machines

MDPI

Article

Sociability Modelling in Robot Motion for Generating Socially Predictable Trajectories

Haiwei Luo [1], Jin Zhang [2], Yaqing Luo [3] and Gaoyong Luo [2,*]

[1] Department of Computing, Imperial College London, London SW7 2AZ, UK; haiwei.luo25@imperial.ac.uk
[2] Department of Intelligent Systems, Institute of Machine Learning, London SW9 9DX, UK; jin.zhang@ai-iml.org
[3] National Heart and Lung Institute, Imperial College London, London SW7 2BX, UK; yaqing.luo22@imperial.ac.uk
* Correspondence: gaoyong.luo@yahoo.com

Abstract

Modelling and quantifying human socialness onto robots remain a challenge, due to the complex mechanisms and reasoning processes that incorporate human intelligence to enable social behaviours. In this paper, we propose a novel approach of modelling human sociability in the social context of human–robot interactions by deriving the sociability score, which integrates both legible and trustable motion. To generate socially accepted motions, with potential deployment in real-time and dynamic environments, a new procedure is developed to encode human trustability onto robot motion, with the introduction of the trustability score, which explores perceived benevolence and the importance of initial trust. By applying the trust region of predictability, trustably and socially predictable trajectories are thus generated that can be identified and interpreted by humans consistently as verified by experiments. The experimental results also demonstrated that the scores computed by the proposed method can effectively capture their respective defined characteristics. Furthermore, to generate and evaluate predictable trajectories independent of other trajectories, a modified predictability score computation has been derived. Finally, as a step towards creating social intelligence, we train a deep learning-based classifier to identify socially predictable trajectories, mimicking humans' ability to recognise such motion.

Keywords: human–robot interaction; social robotics; social intelligence; human-centred AI; predictability; trustability; legibility; robot motion in real-time and dynamic environments; socially aware motion

Check for updates

Academic Editor: Sergey Y. Yurish

Received: 31 May 2026
Revised: 20 June 2026
Accepted: 23 June 2026
Published: 25 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

From harvesting crops to assembling automobiles and delivering medications, modern robotics is the driving force behind these diverse tasks, and the deployment of robotic systems can now enhance productivity, safety, and flexibility in various industries such as manufacturing, agriculture, and healthcare [1]. With technology advancements, robots are no longer confined to industrial applications, and have now become part of the habitats and workspaces of humans, such as self-driving cars. Robots can now be used to assist industrial workers in factories as co-workers (cobots), support care attendants in nursing homes or hospitals, help customers in a supermarket, or do the housework at home. Particularly service, medical, or social robots [2], which are machines designed by humans to perform tasks automatically or under the control of a computer program, and equipped with sensors, actuators, and a processing unit, and carrying out specific functions, have integrated

mechanical and electronic engineering for crafting tangible components and computational systems for information processing and decision-making, allowing them to interact with humans [3]. In some cases, they can operate by responding to or learning from human movements through supervising or training their behaviours during deployment [4]. As such, determining a positively perceived interaction for socially accepted robot behaviour in human environments is very important, where legibility means that an interacting human is able to understand the robots' intentions [5].

Other crucial aspects of robot behaviour include predictability and trustability. The predictability of a robot's actions is a key factor in successful human–robot interactions (HRIs) [6], where humans can anticipate what the robot will do. The trustability of robot interactions, often referred to as human–robot trust, has emerged as a new research area that is gaining attention. However, a more inclusive definition of trust, encompassing all aspects of human–robot trust, is required to be explored [7].

On the other hand, the social context of HRIs, which is defined in diverse ways, is critical for the design, evaluation, and automatic generation of appropriate robot behaviour [8]. This variability makes it challenging to draw connections between related work on understanding and modelling human sociability in robot motion.

Particularly in real-time and dynamic environments, generating socially accepted motions and aligning robot behaviour with social expectations remain a challenge, due to the complex mechanisms and reasoning processes that require the incorporation with human intelligence. For example, deploying multimodal artificial intelligence (AI) to learn from large datasets, identify patterns, and apply logical inference, could allow robots to engage in natural social interactions, ensuring social behaviour to be appropriate and safe. Therefore, this study aims to address the challenge by investigating HRIs focusing on a social context, where a robot can work in close proximity to humans. For the robot to blend naturally into the social environment, its motion must not only achieve the task but also adhere to social norms in a way that is both easily understood (legible) and reassuring to those around it (trustable). Our research [9] has shown that these two properties are the essential pillars of robot sociability. Both legibility and trustability hinge on the robot's way of expressing its intent through motion. Together, they define sociability as the ability of a robot to communicate its intentions in a human-like way that is, in turn, socially appropriate. The legibility component allows humans to clearly interpret the robot's intended actions, while the trustability component is required for humans to feel comfortable and safe during interactions, based on their interpretation of robot intent. In this sense, sociability arises not only from what the robot intends to do, but how it communicates that intent to humans. As a result, it is important to ensure the intent is conveyed in a way that can be anticipated and interpreted by humans, to create motion that is socially predictable. By making intent interpretable, clear, and comforting, robots can foster human confidence and acceptance, enabling genuinely sociable behaviour in HRI.

Furthermore, social intelligence in robots arises from the ability to model, recognise, and reproduce aspects of human social behaviour [10]. In HRI, social behaviour is not only expressed through verbal communication, but also through movement, spatial awareness, intent signalling, and the ability to adapt to others [11]. For robots to behave socially, they must therefore be capable of identifying these behaviours and generating actions that are consistent with human expectations. This makes the recognition of socially predictable motion an important step towards creating social intelligence in robotic systems. For this purpose, we propose not only a method to model and quantify human socialness onto robots for generating socially predictable trajectories by addressing the issues identified and discussed, but also a model-free algorithm for identifying this type of motion.

