**Reviewer 4 (AC)**

**Meta Review**

Summary
This paper focuses on the problem of finding socially acceptable locations for a robot to wait for a shared elevator. Using procedural content generation, the authors generate plausible waiting scenes, and train models indicating acceptable locations for a robot to wait. The models leverage feature maps representing social norms described by prior sociological studies. The paper contains an evaluation of these models, indicating their strengths and weaknesses

Strengths
+All reviewers found the paper to focus on a topic of great interest that deserves further investigation.
+Reviewers also found the paper to be well written, easy to follow, and containing nice graphics.
+All reviewers commended the authors for the depth and rigor of their analysis.
+R3 identified the consideration of OOD / real-world test set as one of the strong points of the paper.
+The idea of cleverly leveraging procedural content generation in HRI is important. Although this is not the first work to add this element to an HRI pipeline, I find this element quite appealing.

Weaknesses
-R2 and R3 find that the specific problem that the authors tackle makes assumptions that require additional motivation. Both reviewers make important points about the implications of these assumptions for real-world deployments.
-R2 and R3 point out the mismatch between authors' claims and reported results, citing specific statements (including the title of the paper) that the authors should consider softening. In particular, it looks like the performance of the best models is not satisfactory as the authors themselves admit. Additionally, both R2 and R3 do not find the technical approach to stand as a strong novel contribution.
-R2 expresses doubts about casting the problem of waiting area selection as a classification problem, rather than a costmap-based planning problem, explaining that no justification for that design choice was given.
-R2 and R3 find that the hand-designed selection of several hyperparameters affecting the results and generalization (e.g., feature maps) is not appropriately discussed. R3 suggests an ablation study as a way to mitigate that concern.
-All reviewers mention the importance of releasing the collected dataset with the community if the authors have approval for that.
-R1 casts doubt on the authors' claim that no dataset with interactions next to elevators exists, citing a recent preprint.
-R1 brings up the question of the scalability of the proposed approach, given the time it takes to perform annotations.
-R1 asks for more details on the data augmentation component of the approach.

Rebuttal suggestions
All reviewers provided in-depth feedback to the paper. In the rebuttal, the authors are encouraged to:
*Revisit some of their claims and more accurately reflect on their contributions, especially addressing the relevant comments by R2 and R3; clearly outline the components of the technical approach that are novel, responding to R2 and R3.
*Add more details and justification about the specific problem setting, the assumptions and simplifications, addressing the detailed comments by R2 and R3.
*Add more context and justification about the technical approach (see comments on costmap vs classification by R2), data augmentation and scalability by R1.
*Consider the release of their datasets with the community. All reviewers agreed that this would substantially strengthen the paper's contribution.

Recommendation
The problem is interesting and relevant to the HRI community, and the paper is well written. However, there are concerns related to the technical novelty (R2, R3), the implementation of the approach (R1, R2), and the assumptions underlying the problem setting (R2, R3). It appears that the authors are also overstating the importance of some of their results. Echoing some of the issues raised by R2 and R3, I am inclined towards rejection but I would like to see how the authors will reframe their contributions in their rebuttal.

**Reviewer 1 (reviewer)**

**Contribution**

This paper explores the idea of "learning to wait for the elevator" where a robot predicts socially acceptable locations to wait for the elevator in the presence of other agents waiting for an elevator. The key contribution claimed in this paper is an algorithm to procedurally generate an artificial scene of waiting for an elevator, which is then hand-annotated by a domain expert to mark regions that are socially acceptable for the robot to wait in. Additionally, the authors propose/utilize features that encode social norms for this shared elevator task. The annotated data is then utilized by a learning-based approach to predict the socially acceptable waiting positions given a scene and target mission. The takeaway message is that a UNet outperforms other machine learning approaches in predicting the socially acceptable location. The contributions claimed in this paper at the start do align with the paper's key findings towards the end.

**Detailed Review**

Summary:

This paper proposes a learning-based framework to predict socially acceptable locations for a robot to wait for an elevator, in a location shared by humans. A procedural environment generation framework is introduced to general synthetic scenes that are annotated by a domain expert for two cases - yielding and priority. Feature maps of social norms are also introduced, which are used as inputs to the learned function (B+N case). The key finding is that using a UNet-based architecture, using informative inputs (social norm features) helps in predicting the socially acceptable locations for the robot to wait.

Strengths:

The paper is well written. The motivation of the problem explored in this paper, approach taken, and the analysis of the results are performed well. I like that the paper takes a structured approach to the problem, first by generating data and considering multiple ways to learn the target, also using explainable machine learning frameworks such as decision trees. The use of social norm features is well motivated in the context of using learning-based solutions that are known to perform well when the input is more informative of the task. The experiments performed seem convincing that learning from annotated expert data is one way to solve this problem.

