Handling Intentional Obstructions in Social Navigation using Foundation Models

Einav Shpiro
Computer Science Department
Bar Ilan University, Israel
Email: einav.shenwald@live.biu.ac.il

Reuth Mirsky
Computer Science Department
Bar Ilan University, Israel
Email: mirskyr@cs.biu.ac.il

***Abstract*—One of the most studied interactions in social navigation is a collision between a human and a robot. Most of these studies focus on collision avoidance: shifting away from close pedestrians or staying still until the conflict is resolved. However, to act socially, avoidance is not always the desired behavior. Consider a staff member in a hospital blocking a delivery robot's path to type in a new delivery request. The robot should not steer away but rather stay put or even get closer to the person. Our recent research paper provided a novel perspective on obstructions in social navigation [1]. That work presented a solution named NIMBLE: Navigational Intentions Model for BLocking Estimation and provided a pipeline for handling intentional obstructions that is general enough to allow for varying implementations while maintaining a clear inference process for intentional obstructions. This paper proposes several approaches to extend NIMBLE using Foundation Models (FM) to improve human-robot interaction.**

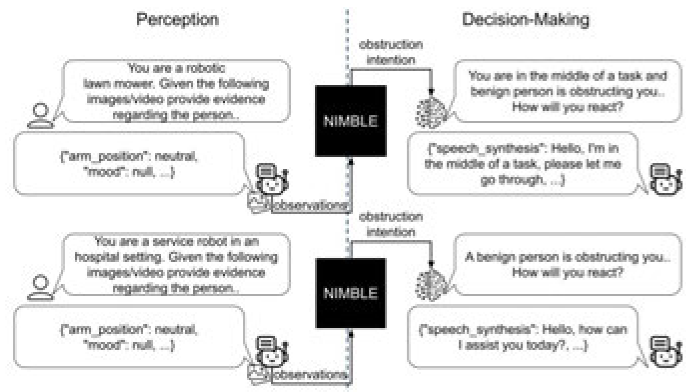

Fig. 1: Proposed extensions for NIMBLE

I. Introduction

Service and assistive mobile robots will soon become integral to our daily tasks. However, there are still many challenges that need to be overcome. One such challenge is the issue of **obstruction**, which refers to a situation where a pedestrian blocks the path of a robot and prevents the robot from reaching its goal. In most cases, this interaction is handled as a by-product of *collision avoidance*, the common policy of navigation algorithms that assume that obstruction should be avoided altogether. The robot should pass smoothly around people, usually without interacting with them directly [12, 14]. However, obstruction is a common issue that happens regularly in human interactions, where people often wish to interact rather than avoid collisions with each other. Similarly, this interaction should be acknowledged and addressed in human-robot interactions. Consider a delivery robot in a hospital on a collision course with a nurse. If the nurse looks directly at the robot, it is more likely that the nurse wishes to type in a delivery request into the robot's interface rather than collide with it. In this case, the robot should not avoid the situation but rather help by steering toward the nurse. Additionally, the robot must also recognize different *intentions* behind obstructions to determine the appropriate response. Contrary to the nurse example, hospitals are also environments known to be prone to cases of aggressive behaviors [7]. When someone approaches the robot intending to damage it, it should move away quickly or sound an alarm. Lastly, if the person is looking to take a selfie with the robot, it should probably act friendlier, either stopping and posing for the selfie or kindly ignoring it and moving on with its task.

We define obstruction as Shpiro and Mirsky [17]. Given two agents, an actor (the pedestrian) and an observer (the robot):

Definition 1: An **obstruction** is a conflict where the actor intentionally wishes the observer to decrease its speed to zero.