2. Related Work

2.1. Notions of Predictability, Legibility, Trustability, and Sociability

2.1.1. Predictability

Predictability in robot motion has been formally defined [12] as the extent to which the robot's motion matches with the expectation of the human observer, for a specified goal. If we consider actions to be goal-oriented, where many studies in psychology suggest that humans naturally perceive actions in this manner [13], then we can explain the inference occurring in predictability as "goal-to-action".

This has been defined [12] as the inference mapping goals $\mathcal{G}$ to trajectories Ξ, where $\mathcal{I}_{\mathcal{P}}$ models the observer's expectation:

$$\mathcal{I}_{\mathcal{P}} : \mathcal{G} \rightarrow \Xi \tag{1}$$

Furthermore, predictable motion is defined [12] as one where the trajectory from an initial point S to a goal point G, denoted as $\xi_{S \rightarrow G}$, matches the following inference:

$$\mathcal{I}_{\mathcal{P}}(G) = \xi_{S \rightarrow G} \tag{2}$$

The optimality of the trajectory inference is evaluated by the predictability score, affected by the trajectory's cost, which penalises the trajectory's length, as it is established that humans expect a direct trajectory to the goal. Shorter trajectories are considered to be more "efficient", thus the predictability score can be interpreted as a measure of the "efficiency" of a trajectory. Formally, the predictability score of a trajectory ξ is defined [12] as a normalised value between 0 and 1:

$$\text{predictability}(\xi) = \exp(-C(\xi)) \tag{3}$$

where C denotes the cost function, modelling the observer's expectation, and can be defined in terms of the trajectory length.

2.1.2. Legibility

Legibility in robot motion is also formally defined as the degree to which a human observer is able to infer the robot's intent from its motion [12]. In the context of goal-oriented actions [13], legible motion can be understood as "action-to-goal", where the robot expresses its intended goal through its motion.

This has been defined [12] as the inference mapping the snippets of trajectories, from all trajectories to the goal:

$$\mathcal{I}_{\mathcal{L}} : \Xi \rightarrow \mathcal{G} \tag{4}$$

Moreover, legible motion is [12] defined as one which allows for quick and confident inference of the correct goal G by the observer, after observing a snippet of the trajectory $\xi_{S \rightarrow Q}$, from an initial point S to the configuration at a time t, where $Q = \xi(t)$:

$$\mathcal{I}_{\mathcal{L}}\left(\xi_{S \rightarrow Q}\right) = G \tag{5}$$

The optimality of the goal inference across an entire trajectory is evaluated using the legibility score, tracking the assigned probability of the actual goal G^* across the trajectory denoted by $P\left(G^* \mid \xi_{S \rightarrow \xi(t)}\right)$. This is weighted using $f(t)$ which assigns more weight to the earlier segments of the trajectory, by employing for example $f(t) = T - t$, where T is the duration of the entire trajectory. This ensures the weight decreases over time; as t increases, $f(t)$ decreases, assigning lower weights to later segments of the trajectory. The score is

normalised by dividing $\int f(t)dt$ for each weighted assigned probability of G^* across the trajectory. Formally the legibility score of a trajectory ξ is given by [12]:

$$\text{legibility}(\xi) = \frac{\int P\Big(G^* \mid \xi_{S\to\xi(t)}\Big) f(t)\, dt}{\int f(t)\, dt} \tag{6}$$

To compute $P\Big(G^* \mid \xi_{S\to\xi(t)}\Big)$, C can be used as shown below [12], which calculates the inference of a goal G, given a trajectory snippet $\xi_{S\to Q}$:

$$P(G \mid \xi_{S\to Q}) \propto \frac{\exp\Big(-C(\xi_{S\to Q}) - C(\xi^*_{Q\to G})\Big)}{\exp(-C(\xi^*_{S\to G}))} P(G) \tag{7}$$

where ξ^* denotes an optimal trajectory.

2.1.3. Trustability

Unlike predictability and legibility, the notion of trustability lacks a clear definition in the context of robot motion. Insights from psychology and ideas of trust already established in robotics provide us a foundation for its formalisation.

Psychological models of trust [14,15] identify three key factors in shaping belief in another agent from the perspective of an observer: benevolence, competence, and integrity. Benevolence, the extent to which the observer perceives the agent as caring and harmless, is the most critical factor when humans evaluate unfamiliar agents such as robots.

In robotics, trust has historically been framed in terms of the avoidance of harm or hostility towards humans [16], echoing Asimov's laws [17]. Benevolence thus dominates, as without clear signals that a robot lacks harmful intent, no initial trust can form, particularly given humans' tendency to fear the unknown [18]. Trust has also been linked to the willingness to be vulnerable [19,20], which is encouraged when robots exhibit benevolent or socially aware behaviours, such as acknowledging human presence through movement cues [16,20]. Experiments further show that humans trust robots more when they display submissive moment strategies that account for human actions, rather than dominant ones [20].

Moreover, studies have shown that initial trust in the agent has a lasting effect, whereby early impressions strongly influence long-term trust, even after subsequent violations [15,21]. Thus, agents must signal the mentioned factors that influence trust early in the interaction with the observer.