Detailed Comments:

1) It would be a great value add to this paper if the authors had shown a real-robot experiment with a questionnaire for humans, to support the contribution of this work.
2) Scalability of the proposed solution is questionable - The authors do mention that annotation takes a long time per sample (2 minutes) and requires a domain expert in the loop, hence, it might be a good extension to think about learning socially acceptable positions for this task using unlabelled large-scale video data mined from the internet, such as from youtube / bilibili
3) The authors claim that "there is no publicly available dataset specifically containing
scenes of people waiting for the elevator". I encourage the authors to check out MuSoHu [1] which does contain scenes of waiting for an elevator.
4) The data augmentation used for training the UNet is unclear. The paper says "We sample the location of the robot", but it is not clear what that means. More explanation is needed. Also, it seems there can be other forms of data augmentation used - such as rotate-in-place. CNNs are not agnostic to rotation augmentations, hence, it may be possible that using rotation augmentation can improve the results, which is not explored in this work.

[1] : "Toward Human-Like Social Robot Navigation: A Large-Scale, Multi-Modal, Social Human Navigation Dataset", Nguyen et al.

Recommendation:

Although the paper has some weaknesses such as no real-world robot experiment with a human study, the contributions in the paper are interesting. I also think the authors should consider additional forms of data augmentations such as

rotate-in-place to see if they get the best from the learning-based approach in this low-data regime. Also, the authors must explain the existing data augmentation used in the paper. I would be inclined to accept this paper with the additional data-augmentation experiments as explained above.

**Reviewer 2 (reviewer)**

**Contribution**

This work explores the question of where robots should wait for elevators in spaces shared with people.
The authors claim that they present a method (PCG+annotation+norm feature maps) to tackle the low-data regime posed by this problem. However, their results do not indicate that these constitute a successful solution, even by their own admission. This is more an exploratory paper that could contribute to the field if their data was made public but there are no indications that it will.

**Detailed Review**

Summary:
This work explores the question of where robots should wait for elevators in spaces shared with people.
The authors claim that they present a method (PCG+annotation+norm feature maps) to tackle the low-data regime posed by this problem. However, their results do not indicate that these constitute a successful solution, even by their own admission. This is more an exploratory paper that could contribute to the field if their data was made public but there are no indications that it will.

Strengths:
This work builds upon strong prior research (especially from Gallo et al), taking a logical next step.
It has in-depth analysis of their results, great figures and tables.
Good formulation of position-based features.
Operationalization of norms as feature maps is a good, interpretable formulation.
I appreciate the post-hoc qualitative analysis of the top models' results to investigate beyond metrics

Areas for improvement:

As the authors themselves note, the performance of their best models is not satisfactory and their research questions are answered in the negative. However, the authors claim success in the paper ( e.g. "We have shown with the experiments the successful combination of PCG and social norms maps in tackling a data scarce robotics application"). These claims must be softened and the paper must stand as an initial exploration of this problem. The title should also reflect that. A commitment to release their data and PCG method publicly would also make the authors' contributions stronger.

In my opinion, the weakest part of this paper is posing the problem of waiting area selection as a classification problem rather than a cost-map based planning problem. Given that the authors use several handcrafted feature maps, I feel a planning problem that combines them to form a costmap is the most natural way to find waiting poses for the robot. This would be a fully interpretable and tuneable solution which would respond in a predictable way to changes in the individual feature maps.
There is no discussion or justification for this choice. In a way, the non-trained baseline is a version of generating a costmap one would plan with but is lacking in details so it is hard to tell if it was a well-executed baseline or not.

Handcrafted parameters: both the choice of feature maps and each map's particular implementation are hyper parameters (how far behind a person to not stand etc.). Inevitably, this excludes some important factors such as: robot orientation while waiting (only robot location is considered here even though full robot pose is mentioned in Sec3 while setting up the problem), reachability/traversability of selected goal locations. I would have liked to see some discussion of what is included and why as well as what is excluded and why.

Motivation for the importance of this problem is a bit lacking. What are the consequences of improper or sub-optimal waiting? The more interesting (and possibly more common) problem occurs when agents (robots/people) are waiting at

a bay of elevators for one among N elevators. Even if this is not explored, the relationship among the single elevator and multi-elevator case deserves some discussion.

Detailed comments and suggestions for improvement:

In the future, please use [sigconf, review] for your latex document prior to submission so that line numbers are used in the submitted manuscript. This helps detailed comments be more precise.

Figure 2: some annotated waiting areas seem smaller than the robot footprint (e.g. P area in Fig 2b). Why is that?

