Spotting the Unfriendly Robot – Towards better Metrics for Interactions

Raphael Wenzel*, Malte Probst*

Abstract— **Establishing standardized metrics for Social Robot Navigation (SRN) algorithms for assessing the quality and social compliance of robot behavior around humans is essential for SRN research. Currently, commonly used evaluation metrics lack the ability to quantify how cooperative an agent behaves in interaction with humans. Concretely, in a simple frontal approach scenario, no metric specifically captures if both agents cooperate or if one agent stays on collision course and the other agent is forced to evade. To address this limitation, we propose two new metrics, a conflict intensity metric and the responsibility metric. Together, these metrics are capable of evaluating the quality of human-robot interactions by showing how much a given algorithm has contributed to reducing a conflict and which agent actually took responsibility of the resolution. This work aims to contribute to the development of a comprehensive and standardized evaluation methodology for SRN, ultimately enhancing the safety, efficiency, and social acceptance of robots in human-centric environments.**

I. PROBLEM STATEMENT

Research on Social Robot Navigation (SRN) is advancing rapidly, leading to the development of new algorithms that enable intelligent, foresighted navigation around humans, even in dense crowds [1]–[3]. However, in order to improve and compare SRN algorithms, clear and agreed upon evaluation protocols and metrics are required. To address this need, recent efforts have been made to establish common ground in the scientific community. Francis et al. [1] provide a broad survey on the current state of Social Robot Navigation, including evaluation. They surveyed the most prevalent simulators and datasets used for evaluation, along with commonly used metrics and a taxonomy for these metrics. Wang et al. [4] propose a condensed set of metrics which evaluate the comfort, naturalness, and sociability along with an evaluation protocol. Gao et al. [2] compile a large number of metrics for navigation performance, human discomfort and sociability.

All these studies report a lack of agreed-upon benchmarks and evaluation criteria. This holds especially true for the evaluation of one of the core aspects of Social Robot Navigation, the interaction between agents. In this sense, SRN can be seen as a sequence of resolving interactions between various agents until the robot reaches its goal. In each interaction, both parties have to resolve a potential conflict [5]. Improving the interactions between robots and humans contributes to various aspects of SRN, such as Safety, Social Compliance, or human discomfort. The ability to measure how cooperative an agent behaves around humans is crucial to the development of SRN algorithms, allowing for direct comparison and quantifiable improvement.

*Honda Research Institute Europe GmbH, Carl-Legien-Str. 30, 63073 Offenbach, Germany, Email:{firstname.lastname}@honda-ri.de

This is aggravated by the fact that humans are very adaptive. Humans will adapt quickly to an unresponsive or uncooperative robot by efficiently taking responsibility for the conflict resolution. Even unsophisticated navigation will, most likely, not lead to a significant number of collisions or similar commonly used success metrics for SRN. As a result, most metrics will not capture the difference in social compliance between good or bad robotic behavior. Specifically, when a human and a robot are evading each other, they do not capture the amount to which an agent contributes to the resolution of the conflict. However, socially inept navigation will create inefficiency and annoyance to the humans around the robot, especially if the penetration rate goes up.

We determine that most metrics do not capture complex aspects of behavior quality such as cooperativeness or social compliance. Furthermore, even if those metrics indicate that an observed interaction was resolved proficiently, most metrics are symmetrical by design, i.e., they do not indicate which agent was responsible for the resolution.

To spark discussion on this topic, we show in Section III using a simple experimental setup with highly adaptive agents, that the most commonly used metrics do not capture the social quality of an algorithm's behavior. In Section IV, we propose two novel metrics: *Conflict Intensity* and *Responsibility*. We assess their insights into the experiments and discuss potential for further research on advanced metrics to evaluate social compliance.

II. EXISTING SOCIAL COMPLIANCE METRICS

In order to evaluate commonly used metrics to assess the performance of SRN algorithms, we use a battery of well-established task-wise metrics. Using the taxonomy established by [1], we focus on Social, Hand-Crafted (Algorithmic) and Task-Wise metrics to evaluate quality of behavior and social compliance. Within this category, we distinguish between kinematic, distance-based, and prediction-based.

Kinematic metrics assess the task performance using the robot's motion data. Commonly used are measures which represent the minimum, average or maximum of the robot's velocity v_{min}, v_{avg}, v_{max}, acceleration a_{min}, a_{avg}, a_{max} and jerk j_{min}, j_{avg}, j_{max} [1]. These measures serve as surrogates for social compliance, based on the assumption that higher social compliance enables the robot to maintain higher speeds or smoother motion profiles.