This line of work provides a unique perspective on obstruction handling because it deals with unexpected intentions, potentially unrelated to navigation, but that can affect a mobile robot's task [16, 18]. Specifically, Singamaneni et al. [18] highlighted the importance of context reasoning during a navigational task. Foundation Models (FMs) excel in context-dependent tasks and thus have a clear benefit for intentional reasoning in social navigation. This work investigates methods to utilize the rapid advancements in Large Language Models (LLMs) and Vision Language Models (VLMs) to enhance the obstruction handling pipeline. While FMs were recently utilized in the context of enhancing robot capabilities [4, 19], here we focus on three tasks specific to navigational interactions:

1) **Design:** LLMs can enhance the designers' and developers' capabilities by providing suggestions and code segments. For example, Vemprala et al. [25] examined GPT in creating high-level functions for robotic use.
2) **Perception:** VLMs can be used online to improve scene

[1]Published in Technological Advancements of HRI (TAHRI) 2024, https://www.tahri.org/program

reasoning and evidence acquisition. For example, in our obstruction recognition framework, NIMBLE may be able to be given more accurate observations.

3) **Decision-Making:** Dynamic decision-making and the generation of suitable responses are often complex tasks highly dependent on context. As such, LLMs and VLMs may be leveraged to analyze a situation and decide on the desired course of action.

While the first task is an intriguing engineering problem, this paper focuses on the latter two, as delineated in Figure 1.

II. NIMBLE: Navigational Intentions Model for BLocking Estimation

Our obstruction deliberation algorithm, Navigational Intentions Model for BLocking Estimation (NIMBLE), consists of three components for three stages of obstruction detection:

1) **Conflict detection** predicts for a pedestrian if they are likely to be on a collision course with the robot.
2) **Conflict classification** classifies whether the conflict is intentional when a pedestrian is identified to conflict with the robot (e.g. if the person is facing the robot).
3) **Intention recognition** identifies the intentions behind each detected obstruction using goal recognition. NIMBLE recognizes three different intentions: *Authorized*, a person with whom an interaction is desired to complete the task (such as the nurse in the hospital example); *Benign*, a person that does not seem to mean harm to the robot, but an interaction with them will not serve the robot's predefined aims; and *Malicious*, a person aiming to harm the robot or hinder its task execution.

The system's functionality is illustrated in Figure 2, depicting the system's decision points flow. It starts at the top with continuous data extraction (such as camera inputs), followed by active evidence acquisition (e.g., pose estimation, speed). Once a conflict is detected and classified as intentional, the intention is inferred based on the collected evidence.

NIMBLE leverages computer vision and machine learning techniques for robust obstruction detection and intention inference. A deep vision model detects people within the robot's environment. Whenever a person is detected, the system analyzes their skeleton data (when available) to assess posture. Additionally, it gathers distance information to estimate the person's velocity. When the person enters the robot's "personal space" [6, 13], these real-time observations are fed into a Bayesian network. Based on the combined data, this model classifies the person's intention (benign, authorized, malicious), enabling the robot to gain valuable insight regarding the approaching interaction that will help the robot react appropriately to the given situation.

For enhancing NIMBLE, we suggest incorporating a Vision Language Model (VLM) within its evidence acquisition module. This VLM extracts semantic observations, offering a deeper understanding of the situation beyond raw sensory data. These enriched observations serve two purposes:

1) **Predicting Person Intention:** Nimble leverages the VLM's insights to predict the person's intention, informing the system's overall decision-making.
2) **Informing LLM Reaction Selection:** The VLM's observations are communicated to the LLM. This context assists the LLM in selecting the most appropriate reaction based on the specific situation. The chosen reaction is then converted into predefined tokens representing actionable options for the robot.