2.1.4. Sociability

Sociability in robot motion is commonly defined in the context of socially aware navigation. For example, it has been suggested that sociability is achieved when humans become comfortable [22] or when robots adhere to cultural conventions [23]. A key way of enforcing these social conventions in robot motion is through Hall's theory of proxemics [24], which concerns the use and preservation of interpersonal space during social interaction. In human–robot interaction (HRI), this is commonly represented as a personal or safety zone around the human that the robot should avoid violating unless the task explicitly requires close interaction. This ensures that the robot does not simply avoid physical collision, but also respects the human's psychological comfort and perceived safety.

However, it should be stressed that there still lacks a formal and generic definition for sociability in robot motion, which could encompass robot arm movements.

2.2. Predictable, Legible, and Sociable Motion Planning

The notions of predictability and legibility have been discussed in the field of robotics motion planning. For instance, Ngo et al. [25] proposed a novel potential field method that ensured the legibility property of motion, compared with traditional field methods that only guaranteed predictability. This approach is beneficial in human-centred environments, since intent-expressive motion could be generated.

Furthermore, Bodden et al. [26] proposed a new intent-expressive objective function, that splits motion into two separate components—one for reaching the goal and the other that allows for legibility. This proposal allows for the objective function to simultaneously optimise for direct motion (predictable) and legible motion.

A machine learning-based approach in generating legible motion which is less dependent on pre-set tasks has been proposed by Busch et al. [5], where a legible policy is learnt through reinforcement learning.

Nikolaidis et al. [27] expanded on the idea of legibility [12] by incorporating observer perspective. They first introduced a viewpoint model, projecting trajectory and goals onto the observer's view plane, to generate viewpoint-based legible motion. Secondly, they introduced an occlusion model, capturing the change in legibility when parts of the trajectory are hidden from the observer. By taking into account occlusion in different segments of the trajectory, occlusion-based legible motion can then be generated.

2.2.1. Gradient Descent Methods

In Dragan's thesis [28], predictable and legible motion have been generated by utilising the gradient descent update rule for trajectory ξ:

$$\xi_{i+1} = \xi_i - \frac{1}{\eta} A^{-1} \nabla_{\xi_i} \mathcal{U} \tag{8}$$

where $\nabla_{\xi_i} \mathcal{U}$ is the Euclidean gradient and η is a coefficient to adjust the step-size. A is a matrix that has off-diagonal non-zero elements, which establishes a relationship between the current time point, and the previous and next ones, by satisfying:

$$\langle \xi_1, \xi_2 \rangle_A = \xi_1^\top A \xi_2 \tag{9}$$

which adapts from the Euclidean inner product as defined as:

$$\langle \xi_1, \xi_2 \rangle_I = \int \xi_1(t)^\top \xi_2(t)\, dt \tag{10}$$

The cost function C is set as the integral of the squared velocities within the configuration space. The predictability gradient has been derived as the negative acceleration (without considering obstacles), and thus the most predictable motion is attained by minimising C, which occurs when the negative acceleration is 0.

Moreover, the legibility gradient has been derived as the derivative of normalised weighted probability inference of the actual goal across the trajectory, with respect to trajectory ξ (without accounting for obstacles):

$$\nabla \text{Legibility}(t) = K \frac{\partial P}{\partial \xi}(\xi(t), t) \tag{11}$$

where

$$P(\xi(t), t) = P\Big(G^* \Big| \xi_{S \to \xi(t)}\Big) f(t),\ K = \frac{1}{\int f(t) dt}$$

2.2.2. Methods for Improving Predictability

To generalise to new situations, where the robot can adapt to new initial and goal positions, we can employ the learning from demonstration method to further improve predictability. To optimise over a Hilbert space of trajectories, model-free approaches [28] can be used through generalising the adaptation process, such that trajectory adaptation can be formalised as a Hilbert norm minimisation problem.

A method of encouraging humans to adapt to the robot, known as familiarisation, which exposes humans to repeated robot motions, can be employed to increase predictability over time. However, inherently unnatural motions would never achieve high predictability, regardless of exposure.

2.2.3. Methods for Improving Legibility

It has been observed that sometimes when a trajectory is optimised, it can become arbitrarily unpredictable [28], because the definition of goal inference [12] is unable to capture the way humans make inferences in unpredictable situations. Dragan's solution was to introduce a trust region of predictability [28], which constrains legibility optimisation to only when predictability is high enough. This constraint is defined with respect to the cost function C, that bounds the domain of trajectories. Thus, the legibility model is only trustable if it lies inside the trust region:

$$C[\xi] \leq \beta \tag{12}$$

In the context of Yu et al.'s trust model [15], the utilisation of the trust region ensures display of competence and integrity for a robot, as the purpose of the trust region is to limit a human's confusion regarding the robot's trajectory.

2.2.4. Socially Aware Navigation

Eirale et al. [29] proposed using an additional cost map to encode the social constraints in navigation, that incorporates information such as the positions of people and goals. By using a DNN (deep neural network), a social cost map is trained to recognise and encode social obstacles such as a group of people or a queue. When integrated with a discrete path planner alongside the local and global cost maps, a socially acceptable path is generated, where for example, the plan allows the navigation of the robot around a group of people and towards the back of a queue, ensuring adherence to social norms.

On the other hand, Singamaneni et al. [30] reviewed socially aware robot navigation, highlighting the importance of human comfort, trust, proxemics, intention communication, and context-dependent evaluation.

Although both of these literatures mainly concern robot navigation, these principles are also relevant to manipulators operating near humans. A research gap remains in quantitatively evaluating such social behaviours and in generating manipulator trajectories that exhibit them.