Metrics based on the robot's distance to pedestrians in each time step capture that close proximity to the robot is inherently dangerous. Francis et al. [1] collected several quality

and social metrics that fall into this category. Measures based on average or maximum *Clearing Distance*, CD_{avg}, CD_{max} directly represent the minimum distance between the robot and an object it encounters. Note that the average/maximum refers to the statistics over multiple encounters with several objects. Other measures, like *Space Compliance/Violation Rate* SVR, evaluate the duration (or rate) for which two agents enter each other's designated space. In this context, we consider the space in question to be the Personal Space, according to the Proxemics model [6]. Truong and Ngo [7] proposed the *Collision Index* CI (also referred to as *Social Individual Index* SSI), which is a distance-based metric parameterized by a standard deviation of the pedestrian's personal space.

Some metrics utilize the measured velocities of the robot and other agents, applying constant-velocity predictions to capture the criticality of observed behavior. The *Minimum Time-To-Collision* TTC_{min} [1], [8] captures the criticality of a collision if both agents continue to move with their current velocity vector. The *Projected Path Duration* PPD [4], [9] evaluates the duration, for which the social safety zones of two agents overlap. Each safety zone is represented by a rectangular area in front of the agent defined by the width of the agent and a velocity-proportional length of the safety zone.

All these metrics aim to capture various aspects of safety, perceived or factual, of behavior computed by SRN algorithms. However, they do not directly capture the key dimensions of social compliance and cooperativity of robot motion planning as the following simulative experiments show. To the best of our knowledge, no existing metrics capture the responsibility assumed by an agent during an interaction.

III. Experiments

In a simple experiment setup, we place two agents 20 meters apart, with their goals set to the other agent's starting position. The objective is to pass each other to reach their goals. Both agents can be either "compliant", i.e., engaging in the interaction by avoiding a collision and accepting a longer path option to reach the goal or "not compliant", by ignoring the other agent and "blindly" showing a constant-velocity behavior. We consider all four permutations of these behavior modes. In the first scenario, both agents follow a direct collision course, passing through each other and reaching their goal on the direct path. While unrealistic, this simulative scenario exemplifies a worst case of bad behavior planning of both agents with the corresponding outcome. In the second scenario, the ego agent (the robot) is non-compliant and proceeds to "barrel through" the other agent to reach its goal. This can be considered bad behavior planning on the side of the robot, as it shows no social compliance whatsoever. In the third scenario, the roles are reversed and the robot is compliant. In Scenarios 2 and 3, no collision occurs due to the high adaptivity and social compliance of one of the agents. In Scenario 4, both agents are compliant, resulting in an efficient resolution of the scenario.

In this experiment, the compliant agents demonstrate foresighted behavior by avoiding each other at an early stage. To achieve this, we parameterize the Social Forces algorithm [10] with highly cooperative parameters, specifically: A = 5.1, $\lambda = 3.0$, γ = 0.35, n = 1 and n' = 3.0). The results of these simulations, evaluated with the commonly used social evaluation metrics described in Section I, are shown in Table III.

A. Finding 1: Evaluating Social Compliance

The experiment reveals that none of the metrics reliably distinguish between Scenario 2 (where the robot behaves uncooperatively) and Scenario 4 (where both agents behave cooperatively). The kinematic metrics (1-5) fail to capture the criticality of the robot's behavior because the other agent completely assumes the responsibility of resolving the conflict. Distance-based measures (6-8), such Space Violation Rate SVR and Collision Index CI, exhibit a slight decrease when both agents are cooperative. Consistently, the Clearing Distance CD_{avg} shows a small increase. However, based on their respective criticality thresholds, none of these metrics would have flagged the robot's behavior as critical. Prediction-based metrics (9-10), such as TTC_{min} and PPD, exhibit the most significant differences, although for incorrect reasons. Although TTC_{min} indicates a difference in criticality, both values are substantially higher than what is normally considered critical. Moreover, the TTC_{min} is only applicable when a collision is predicted, which is too restrictive for movement in 2D space. The PPD does not detect any criticality when both agents are compliant. In conclusion, none of the commonly employed metrics adequately capture the robot's ability (or inability) to resolve interactions.

B. Finding 2: Evaluating Interaction Responsibility

A comparison of the results from Scenario 2 (where the robot behaves uncooperatively) and Scenario 3 (where the human behaves uncooperatively) reveals that none of the evaluated metrics provide a clear indication of which interaction partner took responsibility. The differences in kinematic metrics (1-5) are misleading, showing the uncompliant robot (S2) as moving more smoothly, and will suffer in real-world scenarios that are less sterile. All other metrics (6-10) are, in fact, identical for Scenario 2 and 3. To address this limitation, we introduce two novel metrics in the next section.

