behavioral Robotics*, vol. 12, no. 1, pp. 160–174, 2021.

[35] U. D. of Health, H. Services, *et al.*, "Hipaa security guidance," *Retrieved on December*, vol. 28, p. 6, 2006.

[36] P. Bevilacqua, M. Frego, E. Bertolazzi, D. Fontanelli, L. Palopoli, and F. Biral, "Path planning maximising human comfort for assistive robots," in *IEEE Conf. on Control Applications (CCA)*, 2016.

[37] S. M. Bhagya, P. Samarakoon, M. A. Viraj, J. Muthugala, A. G. Buddhika, P. Jayasekara, and M. R. Elara, "An exploratory study on proxemics preferences of humans in accordance with attributes of service robots," in *Int'l Symp. Rob. Hum. Interactive Comm.*, 2019.

[38] M. Joosse and V. Evers, "Lost in proxemics : Spatial behavior for cross-cultural hri," in *HRI Workshop on Culture Aware Robotics*, 2014.

[39] E. T. Hall, *The silent language*. Anchor books, 1959, vol. 948.

[40] J. Ginés, F. Martín, D. Vargas, F. J. Rodríguez, and V. Matellán, "Social navigation in a cognitive architecture using dynamic proxemic zones," *Sensors*, vol. 19, no. 23, 2019.

[41] J. G. Clavero, F. M. Rico, F. J. R. Lera, J. M. G. Hernández, and V. M. Olivera, "Defining adaptive proxemic zones for activity-aware navigation," *CoRR*, vol. abs/2009.04770, 2020. [Online]. Available: https://arxiv.org/abs/2009.04770

[42] I. Kostavelis, D. Giakoumis, S. Malassiotis, and D. Tzovaras, "Human aware robot navigation in semantically annotated domestic environments," in *Int'l Conf. Universal Access in Human-Computer Int.*, 2016.

[43] N. Mitsunaga, C. Smith, T. Kanda, H. Ishiguro, and N. Hagita, "Adapting robot behavior for human–robot interaction," *IEEE Transactions on Robotics*, vol. 24, no. 4, pp. 911–916, 2008.

[44] R. Mead and M. J. Matarić, "Autonomous human–robot proxemics: socially aware navigation based on interaction potential," *Autonomous Robots*, vol. 41, no. 5, pp. 1189–1201, 2017.

[45] C. Park, J. Ondřej, M. Gilbert, K. Freeman, and C. O'Sullivan, "Hi robot: Human intention-aware robot planning for safe and efficient navigation in crowds," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016.

[46] R. Mead and M. Matarić, "Probabilistic models of proxemics for spatially situated communication in hri," in *HRI 2014*, 2014.

[47] S. M. LaValle, "Rapidly-exploring random trees: A new tool for path planning," Computer Science Deptartment, Iowa State University, Tech. Rep. TR-98-11, October 1998.

[48] S. Karaman and E. Frazzoli, "Sampling-based algorithms for optimal motion planning," *IJRR*, vol. 30, no. 7, pp. 846–894, 2011.

[49] L. E. Kavraki, P. Svestka, J.-C. Latombe, and M. H. Overmars, "Probabilistic roadmaps for path planning in high-dimensional configuration spaces," *T-RO*, vol. 12, no. 4, pp. 566–580, 1996.

[50] A. Varava, J. F. Carvalho, F. T. Pokorny, and D. Kragic, "Caging and path non-existence: a deterministic sampling-based verification algorithm," in *Robotics Research*. Springer, 2020, pp. 589–604.

[51] J. Basch, L. J. Guibas, D. Hsu, and A. T. Nguyen, "Disconnection proofs for motion planning," in *Int'l Conf. Robotics and Autom.*, 2001.

[52] Z. McCarthy, T. Bretl, and S. Hutchinson, "Proving path non-existence using sampling and alpha shapes," in *International Conference on Robotics and Automation*. IEEE, 2012, pp. 2563–2569.

[53] L. Zhang, Y. J. Kim, and D. Manocha, "A hybrid approach for complete motion planning," in *International Conference on Intelligent Robots and Systems*. IEEE/RSJ, 2007, pp. 7–14.

[54] ——, "A simple path non-existence algorithm using c-obstacle query," in *Algorithmic Foundation of Robotics VII*. Springer, 2008.

[55] M. Scheutz, T. Williams, E. Krause, B. Oosterveld, V. Sarathy, and T. Frasca, "An overview of the distributed integrated cognition affect and reflection diarc architecture," *Cognitive architectures*, 2019.

[56] M. Scheutz, G. Briggs, R. Cantrell, E. Krause, T. Williams, and R. Veale, "Novel mechanisms for natural human-robot interactions in the diarc architecture," in *Workshops at the twenty-seventh aaai conference on artificial intelligence*, 2013.

[57] M. Scheutz, "Ade: Steps toward a distributed development and runtime environment for complex robotic agent architectures," *Applied Artificial Intelligence*, vol. 20, no. 2-4, pp. 275–304, 2006.

[58] M. Scheutz, E. Krause, B. Oosterveld, T. Frasca, and R. Platt, "Spoken instruction-based one-shot object and action learning in a cognitive robotic architecture," in *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, 2017.