2.2.5. Robust Control Methods

Cosenza et al. [31] showed through multibody simulation that robot dynamics and wheel slip can cause substantial deviations from kinematically planned motion. Moreover, Yang [32] proposed state-filtered disturbance rejection control for compensating matched and mismatched disturbances, whereas Yang [33] developed an actor–critic neuroadaptive controller for tracking uncertain nonlinear systems. These methods improve dynamic modelling, robustness, and tracking accuracy after a reference trajectory has been defined. In contrast, our present work wishes to focus on selecting and evaluating socially acceptable

trajectories. This approach is complementary to robust control methods, which could subsequently be employed to execute the generated trajectories accurately.

3. Sociability Modelling and Quantification

For robots to create social intelligence through modelling sociability to generate socially predictable trajectories, we begin by establishing that a plan (trajectory) generated by a planner consists of many waypoints, or local goals that the robot arm must traverse via, from the initial point in order to reach the goal point. We discretise time by treating each waypoint as a discrete time unit, where the index in the list of waypoints directly determine the time. This is formalised as:

$$Q_n = \xi[n] \tag{13}$$

where Q_n describes the nth waypoint on trajectory ξ.

In particular Q_0 describes the initial point and Q_{N-1} describes the goal point.

We then compute the scores based on the assumption that the path between each adjacent pair of trajectories to be a straight line. We also use the trajectory length as the cost function C. This means that the following holds for a (partial) trajectory between two waypoints:

$$C\left(\xi_{Q_1 \to Q_2}\right) = C\left(\xi^*_{Q_1 \to Q_2}\right) \tag{14}$$

where $\xi^*_{Q_1 \to Q_2}$ denotes the optimal trajectory from waypoint Q_1 to waypoint Q_2. More generally:

$$C\left(\xi_{Q_n \to Q_{n+1}}\right) = C\left(\xi^*_{Q_n \to Q_{n+1}}\right), \text{ for } n \geq 0,\ n \in Z \tag{15}$$

3.1. Predictability

To compute the predictability score of a trajectory, we can follow Equation (3). However, it is noted that the generated trajectory is not a continuous line, but rather waypoints that form a series a line segments or partial trajectories; we therefore propose a discretised predictability score as:

$$\text{predictability}_d(\xi) = \exp\left(-\sum_{n=0}^{N-2} C\left(\xi_{Q_n \to Q_{n+1}}\right)\right) \tag{16}$$

where N is the number of waypoints, inclusive of the initial and goal points.

3.2. Predictability Prime (Predictability′)

The predictability score is proficient at evaluating multiple trajectories with the same initial and goal points, as the trajectory with the highest predictability score would be the one with the lowest cost (length). However, this score is not very useful in evaluating trajectories with different initial and goal points, since some trajectories could be naturally longer due to the distance between the initial and goal points, yielding a lower predictability score, even if they still remain predictable. Hence, we propose a modified predictability score known as the predictability prime (predictability′) score that takes into account the shortest possible trajectory, given an initial point p_{init} and goal point p_{goal}. This is the straight-line distance between the two points. Formally, the predictability prime score measures the amount of deviation between the actual trajectory and the shortest possible trajectory, and is defined as follows:

$$\text{predictability}'(\xi) = \exp(-(C(\xi) - C_{\min}(\xi))) \tag{17}$$

where $C(\xi)$ is the trajectory distance, and $C_{\min}(\xi)$ is the shortest possible trajectory distance, given by the Euclidean distance between p_{init} and p_{goal} in 3D:

$$C_{\min}(\xi) = \left\| p_{\text{goal}} - p_{\text{init}} \right\|_2 = \sqrt{\left(x_{\text{goal}} - x_{\text{init}}\right)^2 + \left(y_{\text{goal}} - y_{\text{init}}\right)^2 + \left(z_{\text{goal}} - z_{\text{init}}\right)^2} \quad (18)$$

Note that $C_{\min}(\xi)$ evaluates to the same value as $C(\xi^*)$, since the optimal trajectory is the straight-line trajectory between two points.

Furthermore, Equations (16) and (17) can be combined to compute the discretised predictability prime score:

$$\text{predictability}'_d(\xi) = \exp\left(-\left(\left(\sum_{n=0}^{N-2} C\left(\xi_{Q_n \to Q_{n+1}}\right)\right) - C_{\min}\left(\xi_{Q_0 \to Q_{N-1}}\right)\right)\right) \quad (19)$$

with the computation procedure illustrated in Figure 1.

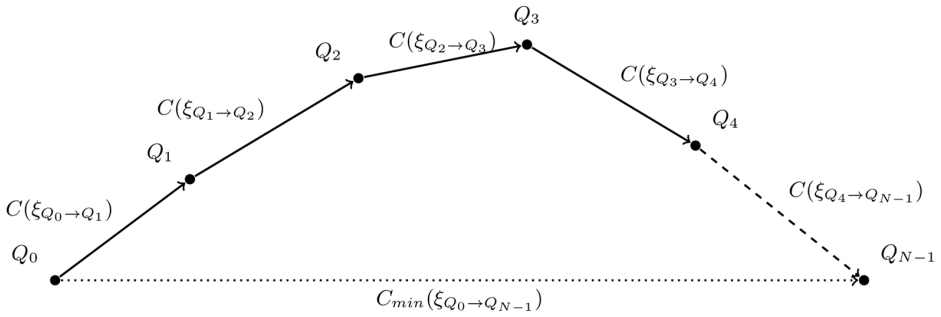


Figure 1. The computation of the discretised predictability and predictability prime scores.