IV. Proposed Metrics

A. Conflict Potential

We begin by assessing the conflict potential in a scenario involving two agents: the robot (referred to as "*ego*") and another agent (referred to as "*other*"). Later, we will extend this theory to evaluate entire interactions and explore its application to multiple agents.

Based on [11], which introduces the concept of Distance at Closest Encounter (DCE), we calculate the predicted Distance at Closest Encounter pDCE as a starting point.

TABLE I

COMMONLY USED METRICS FOR EVALUATING SRN PERFORMANCE

Scenario	Resulting paths	(1.) Avg. vel	(2.) Avg. accel	(3.) Max. accel	(4.) Avg. jerk	(5.) Max. jerk	(6.) Avg. Clearing Distance	(7.) Space Violation Rate	(8.) Collision Index	(9.) Min. TTC	(10.) Proj. Path Duration	**Conflict Intensity**	**Responsibility (R/H)**
S1: Nobody compliant		1.0	0.0	0.0	0.0	0.0	0.0	0.12	1.0	0.0	3.0	10.26	(0, 0)
S2: Robot not compliant		1.0	0.0	0.0	0.0	0.0	1.18	0.11	0.03	3.63	3.0	4.91	(0, 1)
S3: Human not compliant		0.98	0.0	0.14	0.0	0.88	1.18	0.11	0.03	3.63	3.0	4.91	(1, 0)
S4: Both compliant		0.99	0.0	0.05	0.0	0.89	1.54	0.10	0.0	4.79	0.0	3.89	(0.5, 0.5)

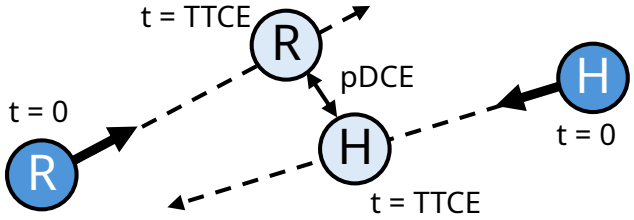

Fig. 1. Construction of the predicted Distance at Closest Encounter (pDCE). Based on their relative position and velocity, the time to closest encounter (TTCE) can be computed. The distance at that time is the pDCE.

The pDCE assesses the minimum distance between the two agents when their movements are predicted based on a constant velocity assumption, as can be seen in Figure 1. This metric indicates the proximity to a potential collision if both agents continue along their current trajectories. Geometrically, this equates to calculating the perpendicular distance, calculated using the following equation:

$$\text{pDCE} = \frac{|\mathbf{r} \times \mathbf{v}|}{|\mathbf{v}|} \tag{1}$$

Here, $\mathbf{r}$ is the relative position vector and $\mathbf{v}$ the relative velocity vector of both agents. Based on the pDCE, the *Conflict Potential* CP of a situation is given as:

$$\text{CP}_{=} \max(0, 1 - \frac{\text{pDCE}}{s_{\text{ego}} + s_{\text{other}}}) \tag{2}$$

Here, s_{ego} and s_{other} represent the radii of both agents. The conflict potential CP indicates the extent of overlap between the two agents at the point of closest encounter. As a result, the conflict potential C is at its max (i.e., equal to 1) in the event of a head-on collision and minimal (i.e., equal to 0) in the case of a near miss. Any motion that alters the agents' trajectory away from a direct collision course will decrease the conflict potential. Therefore, our first proposed metric, the conflict *Intensity* I, is defined as:

$$I = \int \text{CP}(t)\, dt \tag{3}$$

B. Responsibility

We define the *Responsibility* R in an interaction as the extent to which an agent reduces the conflict potential. To determine an agent's responsibility in conflict resolution, we examine how each agent's behavior $\mathcal{B}$ contributes to reducing the conflict potential CP. To approximate this, we calculate the Conflict Contribution CC, caused by the change in behavior $d\mathcal{B}$ of an agent in the last time-step, given by

$$\text{CC}_{\text{agent}} = \frac{d}{d\mathcal{B}}\text{CP} \approx \text{CP} - \text{CP}_{\text{no change, agent}} \quad , \tag{4}$$

where $\text{CP}_{\text{no change, agent}}$ is calculated based on the agent's velocity vector v^{t-1} from the previous time step. By integrating the Conflict Contribution CC over time for each agent, we obtain a measure of the agent's contribution to conflict resolution, as expressed by:

$$R = \frac{1}{\text{CP}_0} \int \text{CC}(t)\, dt \tag{5}$$

Here, CP_0 is the conflict potential at the start of the interaction which must be reduced to resolve the conflict. Both the conflict Intensity I and the agent's Responsibility R are task-wise metrics, providing a single scalar value that characterizes the entire observed interaction. In contrast, the intermediate quantities CP (conflict potential) and CC (conflict contribution) are step-wise metrics, whose evolution over time provides insights into the progression of the interaction between the two agents.

