Examining the Effectiveness of Obfuscatory Planning Strategies through Human Observation

Dakota Sullivan
dsullivan8@wisc.edu
University of Wisconsin-Madison
Madison, Wisconsin, USA

David Porfirio
dporfiri@gmu.edu
George Mason University
Fairfax, Virginia, USA

Bilge Mutlu
bilge@cs.wisc.edu
University of Wisconsin-Madison
Madison, Wisconsin, USA

Laura M. Hiatt
laura.m.hiatt.civ@us.navy.mil
U.S. Naval Research Laboratory
Washington, DC, USA

Abstract

Robots are increasingly relied upon for task completion in privacy-critical human environments. In these environments, it is imperative that a robot's potentially sensitive goals remain *obfuscated*. To address this need, a substantial amount of literature has proposed methods for obfuscatory task planning. These works make many attempts to experimentally or analytically determine whether agents can conceal their goals from observers. While these works make guarantees that resulting plans will conceal an agent's goals, they are often *only* theoretical. Within this work, we develop three obfuscatory task planning strategies inspired by prior literature to evaluate with human observers ($N = 160$). Our preliminary results show that observers struggle to identify a robot's goals at similar levels regardless of whether obfuscatory or optimal task planning strategies are employed. These findings call into question the purported benefits of many obfuscatory task planning strategies.

CCS Concepts

• **Security and privacy** → **Privacy protections**.

Keywords

Privacy, Obfuscation, Task Planning

ACM Reference Format:
Dakota Sullivan, David Porfirio, Bilge Mutlu, and Laura M. Hiatt. 2026. Examining the Effectiveness of Obfuscatory Planning Strategies through Human Observation. In *Companion Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI Companion '26), March 16–19, 2026, Edinburgh, Scotland, UK*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3776734.3794469

1 Introduction

As robots increasingly enter into public and private environments, we gain a better understanding of the risks associated with their day-to-day use. One such risk is *inference-based privacy leakage* [14], where observers infer private information from public signals. Consider a delivery robot that leaves a coffee shop and makes a delivery to the front steps of a home at the same time every day. An observer might infer that the owner of the home enjoys daily coffee. These privacy risks become more acute in critical domains such as health and eldercare, where robots are finding greater use and acceptance [2, 7, 10]. Consider a user requesting a robot to *"deliver medication A to room 3212."* As the robot carries out the task, an observer may infer the robot's goals through its actions and conclude that the patient in room 3212 has a particular medical condition without the robot ever intentionally or directly disclosing that information.

Robots are particularly susceptible to this type of privacy leakage: while a human may discern whether someone is attempting to follow or observe them, existing robot platforms do not possess this level of awareness. Consequently, robots must be designed to protect the sensitive information of the users they serve. While existing works have proposed strategies to reduce privacy violations stemming from data collection, processing, and disclosure [3, 5, 12, 15, 19], these works do not consider unintentional leakage of sensitive information inferred from observing a robot's actions [17].

In this work, we provide a novel approach to obfuscating a robot's goals based purely on plan generation. We present three planning strategies, inspired by existing work on "deceptive" robot-arm motion [4], designed to generate plans that reduce the likelihood of privacy leakage through a robot's actions. The *alternating* strategy generates plans that interleave the actions required to satisfy multiple goals and, therefore, disconnects what may otherwise be predictable consecutive actions to complete individual goals. The *multitasking* strategy generates plans that encourage actions that progress multiple goals. Through this strategy, an action's progression of one goal may be concealed by its progression of another. The *diverting* strategy focuses on adding extraneous actions to the task plan, potentially increasing the set of possible goals the robot could be achieving. While the alternating and multitasking strategies are applicable when a robot is pursuing multiple goals, the diverting strategy can be applied even when the robot has only a single goal.

We evaluate the effectiveness of plans generated by each obfuscation strategy, as well as baseline strategies (*i.e.,* sequential and non-sequential optimal goal completion), in terms of their ability to conceal a robot's goals. To do so, we conducted a large online user study ($N = 160$) asking participants to identify a robot's delivery goals in a simulated robot environment. Through this evaluation,

HRI Companion '26, Edinburgh, Scotland, UK

ACM ISBN 979-8-4007-2321-6/2026/03
https://doi.org/10.1145/3776734.3794469

we examine how effectively these strategies obfuscate goals when observed by a human. Our results show that observers struggle to identify a robot's goals whether the robot's actions are generated by an obfuscatory or baseline strategy. We conclude by discussing the implications of these results for goal obfuscation in human-robot interaction. The contributions of this work are as follows:

- Three task planning strategies for goal obfuscation;
- A comparative evaluation of each strategy's effectiveness;
- Insights into the application of "deceptive" task planning in human-robot interaction.