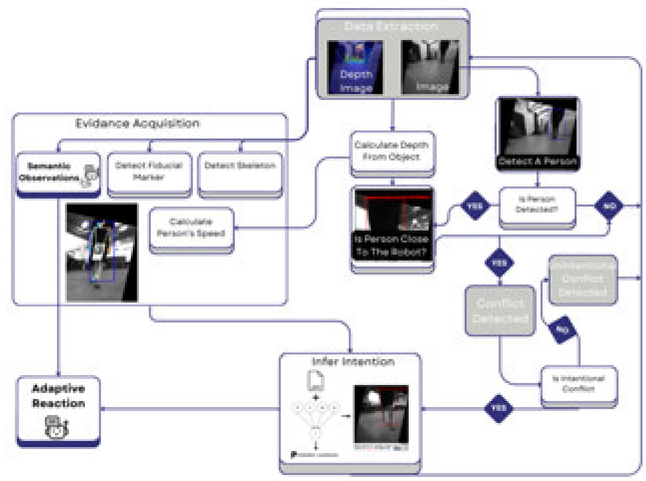

Fig. 2: NIMBLE's visualization with proposed extensions

III. Extending NIMBLE for Perception

Accurate pedestrian intention recognition is crucial for robots to safely navigate dynamic environments. Our previous paper introduced NIMBLE, a baseline solution that leveraged the hardware capabilities of the Boston Dynamics Spot platform and off-the-shelf algorithms to extract observations about pedestrians in the robot's surroundings. To improve NIMBLE's perception capabilities, we can utilize advanced solutions and hardware to achieve more accurate observations. This will also contribute to improving the accuracy of the intention recognition component. We employed the advanced object detection algorithm YOLOv8 [8] for evidence acquisition. Additionally, a foot detector and keypoints information from heels and toes were used to determine if a person was intentionally blocking the robot. However, significant costs and development time are associated with the deployment of advanced hardware and cutting-edge solutions. Therefore, we propose exploring VLMs as an alternative approach. To understand their functionality, VLMs take an image of the robot's surroundings as input and output observations regarding visible pedestrians. These models can be beneficial when searching for particular evidence that is complex or nuanced (e.g., "does the person seem to be angry?"). Moreover, LLMs could be used to provide proactive reasoning: when there is high ambiguity regarding the intention of a pedestrian, the LLM could generate a query and proactively ask the person about their intention.

IV. Extending NIMBLE for Decision-Making

NIMBLE's enhanced perception, detailed previously, allows for a more nuanced understanding of pedestrian intentions. However, effective social interaction requires translating this understanding into actions. Hence, we suggest that the recognition process be followed by a complementary action, or series of actions, suited to the obstructing person's recognized intention. Unlike classic collision avoidance, which relies on objective parameters like distance and velocity, handling obstruction requires considering the human factor. When dealing with people, the robot needs a complex understanding of the environment and current state to assess the situation and determine the best response based on the individual's behavior. A recent technology well-suited to handling such tasks is Foundation Models (FMs). These models, trained on vast amounts of human data, can analyze complex situations using multiple inputs and offer appropriate responses.

We propose the development of a pre-programmed responses library that a robot can employ when encountering an obstruction. These responses will be dynamically selected by a Large Language Model (LLM) based on the specific situational context. To enhance the robot's expressivity during these interactions, we will explore the utilization of various communication modalities, including audio, visual, haptic, and kinesthetic channels (as explored in Su et al. [23]). In this initial phase of our work, we present several potential reaction options that can be employed as the robot's predefined actions:

- **Speech synthesis:** Utilizing speech synthesis as a response modality facilitates smoother communication with users [11] and has been shown to enhance user compliance in certain situations [1].
- **Visual Cues via Light-Emitting Diodes (LEDs):** Employing LEDs to generate visual cues offers an efficient method for the robot to communicate its intentions[5].
- **Human-Inspired Gesture and Movement Generation:** Inspired by human nonverbal communication, the generation of gestures and movements allows the robot to effectively convey specific messages [15, 3].
- **Legible Movement for Intention Communication:** Utilizing clear and intuitive movement patterns for effective intention signaling [2, 9].

Given the robot's ability to execute some or all of these behaviors, this work investigates techniques for prompting the LLM to generate the optimal course of action. This selection will be based on a comprehensive analysis of three key factors: the assigned role of the robot, the environmental setting it operates in, and the specific situational context it encounters.

V. Preliminary Results

Having established the proposals for improving NIMBLE's social interaction capabilities, the following sections present preliminary results from our experiments. These results provide initial insights into the effectiveness of the suggested approach and provide a foundation for further research and development. In the experiments, a higher-ranked FM (GPT-4o) was utilized as both LLM and VLM components.