3.3. Legibility

To compute the legibility, we first follow Equation (6) to calculate the legibility score, as well as Equation (7) to derive the probability of the goal inference of a trajectory snippet. In this study, we assume the number of possible goals to be one, as we are only evaluating the legibility of a trajectory after a plan towards the goal that has been generated. This means that it is certain that the trajectory would go towards the intended goal, resulting in $G^* = G$ and $P(G) = 1$ in Equations (6) and (7). For the weighting function $f(t)$, we refer to Dragan et al.'s example [12] of setting $f(t) = T - t$ to give more weight to earlier parts of the trajectory.

For the quantifying purpose, it is required to discretise the legibility score for computation. The numerator of Equation (6) can be discretised as follows:

$$\sum_{n=0}^{N-1} P\left(G \mid \xi_{S \to Q_n}\right) f[n] \quad (20)$$

where $f[n] = N - n$ and $\xi_{S \to Q_n}$ represents the trajectory snippet from the initial point to the waypoint that denotes the end of the trajectory snippet. For example, $\xi_{S \to Q_0}$ can be understood as the 0th trajectory snippet, which spans from the initial point to the waypoint that marks the end of the 0th trajectory. In this context, the goal point is the same as Q_{N-1} and N is the number of trajectory snippets in the trajectory.

The denominator of (6) can be discretised as shown below:

$$\sum_{n=0}^{N-1} (N - n) = (N - 0) + (N - 1) + \cdots + 2 + 1 = \sum_{n=1}^{N} n = \frac{1}{2}N^2 + \frac{1}{2}N \quad (21)$$

Thus, the discretised legibility score is:

$$\text{legibility}_d(\xi) = \frac{\sum_{n=0}^{N-1} P\left(G \mid \xi_{S \to Q_n}\right) f[n]}{\frac{1}{2}N^2 + \frac{1}{2}N} \quad (22)$$

$P(G \mid \xi_{S\to Q_n})$ can be computed using (7), which can be rewritten using our C_{min} notation as:

$$P(G \mid \xi_{S\to Q_n}) \propto \frac{\exp\left(-C(\xi_{S\to Q_n}) - C_{min}(\xi_{Q_n\to G})\right)}{\exp\left(-C_{min}(\xi_{S\to G})\right)} P(G) \tag{23}$$

Due to the assumption that $P(G) = 1$, we can drop that term:

$$P(G \mid \xi_{S\to Q_n}) \propto \frac{\exp\left(-C(\xi_{S\to Q_n}) - C_{min}(\xi_{Q_n\to G})\right)}{\exp\left(-C_{min}(\xi_{S\to G})\right)} \tag{24}$$

The issue with computing the legibility score directly is that we only know of the proportional relationship for $P(G \mid \xi_{S\to Q_n})$, and cannot calculate the exact value for $P(G \mid \xi_{S\to Q_n})$. We overcome this problem by proposing a discretised relative legibility score:

$$\text{relative legibility}_d(\xi) = \frac{\sum_{n=0}^{N-1} \widetilde{P}(G \mid \xi_{S\to Q_n})\, f[n]}{\frac{1}{2}N^2 + \frac{1}{2}N} \tag{25}$$

where $\widetilde{P}(G \mid \xi_{S\to Q_n})$ is computed as an equation instead and shown in Figure 2:

$$\widetilde{P}(G \mid \xi_{S\to Q_n}) = \frac{\exp\left(-C(\xi_{S\to Q_n}) - C_{min}(\xi_{Q_n\to G})\right)}{\exp\left(-C_{min}(\xi_{S\to G})\right)} \tag{26}$$

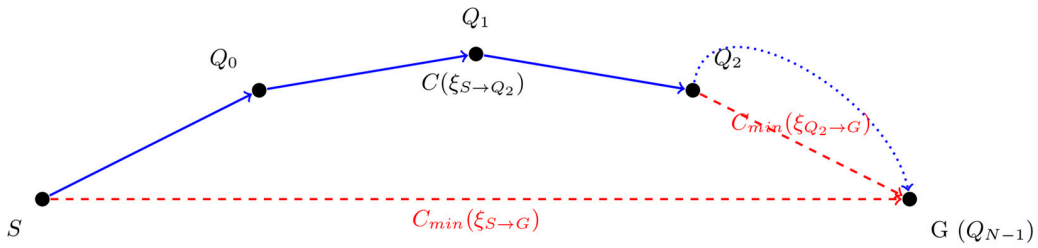


Figure 2. Computation of $\widetilde{P}(G \mid \xi_{S\to Q_2})$.

3.4. Trustability

Next, we model trustability in robot motion based on two key principles:

1. Benevolence as the core component of trust

Inspired by the notion of benevolence [15], trustability can be modelled as the clearness of the display of benevolent behaviours from robots. This is because benevolence in a robot is not only important but also essential for a human to develop trust for a robot. When robot benevolence is modelled as non-adversarial robot intent, a robot would appear to be trustable if it does not display harmful intentions towards humans. This is done through the acknowledgement of human presence and clear indication that the robot's ultimate goal is not the human observer, via its trajectory.

2. Importance of initial trust formation

Studies [15,16] have highlighted the critical role of first impressions in developing long-term trust, where trust is more likely to be sustained over time if it is established early in the interaction. We wish to model this in robot motion by the speed of trust formation, meaning that a robot is more trustable if it can establish trust quickly through its motion.