For the Real Elevator set, were manual annotators authors on this work as well? I am not sure about the validity of system builders being the annotators since they cannot annotate towards the system design instead of annotating independently.

“All the models are short of this number however, rejecting RQ1” -- I think the authors are equating RQs with hypotheses which they are not.

Some typos:
Efficient boarding: this map incentives -> incentivizes
Visibility cone: this map incentives -> incentivizes
Jaccard indexes -> indices

Recommendation

I would argue for probably rejecting this paper.

**Reviewer 3 (reviewer)**

**Contribution**

a. I believe the paper has three main findings and contributions: 1) the formalization of “robot waiting for an elevator in a socially compliant manner” as a segmentation task, along with an expertly labeled dataset to support this; 2) successful and grounded incorporation of hand-crafted, social norm priors into models for increased performance on this task; and 3) usage of procedurally generated training data resulting in a relatively small sim2real gap between out-of-distribution and real-world test sets (although still lacking in overall performance).
b. The paper claims to primarily contribute the following points: 1) an operationalization of social norms in the context of waiting positions for a robot in a shared elevator; 2) remediation of data scarcity via procedural content generation of initially unlabelled scenes; and 3) experimental results on deep learning and traditional ML approaches which have some promise, although noted limitations in overall performance.
c. These key contributions do seem to be aligned, but I would encourage the authors to make an effort to release their created dataset, as the paper didn’t mention whether this would occur or not. As the authors noted there is no equivalent dataset available, the benefits of doing so could be quite high.

**Detailed Review**

**Summary:**

The paper presents an approach for creating and leveraging a dataset for the task of robots finding socially acceptable waiting zones by an elevator. The authors create a labeled dataset by leveraging procedurally generated content with a hand-crafted rejection sampling scheme, having experts annotate the acceptable zones for a given scenario, and creating test sets of both out-of-distribution (OOD) scenes and real-world scenes from video recordings. The paper focuses extensively on embedding prior norms via theoretically grounded feature maps for both social compliance (e.g.

proxemics) as well as task-specific heuristics (e.g. efficient boarding). When experimenting with both grid-based image-segmentation techniques (like U-Net) and cell-based feature vector approaches (like Random Forest classifiers), the authors achieved a small sim2real gap between the OOD and real-world test sets. However the overall Jaccard Index achieved by the authors' best approaches was still relatively low, given what the authors believe should be possible.

**Strengths:**

- The authors create a dataset for evaluating the specified waiting zone segmentation task, where no such dataset has been created before.
- The evaluation methodology of having an OOD and a real-world test set is very good, and assuages fears of model overfitting to training set.
- The procedurally-generated content and rejection sampling scheme seems to produce reasonable enough data to train a model with limited sim2real gap.
- The incorporation of the prior belief maps helps for all models and tasks, justifying its inclusion.
- The Likert scale and compliance-score analyses provide additional insight besides just raw intersection-over-union (IoU) numbers, validating the model's Jaccard Index with expert's perception of the model's performance and justifying the selection of feature maps.

**Weaknesses:**

- The overall task of finding waiting spots for an elevator and applying hand-crafted features therein seems limited, compared to broader social waiting and navigation tasks.
- The hand-crafted, task-specific feature maps may reduce applicability of author's approach to other tasks.
- Some assumptions in the task seem too strong, such as assuming a static scene (i.e. all people and group locations are fixed throughout), ignoring navigation aspects of social compliance, assuming perfect information and sensing is available to the robot, and only considering scenes with a single elevator across three total room setups.
- A more thorough ablation of which maps are beneficial in downstream training would be beneficial, especially since social and task heuristics are grouped together.
- Splitting the tasks into two extreme missions (i.e. priority and yielding), while established by prior work of Gallo et al. [1], is still a large assumption.
- The overall task performance, according to Jaccard Index, is below what could be reasonably attained, which is particularly concerning given the authors do not provide any novelty in model development.

**Detailed Comments:**

There are many positive aspects of the paper. As discussed, the creation of this elevator-waiting zone dataset is a significant contribution and should be highlighted more; commitment by authors to release this dataset publicly would be a benefit to the HRI community. With the procedurally generated content, although some of the decisions and hand-crafted functions are a bit ad hoc, the fact that it still resulted in decent performance is to be commended. This is also reinforced by the decision to test on an explicitly OOD test set (in addition to the recorded real-world test set), relative to the parameters used to generate the training set. Furthermore, the incorporation of the prior maps in the first place seems to be very effective, especially given that the authors are operating under a limited-data availability paradigm. Thus, on many aspects of the rubric, including track relevance, and theoretical and technical grounding, the contribution is quite decent.