2 Technical Approach

We next present our planning-based approach to goal obfuscation, detailing the algorithmic approach of each strategy. Figure 1 depicts a visual comparison of each strategy to a non-obfuscatory approach.

2.1 Alternating Strategy

The *alternating* strategy generates plans that interleave actions that satisfy different goals (see Figure 1A). This strategy is based on the switching robot motion-generation strategy [4], in which a robot arm switches between moving towards two different items until it settles on a final target to approach. This motion deceives observers by making it difficult to determine which item the arm is targeting.

We adapt this concept to obfuscate the robot's goals while executing actions. Our approach to the alternating strategy constrains plan generation such that no goal is progressed by more than one action in a row. Therefore, while many non-deceptive optimal task plans may naturally alternate task progression to some extent, the alternating strategy ensures that the entire plan is interleaved. Formally, an alternating plan will consist of actions $a_1, a_2, ...a_n$ with the constraint that no a_i and a_{i+1} exist such that $goal(a_i) = goal(a_{i+1})$.

2.2 Multitasking Strategy

Our second obfuscation strategy, *multitasking*, involves the progression of multiple goals by a single action (see Figure 1B). This strategy is inspired by the ambiguity robot motion-generation strategy from Dragan et al. [4], in which a robot's end effector moves equidistantly toward two items for as long as possible before moving directly toward the true target. The robot's goal is ambiguous for most of its trajectory because both targets appear equally likely.

In the spirit of this strategy and similar approaches [8, 11], the *multitasking* strategy maximizes the number of actions in the plan that progress multiple goals. While a non-obfuscatory, optimal plan may include multitasking if it makes the plan more efficient, the multitasking strategy will do so even when the overall plan may be longer. The multitasking strategy operates by discounting actions' costs for each "extra" goal that they simultaneously progress, encouraging multitasking actions to occur whenever reasonably possible. Thus, formally, given a set of goals G, a multitasking plan will find the sequence $a_1, a_2, ...a_n$ that maximizes the number of actions where $|goal(a_i)| = |G|$, then $|goal(a_i)| = |G| - 1$, and so on.

2.3 Diverting Strategy

The third obfuscation strategy, *diverting*, adds extraneous actions to the task plan (see Figure 1C). This strategy is inspired by the exaggeration robot motion-generation strategy, which generates end-effector paths that move toward a false target and away from a true target before moving directly toward the true target at the end of the trajectory [4]. The robot's goal appears to be the false target because such a path toward the true target appears unreasonable.

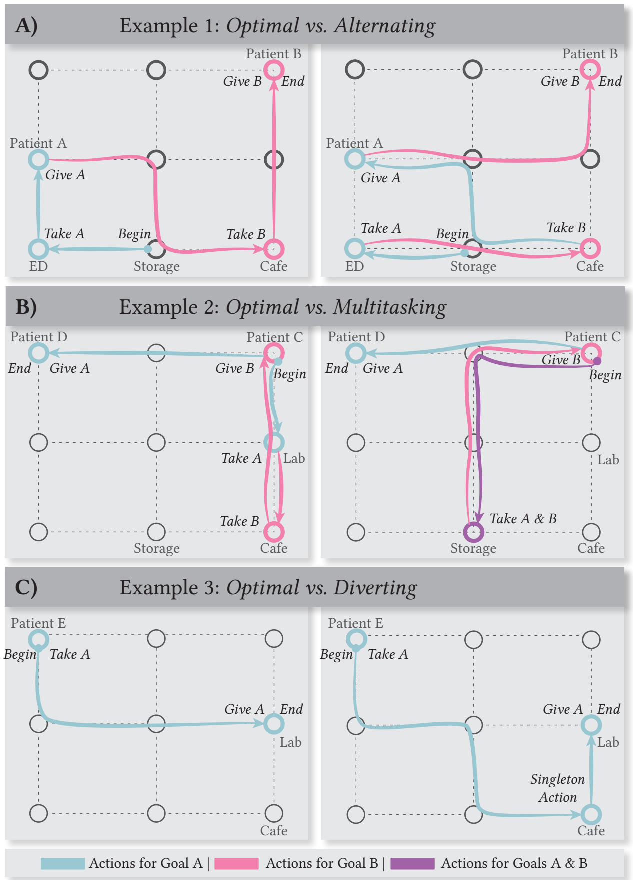

Figure 1: ***Deception Strategies*** – We present a simple comparison between a non-deceptive, optimal plan and a plan generated using one of each of the *alternating*, *multitasking*, and *diverting* strategies.