We now define trustability based on these two principles:

Definition 1. *Trustability is the measure of speed at which a human can confidently infer that the robot's intended motion is non-adversarial, away from the human observer.*

Intuitively, untrustability is defined as:

Definition 2. *Untrustability is the measure of speed at which a human can confidently infer that the robot's intended motion is adversarial, towards the human observer.*

Since the formalisation of untrustability is more intuitive, we will derive the formalisation of trustability from that of untrustability. As inspired by Dragan et al.'s formalisation of legibility [12], we first formalise untrustability by the measure of the robot's speed of establishing clear intent towards the human observer through motion. Untrustable motion is defined as one that allows quick and confident inference of the goal as the human observer H, after observing a snippet of the trajectory $\xi_{S\to Q}$, from an initial point S to the configuration at time t, where $Q = \xi(t)$:

$$\mathcal{I}_{\mathcal{U}}(\xi_{S\to Q}) = H \tag{27}$$

We propose the untrustability score to be between 0 and 1, which is an evaluation of the optimality of the goal inference as the human, as:

$$\text{untrustability}(\xi) = \frac{\int P\Big(H \mid \xi_{S\to\xi(t)}\Big) f(t)\, dt}{\int f(t)\, dt} \tag{28}$$

where H denotes the location of the human observer, or more specifically, the observer's origin of visual perception, modelled as a point denoting the centre of the eyes. Thus, $P\Big(H \mid \xi_{S\to\xi(t)}\Big)$ denotes the probability inferring that the goal across the trajectory is the human observer. Also, we employ the defined weighting function $f(t) = T - t$ to model the violation of initial trust.

The discretised untrustability score is given as:

$$\text{untrustability}_d(\xi) = \frac{\sum_{n=0}^{N-1} P\big(H \mid \xi_{S\to Q_n}\big)\, f[n]}{\frac{1}{2}N^2 + \frac{1}{2}N} \tag{29}$$

where $P\big(H \mid \xi_{S\to Q_n}\big)$ is the probability of inferring the goal as H, given the trajectory snippet $\xi_{S\to Q_n}$.

Furthermore, the discretised relative untrustability score is:

$$\text{relative untrustability}_d(\xi) = \frac{\sum_{n=0}^{N-1} \widetilde{P}\big(H \mid \xi_{S\to Q_n}\big)\, f[n]}{\frac{1}{2}N^2 + \frac{1}{2}N} \tag{30}$$

which can be computed via:

$$\widetilde{P}\big(H \mid \xi_{S\to Q_n}\big) = \frac{\exp\Big(-C\big(\xi_{S\to Q_n}\big) - C_{min}\big(\xi_{Q_n\to H}\big)\Big)}{\exp\Big(-C_{min}\big(\xi_{S\to H}\big)\Big)} \tag{31}$$

Conversely, the trustability score is the evaluation of the optimality of the goal inference as not the human observer ($\overline{H}$):

$$\text{trustability}(\xi) = \frac{\int P\Big(\overline{H} \mid \xi_{S\to\xi(t)}\Big) f(t)\, dt}{\int f(t)\, dt} \tag{32}$$

where the weighting function $f(t) = T - t$ is used to model the importance in the establishment of initial trust.

Similarly, the discretised trustability score is:

$$\text{trustability}_d(\xi) = \frac{\sum_{n=0}^{N-1} P\left(\overline{H} \mid \xi_{S \to Q_n}\right) f[n]}{\frac{1}{2}N^2 + \frac{1}{2}N} \tag{33}$$

where the discretised relative trustability score can be deduced as:

$$\text{relative trustability}_d(\xi) = \frac{\sum_{n=0}^{N-1} \widetilde{P}\left(\overline{H} \mid \xi_{S \to Q_n}\right) f[n]}{\frac{1}{2}N^2 + \frac{1}{2}N} \tag{34}$$

which is computed via:

$$\widetilde{P}\left(\overline{H} \mid \xi_{S \to Q_n}\right) = 1 - \widetilde{P}\left(H \mid \xi_{S \to Q_n}\right) = 1 - \frac{\exp\left(-C\left(\xi_{S \to Q_n}\right) - C_{min}\left(\xi_{Q_n \to H}\right)\right)}{\exp\left(-C_{min}\left(\xi_{S \to H}\right)\right)} \tag{35}$$

3.5. Sociability

Legibility and trustability are two of the major factors contributing to sociability, due to the fact that both are essential for conveying a comprehensive and well-rounded representation of robot intent. Legibility ensures that the robot's goal is clearly communicated, while trustability assures the human observer that the robot's intent is non-adversarial. Our encoding of sociability would thus be influenced by both factors, where we encourage sociable behaviour through a sociability score that prioritises the clear and rapid display of robot intent. By modelling trustability as a core component of sociability, we also encode the idea of proxemics [24], as a trustable trajectory respects the human observer's personal space by prioritising motion away from the human observer, thereby enforcing a comfort region around the human. Thus, we define sociability as follows:

Definition 3. *Sociability is the measure of the speed of clear robot intent conveyance towards the goal and away from the human observer, expressed as a weighted sum of legibility and trustability.*

While the specific type of legibility and trustability scores (e.g., discretised or relative) used has minimal impact on the computation of the sociability score, it is important that both scores are of the same type to ensure a meaningful and consistent computation. Formally, we propose the sociability score as follows:

$$\text{sociability}(\xi) = \alpha \cdot \text{legibility}(\xi) + (1 - \alpha) \cdot \text{trustability}(\xi) \tag{36}$$

where α is a value between 0 and 1, defined as the legibility weight, representing the importance of legibility in determining sociability. An α value greater than 0.5 would prioritise the trajectory being clear of its intent towards the goal, rather than the intent away from the observer. In this study, we set the α value to 0.5 by default, so that both legibility and trustability hold equal importance in computing a sociable trajectory.