A. Augmenting Perception

We first evaluated the capability of a VLM to extract useful observations that will enhance the obstruction recognition process. We used high-quality images representing Benign, Authorized, and Malicious behaviors (one image per intention type). Additionally, we utilized images from Shpiro and Mirsky [17] taken during experiments with a Boston Dynamics Spot's camera. Openly available images allowed us to test an ideal setting in which our robotic platform could capture high-resolution and precise images. To challenge the VLM's robustness, we also utilized real images from a robot, typically of lower quality than the controlled setting images.As expected, the VLM could not generate correct observations on most of Spot's images. We specifically asked the model to output a JSON format response to eliminate the noise that an FM can generate by providing cumbersome texts. We further reduced the possible observations to four main observation types – arm position, mood, staff member, and suspicious object – as used in [17]. Given the following prompt, GPT accurately identified the observations from the high-quality images.

Prompt

You are a service robot and the person in the image is blocking your way. Give all the possible evidence regarding the person that will help you to infer the intention of the person provide your observations in the following format: {arm_position: neutral/offensive, mood: neutral / upset, is_staff_member: yes/no, suspicious_object: yes/no}. Return your response in a JSON format. If you can't observe a parameter, put null

B. Augmenting Decision-Making

In this section, we tested the ability of an LLM to provide a coherent course of action for every defined situation.

Prompt

Imagine you are a service robot in a hospital setting, your tasks include transferring medicine to different departments in the hospital. During your work, a pedestrian will block you for one of these three intentions: **Benign** a curious person who is interested in you. **Authorized** a person (usually a staff member) who wishes to initiate one of your defined tasks. **Malicious** a person with the intention to harm you or interfere with your work. You, as a service robot, have the following possible reactions to interact with people: **Speech** generate speech that is relevant to the situation. **LED** use LED for signaling. **Alarm** start an alarm. **Move** apply one of your possible movements.
Given a situation, you need to supply a dictionary with the reactions that you need to generate to resolve the obstruction situation optimally. From now on, please provide your answers in a JSON format.

Below are some prompts describing a potential obstruction scenario and GPT's suggested responses (in gray boxes).

Benign Obstruction

Prompt

A person with a benign intention is blocking you. How will you react?

```
{
    "Speech": "Hello! I am on duty. Could
        you please step aside?",
    "LED": "Blinking Blue"
}
```

Authorized Obstruction

Prompt

A person with an authorized intention is blocking you. How will you react?

```
{
    "Speech": "Hello! How can I assist
      you?",
    "LED": "Solid Green"
}
```

Malicious Obstruction

Prompt

A person with a malicious intention is blocking you. How will you react?

```
{
    "Alarm": "Activate",
    "Move": "Reverse and turn"
}
```

Ambiguous Obstruction

Prompt

A depressed person is blocking you. How will you react?

```
{
    "Speech": "Hello. I can see you might
        be feeling down. Can I assist you
        in any way?",
    "LED": "Solid Yellow"
}
```

GPT was capable of properly responding in most cases, and even when the situation wasn't defined in the original prompt, it generated well. Notably, the responses exhibited slight variations each time, mimicking natural human conversation and making the experience feel more human-like. This characteristic, along with GPT's ability to adapt, suggests exciting possibilities. For instance, a robot encountering a person displaying signs of distress could leverage a VLM to observe the emotional state and collaborate with an LLM to generate an appropriate response, even for unexpected scenarios that the designers hadn't originally anticipated.

C. *Evaluation Methods Suggestions*

One way to quantitatively evaluate the effectiveness of LLM-generated reactions could be through a user study. Such a study would contain various scenarios and possible LLM-generated reactions. Then, users would be asked about their most preferred response. Evaluating a VLM's perception ability in the context of obstructions is challenging due to a lack of labeled data. However, despite this limitation, we have identified two possible approaches. First, we can assess the VLM's capabilities on datasets not directly related to our domain but where the possible observations are somewhat relevant. For example, a dataset for classifying violent/non-violent behaviors ([20]). The second approach involves constructing a new dataset, either simulated or real-life-based, specifically designed for our problem definition [24]. The dataset would consider the interaction of obstructions from the robot's point of view. While this approach might be challenging and time-consuming, it can provide a tailored evaluation for our domain.