[59] J. Dzifcak, M. Scheutz, C. Baral, and P. Schermerhorn, "What to do and how to do it: Translating natural language directives into temporal and dynamic logic representation for goal management and action execution," in *Int'l Conference on Robotics and Automation*, 2009.

[60] M. Steedman and J. Baldridge, "Combinatory categorial grammar," *Non-Transform. Syntax: Formal and Expl. Models of Gram.*, 2011.

[61] J. K. Gundel, N. Hedberg, and R. Zacharski, "Cognitive status and the form of referring expressions in discourse," *Language*, 1993.

[62] T. Williams and M. Scheutz, "A givenness hierarchy theoretic approach," *The Oxford handbook of reference*, p. 457, 2019.

[63] T. Williams, G. Briggs, B. Oosterveld, and M. Scheutz, "Going beyond literal command-based instructions: Extending robotic natural language interaction capabilities," in *29th AAAI Conf. on Artificial Intel.*, 2015.

[64] G. Briggs, T. Williams, and M. Scheutz, "Enabling robots to understand indirect speech acts in task-based interactions," *Journal of Human-Robot Interaction*, vol. 6, no. 1, pp. 64–94, 2017.

[65] T. Williams and M. Scheutz, "A framework for resolving open-world referential expressions in distributed heterogeneous knowledge bases," in *Proc. AAAI Conference on Artificial Intelligence*, 2016.

[66] ——, "Power: A domain-independent algorithm for probabilistic, open-world entity resolution," in *Int'l Conf. Intel. Rob. & Sys. (IROS)*, 2015.

[67] T. Williams, E. Krause, B. Oosterveld, and M. Scheutz, "Towards givenness and relevance-theoretic open world reference resolution," in *RSS WS on models and reps. for natural human-robot comm.*, 2018.

[68] T. Williams, S. Acharya, S. Schreitter, and M. Scheutz, "Situated open world reference resolution for human-robot dialogue," in *11th ACM/IEEE Int'l Conf. on Human-Robot Interaction (HRI)*, 2016.

[69] T. Brick and M. Scheutz, "Incremental natural language processing for hri," in *ACM/IEEE Int'l Conf. on Human-robot interaction*, 2007.

[70] S. M. LaValle, *Planning algorithms*. Cambridge university press, 2006.

[71] J. J. Kuffner and S. M. LaValle, "RRT-connect: An efficient approach to single-query path planning," in *ICRA*, vol. 2, 2000, pp. 995–1001.

[72] P. Gao and H. Zhang, "Long-term Place Recognition through Worst-case Graph Matching to Integrate Landmark Appearances and Spatial Relationships," in *Int'l Conf. Robotics and Automation (ICRA)*, 2020.

[73] S. Siva and H. Zhang, "Omnidirectional multisensory perception fusion for long-term place recognition," in *International Conference on Robotics and Automation*, 2018.

[74] S. Siva, Z. Nahman, and H. Zhang, "Voxel-Based Representation Learning for Place Recognition Based on 3D Point Clouds," in *International Conference on Intelligent Robots and Systems*, 2020.

[75] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Conference on Computer Vision and Pattern Recognition*, 2016.

[76] P. Fankhauser, M. Bloesch, and M. Hutter, "Probabilistic terrain mapping for mobile robots with uncertain localization," *Robotics and Automation Letters*, vol. 3, no. 4, pp. 3019–3026, 2018.

[77] T. M. Ciolek and A. Kendon, "Environment and the spatial arrangement of conversational encounters," *Sociological Inquiry*, vol. 50, 1980.

[78] A. K. Pandey and R. Gelin, "A mass-produced sociable humanoid robot: Pepper: The first machine of its kind," *IEEE Robotics & Automation Magazine*, vol. 25, no. 3, pp. 40–48, 2018.

[79] T. Williams, D. Szafir, T. Chakraborti, and H. Ben Amor, "Virtual, augmented, and mixed reality for human-robot interaction," in *Comp. 2018 ACM/IEEE Int'l Conference on Human-Robot Interaction*, 2018.

[80] R. S. Andersen, O. Madsen, T. B. Moeslund, and H. B. Amor, "Projecting robot intentions into human environments," in *Int'l Symp. Robot and Human Interactive Communication (RO-MAN)*, 2016.

[81] A. Hietanen, R. Pieters, M. Lanz, J. Latokartano, and J.-K. Kämäräinen, "Ar-based interaction for human-robot collaborative manufacturing," *Robotics and Computer-Integrated Manufacturing*, vol. 63, 2020.

[82] F. Setti, C. Russell, C. Bassetti, and M. Cristani, "F-formation detection: Individuating free-standing conversational groups in images," *PloS one*, vol. 10, no. 5, 2015.

[83] K. Winkle, D. McMillan, M. Arnelid, M. Balaam, K. Harrison, E. Johnson, and I. Leite, "Feminist human-robot interaction: Disentangling

power, principles and practice for better, more ethical hri," in *Proc. ACM/IEEE Int'l Conf. Human-Robot Interaction*, 2023.