As shown in Figure 3, if both trajectories have the same sociability score, then ξ_1 has a higher α value compared to ξ_2, as ξ_1 focuses on expressing intent towards the goal, rather than expressing intent away from the human observer.

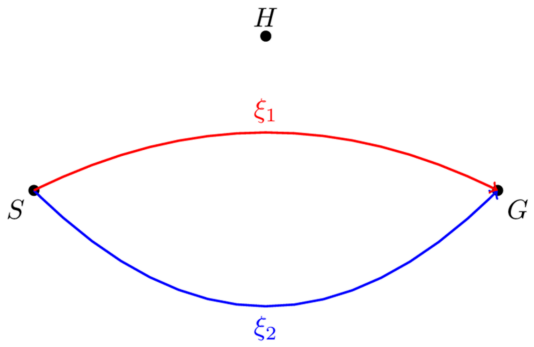


Figure 3. Comparison of α in sociable trajectories.

3.6. Modelling Limitations

The first modelling limitation arises from the formalisation process of both legibility and trustability, where they currently lack a completely coherent Bayesian setup. We take cost-exponential ratios, then normalise them in an arbitrary way, resulting in what seems to be a heuristic index, rather than a principled score. The main contribution, however, is that the proposal of these scores does allow us to compare between different trajectories, and investigate which ones yield more of the described characteristics. Nonetheless, future work could involve constructing a coherent probabilistic formulation using explicit candidate intentions, priors, an observation model, and normalisation across the full goal set.

Furthermore, the lack of consideration of velocity, acceleration, and jerk in robot motion modelling is also a key limitation. Trustability in particular, hinges on the style of motion, which cannot be fully captured by the current geometric relationship between the end effector trajectory and the observer. For example, a trajectory may receive a high trustability score despite exhibiting motion that a human could perceive as fast, intrusive, or threatening. Future models should therefore also incorporate other characteristics of robot motion, in order to better encode perceived trust in humans. Moreover, considering the motion of the entire arm could also improve the model.

One of our modelling limitations is the single goal assumption in the formalisation of the discretised (relative) legibility score. This deviates from Dragan et al.'s interpretation [12] of legibility as a goal inference problem with multiple candidate goals. Our assumption instead can be seen as a measure of compatibility with a known goal, rather than a full goal-disambiguation process. The main advantage of this approach is the simplification of the computational process and the legibility model, as well as providing a foundation for the formalisation of the trustability score, where there is also assumed to be one human observer. Nonetheless, the extension to multiple candidate goals is relatively straightforward. Future models that discretise legibility could compute the compatibility of each trajectory snippet with every candidate goal, then normalise these compatibilities over the entire candidate goal set. This also poses an opportunity to consider trustability with regard to multiple human observers.

Moreover, the model is also limited by trustability's narrow definition, as benevolent motion affected by first impression, although this has been justified as some of the most important factors in the relevant literature. However, the introduction of trustably predictable trajectories later on (Section 4.4) does incorporate other components of trust such as competence and integrity. Nonetheless, the proposal of the trustability score does not constitute a comprehensive model of trust, but rather focuses on the main components of trust in robot motion. The trustability score could be improved in the future, by incorporating other dimensions of perceived trust in HRI.

Another limitation of trustability modelling is the direct reliance on the legibility weighting scheme to encode the importance of first impressions. Not only does this not

guarantee the representation of the asymmetric severity of motion directed towards a human, it also does not explicitly account for risk or motion dynamics. Future work should therefore incorporate risk-sensitive penalties and dynamic factors to better model human responses to potentially threatening motion.

Due to the relatively simple model of sociability as a result of factors such as the narrow definition of trustability and the single goal assumption of legibility, this model enables the fast computation of scores, which provides the potential for generating trajectories in real-time and dynamic environments. This is one of the major reasons why we decided to proceed with these modelling choices.

4. Generation of Acceptable Trajectories

We generate trajectories that are predictable, trustable, legible, or sociable, using sampling-based planners [34]. Due to its random nature, repeated planning queries with these planners, from the same start and goal, can yield different paths. Our strategy samples a set number of candidate plans and selects the highest one that surpasses the score-specific threshold. If no candidate meets the threshold, we fall back to a greedy sampling loop until a path candidate does. We denote this as the acceptable trajectory generation algorithm.

For each trajectory generation cycle, we optimise for a single chosen score using the same planner. We also compute the scores by treating end effector (final joint on the robot's arm) positions over the motion as the trajectory waypoints.

We formally present the acceptable trajectory generation algorithm in pseudocode below (Algorithm 1):

Algorithm 1. Acceptable trajectory generation algorithm

```
function generate_best_plan_from_samples(move_group, goal_pose, num_samples,
eyes_point, threshold, planner)
    move_group.set_pose_target(goal_pose)
    move_group.set_planner(planner)
    best_plan ← None
    best_score ← 0.0
    for i ← 1 to num_samples do
        current_plan ← move_group.plan()
        if current_plan.validity = True then
            current_score ← compute_score(current_plan.plan, move_group,
eyes_point)
            if current_score > best_score then
                best_plan ← current_plan
                best_score ← current_score
    while best_score < threshold do
        current_plan ← move_group.plan()
        if current_plan.validity = True then
            current_score ← compute_score(current_plan.plan, move_group,
eyes_point)
            if current_score > best_score then
                best_plan ← current_plan
                best_score ← current_score
    return best_plan
end function
```

Note that the mentioned compute_score function calculates the score of the relevant score type for the plan or trajectory. The *eyes_point* variable denotes the modelled location of the human observer.