That being said, there are still some concerns that I have with this paper. One key issue revolves around the strong assumptions used when defining and addressing the task. Scenes are assumed to be static, where humans are held as simple, non-responsive obstacles to coordinate amongst. As addressed extensively in Mavrogiannis et al. [2], this is a very limited perspective; in fact, humans navigate and interact in a coordinated, back-and-forth manner, where a robot's actions influence a human's and vice versa. This is exacerbated by the decision to ignore socially compliant navigation as an aspect of this task altogether, both in the labeling of the generated scenes, as well as the model development itself. While the authors do acknowledge this point, simply including additional maps would not address the fundamental lack of dynamicism, where the path a robot takes to a goal could itself affect whether that goal should be considered as a socially compliant waiting zone in the first place. Furthermore, the authors assume that the robot has access to perfect sensing information of peoples' locations and orientation. As discussed in Stoler et al. [3], downstream task

performance of mobile robots (in navigation) is extremely influenced by the perception quality and first-person view errors which occur with realistic sensing; again, the authors mention this point, but leave it as future work which is unsatisfying. These limitations overall reduce the motivation and field-relevance of the contribution.

There are additionally concerns regarding the usage of the prior maps. Specifically, while some norm maps, such as proxemics and visibility cone, seem to be applicable to many social navigation tasks, many of the others (e.g. as far as furthest person, avoid blocking the door, etc.) seem to be constrained to this specific elevator task. While aspects of this could be applied to related tasks, such as joining a waiting group to cross the street at a crosswalk, the usage of these priors limits the generalizability to less constrained social navigation and waiting scenarios. Additionally, although the authors split the set of maps into a “basic” and “basic + norm” set when training models as an ablation, the “norm” set still includes a wide variety of maps, from social priors to task heuristics, making it difficult to draw conclusions as to which maps were the most beneficial. Thus, this overall limits the evaluation fit and rigor of this paper.

The final large issue lies with the model contribution and results themselves. As stated by the authors, they reasonably expected to attain Jaccard Indexes of around 0.5, based on the internal agreement among the experts annotators, and scores on similar tasks, yet their best models attained scores of around 0.25 only. For the image segmentation model, the authors utilize the widely studied, and relatively modern, U-Net, alongside more traditional cell-based approaches such as Random Forests and Decision Trees from scikit-learn. The only model novelty exhibited is using a smaller version of U-Net, with fewer layers and trainable parameters. Because the resulting Jaccard score is less than ideal, this is especially disappointing that the authors only used these off-the-shelf approaches.

**Suggestions for Improvement:**

There are many actionable items which could be performed to increase the quality of this work. Namely, addressing some of the limitations of the task assumptions would go a long way in improving the contribution. For instance, the authors could move away from the limited static perspective of scenarios, and instead simulate more dynamic, interactive scenes which include social navigation. The authors could explore starting with the same goal locations, as determined by the expert labeler in training, but then could incorporate proxy metrics based on the provided ground truth to determine new goal locations in a semi-supervised manner. Additionally, the authors could perform further ablation studies breaking down which maps are beneficial to the model at a finer granularity than just “basic” and “basic + norm”, which would produce more confidence in each maps’ importance in being included in the proposed approach. Finally, to address the performance limitations, the authors could consider a variety of strategies: 1) Using a larger U-Net model, but applying transfer learning from a model pre-trained on a more general semantic segmentation task, to still have effectively fewer trainable parameters; This could also enable using more advanced, modern architectures such as in Gu et al. [4]; 2) Extending the amount of data and variety of room layouts used, to prevent overfitting and allow for better training. Since annotations take ~2 minutes on average, this could also be accelerated by exploring smaller grid sizes.

**********Recommendation**********

While the work presented in this paper is promising, it ultimately has major weaknesses in both task formulation and model development which make it not yet ready for publication. With additional work discussed in the suggested work, beyond the scope of the rebuttal period, this paper could be significantly improved and authors are encouraged to do so.

**References**

1. Gallo, Danilo, et al. "Investigating the Integration of Human-Like and Machine-Like Robot Behaviors in a Shared Elevator Scenario." *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 2023.
2. Mavrogiannis, Christoforos, et al. "Core challenges of social robot navigation: A survey." *ACM Transactions on Human-Robot Interaction* 12.3 (2023): 1-39.
3. Stoler, Benjamin, et al. "T2FPV: Dataset and Method for Correcting First-Person View Errors in Pedestrian Trajectory Prediction.”
4. Gu, Zaiwang, et al. "Ce-net: Context encoder network for 2d medical image segmentation." *IEEE transactions on medical imaging* 38.10 (2019): 2281-2292.