We apply this concept to the robot's task execution by creating diversions during the robot's progression of its goals. Our approach to the *diverting* strategy adds a singleton action as an extra "stop" whenever the robot is on the way to a stop that is in service of its goals. Formally, a diverting plan will include actions $a_1, a_2, ...a_n$ with the constraint that a singleton action a_i where $|goal(a_i)| = 0$ exists before each action a_j where $loc(robot, pre(a_j)) = loc(robot, post(a_j))$, and $|goal(a_j)| > 0$. These singleton actions have the additional constraint that $loc(robot, pre(a_i)) \neq loc(robot, pre(a_j))$ for any action a_j preceding or following a_i.

3 Evaluation

To evaluate our proposed planning strategies, we conducted a large online user study ($N = 160$) where participants observed a robot completing delivery tasks in a simulated hospital domain. This study was approved by the authors' Institutional Review Board, and all participants provided consent prior to participation.

3.1 Simulator & Videos

We use the Planning Domain Definition Language (PDDL) [6] to implement our planning strategies. Additional non-obfuscatory baseline strategies were also developed, including an *optimal* strategy (*i.e.,* purely using optimal planning) and a *sequential* strategy (*i.e.,* generating plans that fully achieve one goal before beginning the next). We use SymK as our planning engine [16].

Each goal a plan achieves involves a robot delivering an item to a person or storage container within a hospital environment. We developed a total of 18 unique delivery goals and sampled 20 sets of three goals for use in our evaluation. Each obfuscatory and baseline strategy was then used to generate a plan for each set of goals (*i.e.,* 100 plans in total). To visualize each of our generated plans, we utilize a top-down two-dimensional simulator that depicts the robot as it makes deliveries. As the robot executes actions, a text box appears describing the action and what items are present in the robot's current location, when relevant.

To present these simulations to participants, a total of 100 videos were created, each showing one of the five planning strategies being used to achieve one of the 20 goal sets. Each video allows the viewer to see the locations the robot visits and the individuals with whom the robot interacts. The items that are transferred between the robot and others are not visible because an observer with this information would immediately be able to identify the robot's goals.

3.2 Procedure

To begin the Qualtrics survey, participants read a brief introduction to the study's purpose and a consent form. By continuing with the study, participants consented to participate. Next, participants studied a set of images explaining the robot behaviors they would view in the simulator. These images were followed by two practice questions similar to those that would be encountered after each video. Participants then viewed the series of five videos, each depicting the robot completing a set of three delivery tasks using one of the planning strategies. Each participant viewed the five strategies in a random order, and the goal set achieved by any strategy was randomly assigned. After each video, participants were asked to report the three deliveries they believed the robot was making using a series of drill-down options. For each delivery, participants chose the provider and recipient of the item, and the item itself. Next, participants rated their agreement with the statement "*I feel confident that I was able to identify the correct deliveries*" on a five-point Likert scale. Finally, participants provided demographic data.

3.3 Participants

This study utilized Prolific to recruit 160 participants. All participants were located within the U.S. and fluent in English. The participant population was composed of 46.25% women, 51.25% men, 1.88% non-binary individuals, and 0.63% individuals who selected "prefer not to say". Participant ages ranged from 19 to 82 years ($M = 41.14$, $SD = 11.92$). All participants who successfully completed the survey received $7.57 for the 30-minute study.

3.4 Goal Recognition

For each participant-video pair, we first scored the accuracy of participants' predicted robot delivery goals. These results were then sorted based on the planning strategy used in each video. The results of this analysis can be seen in Figure 2. *Alternating* and *diverting* produced plans that led to the greatest percentage of participants being unable to identify any goals (74.38% and 75.00%, respectively). These strategies were followed by *multitasking* and *optimal*, which yielded zero goals identified by 69.38% and 67.50% of participants, respectively. The *sequential* strategy led to the lowest percentage of participants recognizing zero goals with 44.38%.