VI. Conclusion

This paper reports some preliminary results on using foundation models to improve the performance of robots in obstruction recognition and management as part of social navigation. The strengths of these approaches shine in such challenges, where a slightly different context can lead to different responses. For example, a robot in the middle of an urgent task might not respond to a benign obstruction the same way it would when it is not in a hurry.

However, FM use in such settings should be carefully designed. Most importantly, the tendency of these models to generate false information or imprecise descriptions can lead to undesired robot behavior or even raise safety issues [22]. One potential approach is to deploy existing verification and diagnosis techniques and adapt them to LLM-robot systems [10, 21] Additionally, given these models' tendency for verbosity, a balance between a short and concise answer and a robust one is needed.

Acknowledgments

This work took place in the Goal Optimization using Learning and Decision-making (GOLD) Lab at Bar Ilan University. The Israeli Innovation Authority supported this research as part of the HRI Consortium.

References

[1] Franziska Babel, Johannes Kraus, Philipp Hock, and Martin Baumann. Verbal and non-verbal conflict resolution strategies for service robots. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1030–1037, 2022. doi: 10.1109/RO-MAN53752.2022.9900807.

[2] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 301–308. IEEE, 2013.

[3] Kanghui Du, Drazen Brsčić, Yuyi Liu, and Takayuki Kanda. Can't you see i am bothered? human-inspired suggestive avoidance for robots. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 184–193, 2024.

[4] Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, Brian Ichter, Danny Driess, Jiajun Wu, Cewu Lu, and Mac Schwager. Foundation models in robotics: Applications, challenges, and the future. *ArXiv*, abs/2312.07843, 2023. URL https://api.semanticscholar.org/CorpusID:266191787.

[5] Michael Fulton, Chelsey Edge, and Junaed Sattar. Robot communication via motion: A study on modalities for robot-to-human communication in the field. *J. Hum.-Robot Interact.*, 11(2), feb 2022. doi: 10.1145/3495245. URL https://doi.org/10.1145/3495245.

[6] Edward Twitchell Hall. *The hidden dimension*, volume 609. Garden City, NY: Doubleday, 1966.

[7] Rh Harwood. How to deal with violent and aggressive patients in acute medical settings. *Journal of the Royal College of Physicians of Edinburgh*, 47(2): 176–182, 2017. ISSN 14782715, 20428189. doi: 10.4997/JRCPE.2017.218.

[8] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolo, 01 2023. URL https://ultralytics.com. If you use this software, please cite it using the metadata from this file.

[9] Ryo Kitagawa, Yuyi Liu, and Takayuki Kanda. Human-inspired motion planning for omni-directional social robots. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pages 34–42, 2021.

[10] Michael Kuchnik, Virginia Smith, and George Amvrosiadis. Validating large language models with relm. *Proceedings of Machine Learning and Systems*, 5:457–476, 2023.

[11] Matthew Marge, Carol Espy-Wilson, Nigel G. Ward, Abeer Alwan, Yoav Artzi, Mohit Bansal, Gil Blankenship, Joyce Chai, Hal Daumé, Debadeepta Dey, Mary Harper, Thomas Howard, Casey Kennington, Ivana Kruijff-Korbayová, Dinesh Manocha, Cynthia Matuszek, Ross Mead, Raymond Mooney, Roger K. Moore, Mari Ostendorf, Heather Pon-Barry, Alexander I. Rudnicky, Matthias Scheutz, Robert St. Amant, Tong Sun, Stefanie Tellex, David Traum, and Zhou Yu. Spoken language interaction with robots: Recommendations for future research. *Computer Speech & Language*, 71:101255, 2022. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2021.101255. URL https://www.sciencedirect.com/science/article/pii/S0885230821000620.