V. RESULTS FOR PROPOSED METRICS

For Scenario 1, a head-on collision scenario, the *Conflict Intensity* for both agents is $I = 10.255$. If either of the agents takes steps towards resolving the interaction, the conflict intensity decreases to $I = 4.910$, and decreases to $\text{I} = 3.889$ if both agents cooperate. The relativity between these values appears to be consistent, reflecting the fact that the conflict is present in all cases, but the resolutions differ. We argue that this is beneficial for measuring compliance in social robot navigation, as it reflects the persistence of underlying conflicts despite mitigation by one of the agents. The reduction of Intensity from Scenarios 2/3 to 4 is relatively small. This is plausible, as one agent would have resolved the conflict

independently. However, when both agents share the burden, the intensity is reduced further, albeit only slightly.

In contrast, other metrics seem to be less consistent across scenarios. CD_{avg} and TTC_{min} show a drastic reduction in criticality between Scenario 1 vs. Scenario 2/3. Metrics like PPD and CI decrease sharply when both agents resolve the conflict, despite all scenarios beginning with the same initial conflict. The SVR shows only a very small relative change across all 4 Scenarios. This makes it harder to compare different scenarios or algorithms.

The Responsibility metric shows a clear difference between the four scenarios: In the first scenario, neither agent takes responsibility for avoiding the collision, resulting in a collision. The Responsibility for both agents evaluates to $R_R = R_H = 0.0$. In Scenarios 2 and 3, the agent that takes responsibility earns the full share of the Responsibility metric. Specifically, $R_H = 1.0$ (in Scenario 3) and $R_R = 1.0$ (in Scenario 2), respectively. In the fourth scenario, where both agents share the responsibility equally, the Responsibility metric clearly indicates the cooperation: $R_R = R_H = 0.5$. The equal 50% share of responsibility between the two agents is a logical consequence of the symmetrical initial conditions and the identical behavior planners used for both agents.

The proposed step-wise quantities Conflict Potential (CP) and Conflict Contribution (CC) also provide insights into the conflict resolution in these exemplary scenarios. Figure 2 provides the progression of CP and CC values in all 4 Scenarios. The top right plot shows that, while all scenarios begin with the same Conflict Potential, the rate of conflict resolution varies significantly. In Scenarios 2 and 3, the conflict potential decreases at the same rate. However, in Scenario 4, both agents cooperate to resolve the interaction, resulting in a steeper decline in Conflict Potential and an earlier resolution. Since the Intensity of the conflict is equivalent to the area under the Conflict Potential curves, this relationship is also reflected in the Intensity metric. Similarly, the Conflict Contributions (lower left: "ego, lower right: "other") is smaller in Scenario 4, where both agents cooperate. Consequently, the responsibility assigned to both agents is smaller, as it is normalized by the initial overall intensity of the interaction.

VI. Discussion

To effectively use the proposed metrics as social quality metrics for SRN, they must be applied not only to pairwise interactions but also consider the presence of all other pedestrians in the scene. To achieve this, average values of Intensity and Responsibility can be used to assess how a given algorithm's behavior affects conflict intensity, either by reducing or increasing it. Similarly, the average Responsibility can serve as an indicator of the robot's contribution to conflict resolution.

Importantly, both metrics can be parameterized with the agent radii to focus on social compliance regarding collision or invasion of personal space. This enables a more nuanced assessment of agent behavior, based on various environmental conditions, such as crowd density.

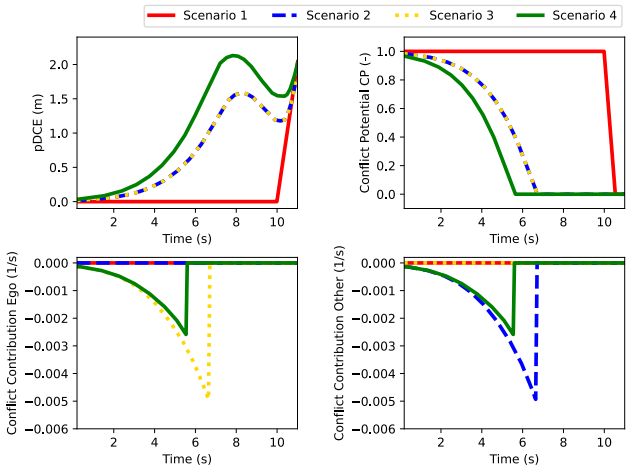

Fig. 2. Progression in all 4 Scenarios of various step-wise metrics used in this work to derive the Intensity and Responsibility metrics.