We additionally conducted a Friedman omnibus test and Wilcoxon pairwise comparisons with Holm–Bonferroni correction to assess whether the differences in recognized goals between strategies are statistically significant. Our results show a statistically significant difference between each of the four non-sequential strategies and *sequential*. While the differences between pairs of strategies are highly significant ($p < 0.001$), no other statistically significant differences can be found between comparisons.

These results highlight a major difference between our baseline planning strategies. While the plans generated through obfuscation strategies led to fewer recognized goals than plans generated with the *optimal* strategy, the difference is minimal. As compared to the *sequential* strategy, however, the difference is far more pronounced. While plans generated through the *sequential* strategy are relatively easy to follow (*i.e.,* one full task is completed before the next), those generated through the *optimal* strategy are much more opaque. These results suggest that when multiple tasks are available, optimal planning can be used to conceal a robot's goals from an observer without specialized obfuscation algorithms.

3.5 Confidence

Participants also reported confidence levels in their goal recognition responses for each video viewed. We sorted these confidence levels based on the planning strategies to which they corresponded. These results can be seen in Figure 3. Like our goal recognition analysis, we see a similar pattern emerge among participant confidence levels. *Diverting* and *alternating* produced plans that yielded the lowest participant confidence with 58.75% and 50.63% of participants, respectively, responding "strongly disagree", "somewhat disagree", or "neither agree nor disagree" to feeling confident about their goal recognition responses. *Optimal* and *multitasking* followed with 47.50% and 42.50% participants, respectively, again either disagreeing or neither agreeing nor disagreeing. *Sequential* led to the lowest percentage of participants with low confidence at 36.88%.

We again conducted a Friedman omnibus test and Wilcoxon pairwise comparisons with Holm–Bonferroni correction to assess whether statistically significant differences existed in confidence levels between strategies. Significant differences were found between the *diverting* strategy and each of the *multitasking* ($p < 0.001$), *optimal* ($p < 0.01$), and *sequential* ($p < 0.001$) strategies. This outcome may be due to the fact that plans generated from the *diverting* strategy are longer than those of any other strategy, and therefore demand more of an observer. Interestingly, no significant difference was found between *diverting* and *alternating*. Given that the *alternating* strategy forces individual components of

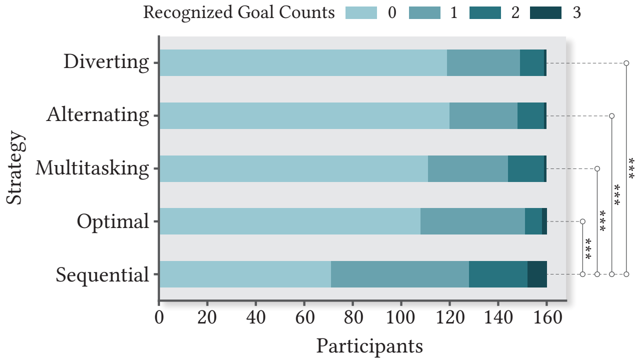

Figure 2: ***Goal Recognition*** – This plot shows the number of goals participants recognized in each evaluation condition. Statistically significant differences were found between each non-sequential strategy and *sequential* ($p < 0.05^*$, $p < 0.01^{**}$, $p < 0.001^{***}$).

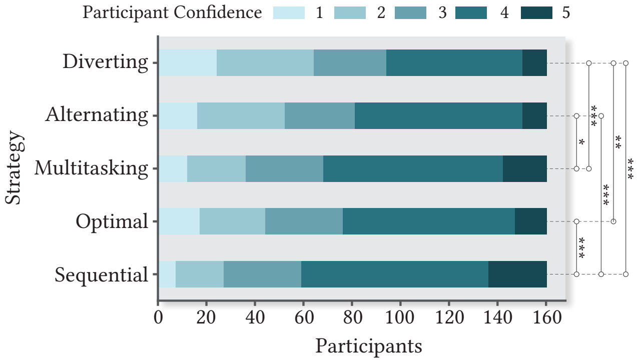

Figure 3: ***Participant Confidence*** – This plot shows participants' confidence levels through a five-point Likert scale (*i.e.,* 1 indicates low confidence and 5 indicates high confidence). Statistically significant differences were found between several pairwise comparisons ($p < 0.05^*$, $p < 0.01^{**}$, $p < 0.001^{***}$).

plans to be disconnected through interleaving, these strategies may lead to similarly high levels of cognitive demand and low levels of confidence, though by different means.