[12] Christoforos Mavrogiannis, Francesca Baldini, Allan Wang, Dapeng Zhao, Pete Trautman, Aaron Steinfeld, and Jean Oh. Core challenges of social robot navigation: A survey. *ACM Transactions on Human-Robot Interaction*, 12(3):1–39, 2023.

[13] Ross Mead and Maja J Matarić. Perceptual models of human-robot proxemics. In *Experimental Robotics*, pages 261–276. Springer, 2016.

[14] Reuth Mirsky, Xuesu Xiao, Justin Hart, and Peter Stone. Conflict avoidance in social navigation—a survey. *ACM Transactions on Human-Robot Interaction*, 13(1):1–36, 2024.

[15] Kazuki Mizumaru, Satoru Satake, Takayuki Kanda, and Tetsuo Ono. Stop doing it! approaching strategy for a robot to admonish pedestrians. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 449–457, 2019. doi: 10.1109/HRI.2019.8673017.

[16] Astrid Marieke Rosenthal-von der Pütten, David Sirkin, Anna Abrams, and Laura Platte. The forgotten in hri: Incidental encounters with robots in public spaces. pages 656–657, 03 2020. doi: 10.1145/3371382.3374852.

[17] Einav Shpiro and Reuth Mirsky. Recognition and identification of intentional blocking in social navigation. In *Proceedings of the 2024 International Symposium on Technological Advances in Human-Robot Interaction*, TAHRI '24, page 101–110, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400716614. doi: 10.1145/3648536.3648549. URL https://doi.org/10.1145/3648536.3648549.

[18] Phani Teja Singamaneni, Pilar Bachiller-Burgos, Luis J. Manso, Anaís Garrell, Alberto Sanfeliu, Anne Spalanzani, and Rachid Alami. A survey on socially aware robot navigation: Taxonomy and future challenges. *The International Journal of Robotics Research*, 0(0):02783649241230562, 0. doi: 10.1177/02783649241230562. URL https://doi.org/10.1177/02783649241230562.

[19] David Sobrín-Hidalgo, Miguel Ángel González-Santamarta, Ángel Manuel Guerrero-Higueras, Francisco Javier Rodríguez-Lera, and Vicente Matellán-Olivera. Enhancing robot explanation capabilities through vision-language models: a preliminary study by interpreting visual inputs for improved human-robot interaction, 2024.

[20] Mohamed Mostafa Soliman, Mohamed Hussein Kamal, Mina Abd El-Massih Nashed, Youssef Mohamed Mostafa, Bassel Safwat Chawky, and Dina Khattab.

Violence recognition from videos using deep learning techniques. In *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*, pages 80–85, 2019. doi: 10.1109/ICICIS46948.2019.9014714.

[21] Tran Cao Son, William Yeoh, Roni Stern, and Meir Kalech. Multi-agent planning and diagnosis with commonsense reasoning. In *Proceedings of the Fifth International Conference on Distributed Artificial Intelligence*, pages 1–9, 2023.

[22] Kaya Stechly, Matthew Marquez, and Subbarao Kambhampati. Gpt-4 doesn't know it's wrong: An analysis of iterative prompting for reasoning problems. *arXiv preprint arXiv:2310.12397*, 2023.

[23] Hang Su, Wen Qi, Jiahao Chen, Chenguang Yang, Juan Sandoval, and Med Amine Laribi. Recent advancements in multimodal human–robot interaction. *Frontiers in Neurorobotics*, 17:1084000, 2023.

[24] Nathan Tsoi, Mohamed Hussein, Jeacy Espinoza, Xavier Ruiz, and Marynel Vázquez. Sean: Social environment for autonomous navigation. In *Proceedings of the 8th international conference on human-agent interaction*, pages 281–283, 2020.

[25] Sai Vemprala, Rogerio Bonatti, Arthur Bucker, and Ashish Kapoor. Chatgpt for robotics: Design principles and model abilities. *IEEE Access*, PP:1–1, 01 2024. doi: 10.1109/ACCESS.2024.3387941.