A key insight from the construction of collision intensity is that it is an integral over the collision potential. This indicates that early conflict resolution significantly reduces the intensity of a conflict compared to a late resolution. This aligns with intuition, where a foresighted action reduces the overall criticality compared to a late reaction, even though the final outcome is the same. Moreover, the responsibility metric would also attribute a significant share to an agent executing an early action.

The proposed metrics are designed to be generalizable. The use of pDCE as a base for constructing the metrics makes the metrics meaningful for all motion in 2D space. Additionally, the metrics can capture various types of behavior $\mathcal{B}$. In an additional experiment, we observed that two Social Force agents with the same parameters crossing at a 90° angle showed different reactions to mitigate the collision. One agent slowed down, resolving the interaction via a change in speed. The other agent evaded by veering away, resolving the interaction by changing direction. In this case, the responsibility according to Eq. 5 showed an almost equal share of responsibility.

VII. Conclusion

This paper addresses the need for advanced evaluation metrics that capture social compliance and cooperativity in Social Robot Navigation. Humans are highly adaptive around robots, which is one of the main reasons why commonly used metrics are insufficient. Through simulative experiments, we demonstrate that typical metrics struggle to distinguish between simplistic and socially compliant navigation algorithms when interacting with cooperative partners. We identify a need for additional metrics that capture the reduction of conflict intensity and the allocation of responsibility between agents. We propose two metrics, the *Conflict Intensity* and the *Responsibility* for an agent's observed behavior. We show that these metrics effectively capture the observed effects in our simulative experiments. Our goal is to initiate a discussion on the development of suitable metrics for evaluating the performance of Social Robot Navigation algorithms in human-robot interactions.

REFERENCES

[1] A. Francis, C. Pérez-d'Arpino, C. Li, F. Xia, A. Alahi, R. Alami, A. Bera, A. Biswas, J. Biswas, R. Chandra *et al.*, "Principles and guidelines for evaluating social robot navigation algorithms," *ACM Transactions on Human-Robot Interaction*, vol. 14, no. 2, pp. 1–65, 2025.

[2] Y. Gao and C.-M. Huang, "Evaluation of socially-aware robot navigation," *Frontiers in Robotics and AI*, vol. 8, p. 721317, 2022.

[3] P. T. Singamaneni, P. Bachiller-Burgos, L. J. Manso, A. Garrell, A. Sanfeliu, A. Spalanzani, and R. Alami, "A survey on socially aware robot navigation: Taxonomy and future challenges," *The International Journal of Robotics Research*, vol. 43, no. 10, pp. 1533–1572, 2024.

[4] J. Wang, W. P. Chan, P. Carreno-Medrano, A. Cosgun, and E. Croft, "Metrics for evaluating social conformity of crowd navigation algorithms," in *2022 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*. IEEE, 2022, pp. 1–6.

[5] R. Mirsky, X. Xiao, J. Hart, and P. Stone, "Conflict avoidance in social navigation—a survey," *ACM Transactions on Human-Robot Interaction*, vol. 13, no. 1, pp. 1–36, 2024.

[6] E. T. Hall, *The hidden dimension*. Anchor, 1966, vol. 609.

[7] X.-T. Truong and T.-D. Ngo, "Dynamic social zone based mobile robot navigation for human comfortable safety in social environments," *International Journal of Social Robotics*, vol. 8, pp. 663–684, 2016.

[8] D. N. Lee and R. Lishman, "Visual control of locomotion," *Scandinavian journal of psychology*, vol. 18, no. 1, pp. 224–230, 1977.

[9] J. Jin, N. M. Nguyen, N. Sakib, D. Graves, H. Yao, and M. Jagersand, "Mapless navigation among dynamics with social-safety-awareness: a reinforcement learning approach from 2d laser scans," in *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2020, pp. 6979–6985.

[10] M. Moussaïd, D. Helbing, S. Garnier, A. Johansson, M. Combe, and G. Theraulaz, "Experimental study of the behavioural mechanisms underlying self-organization in human crowds," *Proceedings of the Royal Society B: Biological Sciences*, vol. 276, no. 1668, pp. 2755–2762, 2009.

[11] J. Eggert, "Predictive risk estimation for intelligent adas functions," in *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2014, pp. 711–718.