Additionally, significant differences were shown to exist between the *sequential* planning strategy and both the *optimal* ($p < 0.001$) and *alternating* ($p < 0.001$) strategies. These results are unsurprising given that sequential goal completion is easier to comprehend than task execution through the other four strategies. Surprisingly, no significant difference was seen between *sequential* and *multitasking*, but a significant difference was found between *multitasking* and *alternating* ($p < 0.05$). Given their differences in function and goal recognition outcomes, the cause of these results is unclear. Further analysis is needed to understand these differences in confidence.

4 Discussion

From our findings, we share two key insights: (1) while prior works have made theoretical guarantees about the effectiveness of obfuscation strategies, evaluation with human observers is needed to understand whether these guarantees hold; and (2) optimal planning may be sufficient for goal obfuscation under specific conditions.

4.1 Necessity of Human Evaluation

While many obfuscation strategies are present in the planning literature [1, 9, 13, 18, 20], few works have conducted evaluations with users [4]. Although purely computational evaluations can provide insights into strategy performance, without human evaluation, true effectiveness is unknown. Additionally, the implicit assumption of these works is that without such obfuscation strategies, a human observer will be able to infer sensitive information. Whether or not that assumption is true, the degree to which these strategies provide benefits is unknown. As we saw in our user evaluation, our obfuscation strategies appeared to perform well in that participants recognized few goals. In the context of our *optimal* strategy, however, the benefits of the obfuscation strategies appear reduced.

Furthermore, human observers may form highly variable assumptions or perceptions about an agent or task during their observation. As we have seen in this work, for example, some individuals may confidently draw incorrect conclusions. While we were not able to capture participants' thought processes, it is also possible that some arrived at the correct conclusions about the robot's deliveries for the wrong reasons. Purely computational or simulated approaches to assessing the effectiveness of these strategies do not capture these critical elements about an observer. Therefore, we argue that additional human-subjects research is needed to better understand existing and future obfuscation strategies. This effort is of particular importance in the context of sensitive human environments where private information may be leaked to observers.

4.2 Obfuscation Through Optimal Planning

The results of this work suggest that optimal planning can be used to conceal a robot's actions without obfuscation algorithms. When an optimal planner generates a plan's sequence of actions, it necessarily interleaves unrelated components of multiple tasks when such tasks are available. This process often achieves the very same outcome we attempt to enforce in the *alternating* strategy: disconnecting predictable consecutive actions to complete individual goals. In doing so, a robot's goals can become much more difficult to comprehend. This outcome can be seen in Figure 2 when comparing the *optimal* and *sequential* strategies.

Under the appropriate conditions (*e.g.,* multiple tasks are available and the adversary is an average human observer), a purely optimal strategy may produce plans that are sufficiently difficult to discern without the added cost that may be required under obfuscation strategies. In cases where only a single task is available, or the adversary is a sophisticated observer with computational assistance, specialized obfuscation strategies can be utilized. We hope that future research can further evaluate these concepts by investigating the specific contexts in which specialized strategies substantially outperform optimal planning. Future efforts could focus on observers with technological assistance and expand to include additional critical domains. Through such investigation, we can better understand the tradeoffs of planning strategies and further reflect on existing computational evaluations.

References

[1] Sara Bernardini, Fabio Fagnani, and Santiago Franco. 2020. An Optimization Approach to Robust Goal Obfuscation. In *Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning*. 119–129. doi:10.24963/kr.2020/13

[2] Elizabeth Broadbent, Rebecca Stafford, and Bruce MacDonald. 2009. Acceptance of healthcare robots for the older population: Review and future directions. *International journal of social robotics* 1 (2009), 319–330. doi:10.1007/s12369-009-0030-6

[3] Anna Chatzimichali, Ross Harrison, and Dimitrios Chrysostomou. 2021. Toward privacy-sensitive human–robot interaction: Privacy terms and human–data interaction in the personal robot era. *Paladyn, Journal of Behavioral Robotics* 12, 1 (2021), 160–174. doi:10.1515/pjbr-2021-0013

[4] Anca Dragan, Rachel Holladay, and Siddhartha Srinivasa. 2014. An Analysis of Deceptive Robot Motion. In *Proceedings of Robotics: Science and Systems (RSS '14)*. https://www.ri.cmu.edu/pub_files/2014/7/deception.pdf

[5] Eduard Fosch-Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. 2020. Gathering expert opinions for social robots' ethical, legal, and societal concerns: Findings from four international workshops. *International Journal of Social Robotics* 12, 2 (2020), 441–458. doi:10.1007/s12369-019-00605-z

[6] Maria Fox and Derek Long. 2003. PDDL2. 1: An extension to PDDL for expressing temporal planning domains. *Journal of artificial intelligence research* 20 (2003), 61–124. doi:10.1613/jair.1129

[7] Amanda Hall, Uba Backonja, Ian Painter, Maya Cakmak, Minjung Sung, Timothy Lau, Hilaire Thompson, and George Demiris. 2017. Acceptance and perceived usefulness of robots to assist with activities of daily living and healthcare tasks. *Assistive Technology* 31 (11 2017). doi:10.1080/10400435.2017.1396565

[8] Anagha Kulkarni, Matthew Klenk, Shantanu Rane, and Hamed Soroush. 2018. Resource bounded secure goal obfuscation. In *AAAI Fall Symposium on Integrating Planning, Diagnosis and Causal Reasoning*. https://anaghak.github.io/files/parc-obfuscation-fss.pdf

[9] Anagha Kulkarni, Siddharth Srivastava, and Subbarao Kambhampati. 2019. A unified framework for planning in adversarial and cooperative environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2479–2487. doi:10.1609/aaai.v33i01.33012479

[10] I. H. Kuo, J. M. Rabindran, E. Broadbent, Y. I. Lee, N. Kerse, R. M. Q. Stafford, and B. A. MacDonald. 2009. Age and gender factors in user acceptance of healthcare robots. In *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*. 214–219. doi:10.1109/ROMAN.2009.5326292

[11] Zhengshang Liu, Yue Yang, Tim Miller, and Peta Masters. 2021. Deceptive reinforcement learning for privacy-preserving planning. *arXiv preprint arXiv:2102.03022* (2021). doi:10.48550/arXiv.2102.03022

[12] Christoph Lutz, Maren Schöttler, and Christian Pieter Hoffmann. 2019. The privacy implications of social robots: Scoping review and expert interviews. *Mobile Media & Communication* 7, 3 (2019), 412–434. doi:10.1177/2050157919843961

[13] Peta Masters and Sebastian Sardina. 2017. Deceptive Path-Planning.. In *IJCAI*. 4368–4375. doi:10.24963/ijcai.2017/610

[14] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2019. Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning. In *2019 IEEE Symposium on Security and Privacy (SP)*. 739–753. doi:10.1109/SP.2019.00065

[15] Rahul Shome, Zachary Kingston, and Lydia E. Kavraki. 2023. Robots as AI Double Agents: Privacy in Motion Planning. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2861–2868. doi:10.1109/IROS55552.2023.10341460

[16] David Speck, Robert Mattmüller, and Bernhard Nebel. 2020. Symbolic Top-k Planning. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020)*, Vincent Conitzer and Fei Sha (Eds.). AAAI Press, 9967–9974. doi:10.1609/aaai.v34i06.6552

[17] Dakota Sullivan and Bilge Mutlu. 2025. Protecting User Data Through Privacy-Sensitive Robot Design. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 1891–1893. doi:10.1109/HRI61500.2025.10974013

[18] Hideaki Takahashi and Alex Fukunaga. 2024. On the Transit Obfuscation Problem. doi:10.48550/arXiv.2402.07420

[19] Brian Tang, Dakota Sullivan, Bengisu Cagiltay, Varun Chandrasekaran, Kassem Fawaz, and Bilge Mutlu. 2022. Confidant: A Privacy Controller for Social Robots. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 205–214. doi:10.1109/HRI53351.2022.9889540

[20] Kai Xu, Yunxiu Zeng, Long Qin, and Quanjun Yin. 2020. Single real goal, magnitude-based deceptive path-planning. *Entropy* 22, 1 (2020), 88. doi:10.3390/e22010088

Received 2025-12-08; accepted 2026-01-12