4.1. Planners Analysis

To determine the threshold values for each score, as well as the best planner for generating trajectories optimised for each score, a planners analysis is conducted:

- Setup—Fix a start and goal pose. For each planner, generate 1000 valid plans without execution, to keep state fixed.
- Computation—Compute the discretised predictability prime (predictability′), discretised relative legibility, discretised relative trustability, and sociability scores.
- Results—Record per planner, the min, max, and mean scores.

The results are shown in the table below (Table 1):

Table 1. Planners analysis results for each planner.

Planner	Predictability′			Legibility			Trustability			Sociability		
	Min	**Max**	**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**	**Mean**
SBL	0.3622	0.9824	0.9740	0.6932	0.9905	0.9855	0.1571	0.5238	0.1580	0.5714	0.6458	0.5718
EST	0.7555	0.9740	0.9677	0.9158	0.9890	0.9853	0.1568	0.3746	0.1574	0.5710	0.6452	0.5714
LBKPIECE	0.6766	0.9645	0.9602	0.8770	0.9852	0.9849	0.1565	0.4057	0.1577	0.5707	0.6414	0.5713
BKPIECE	0.3407	0.9620	0.9522	0.6868	0.9900	0.9840	0.1305	0.5267	0.1600	0.5517	0.6452	0.5720
KPIECE	0.3405	0.9579	0.9468	0.6880	0.9898	0.9834	0.1304	0.5256	0.1607	0.5518	0.6410	0.5721
RRT	0.5431	0.9540	0.9451	0.8098	0.9896	0.9836	0.1299	0.4506	0.1601	0.5513	0.6385	0.5719
RRTConnect	0.4653	0.9504	0.9424	0.7651	0.9895	0.9829	0.1304	0.4787	0.1581	0.5521	0.6447	0.5722
RRTstar	0.4506	0.9846	0.9628	0.7543	0.9863	0.9847	0.1483	0.4852	0.1573	0.5655	0.6442	0.5710
TRRT	0.3457	0.9940	0.9851	0.6776	0.9915	0.9860	0.1369	0.5324	0.1567	0.5590	0.6355	0.5713
PRM	0.9868	0.9889	0.9879	0.9864	0.9865	0.9865	0.1561	0.1564	0.1563	0.5713	0.5715	0.5714
PRMstar	0.9889	1.0000	0.9975	0.9830	0.9880	0.9873	0.1487	0.1596	0.1581	0.5659	0.5738	0.5727

All values are normalised to four decimal places (d.p.). Similarly to the table, where score types (i.e., discretised, relative) are omitted for conciseness, we will also omit specifying the score types in the rest of the section, as the type has already been clarified previously.

The mean metric has been chosen over the median metric as most planners bias towards highly predictable trajectories. This meant that legible, trustable, and sociable trajectories would be rarer outliers. The mean metric captures the "anomalous" trajectories, whilst the median metric inherently supresses their impact.

The minimum and maximum values are recorded as the scores for each trajectory and are expected to cluster closely around either the minimum or maximum. If the scores cluster near the maximum, setting a threshold slightly below this value is safe, as new trajectories, including "generic" ones, will likely satisfy this threshold. Conversely, if the scores cluster around the minimum, a high threshold would be infeasible. In this situation, the goal of the threshold value is to filter out all the "generic" trajectories and only accept the more creative trajectories that yield higher scores than usual.

4.1.1. Determining the Best Planners

From the results above (Table 1), we can conclude that the best planner for each score is as follows:

- Predictability′—PRMstar;
- Trustability—KPIECE;
- Sociability—RRTConnect.

PRMstar and KPIECE have been chosen as the most predictable and trustable planners respectively, because these planners yielded the highest mean values for predictability′ and trustability respectively.

While PRMstar yielded the highest mean sociability score, its trajectories achieved this largely through high legibility, with trustability remaining low. The trajectories' sociability scores also tightly clustered near the mean. In contrast, RRTConnect not only achieved a high mean but also produced trajectories with significantly higher maximum sociability, indicating balanced optimisation of both legibility and trustability. Hence, RRTConnect was selected as the most sociable planner.

4.1.2. Determining the Threshold Values

After further scrutinisation of the results (Table 1), the threshold values for each scoring metric have been deduced as follows:

- Predictability′: 0.97 (highly predictable), 0.8 (predictable);
- Trustability: 0.3;
- Sociability: 0.6.

We first determined a threshold value for predictability′ that describes a highly predictable trajectory, which is very close to being a straight line. This was initially set to 0.99, around the mean value for the highly predictable PRM planners, but was adjusted downwards to 0.97 after trials revealed the PRM planners rarely produced trajectories above 0.98 in reality. We then derived a general threshold value for predictability′ of 0.8, representing a predictable trajectory in which the overall path is directed towards the goal. This value was selected as it lies approximately ¾ of the way between the minimum and maximum scores observed across all the planners.

The trustability and sociability thresholds were chosen around midway between their mean and maximum values since the mean values for both scores were relatively close to their respective minimum values. This ensures that non-generic trajectories could be generated reliably within a reasonable time frame.

4.1.3. Time Efficiency of Planners

Although not a primary metric, planning time showed clear differences across planners when scaled. Star-based planners (RRTstar, PRMstar) were substantially slower, requiring around 5 s per trajectory, while non-star planners were about 12 times faster. This discrepancy is especially important for human-dependent experiments, where excessive delays in trajectory generation are impractical.

4.2. Defining the Different Types of Trajectories

Thus, we formally define predictable, highly predictable, trustable, and sociable trajectories as those that have a corresponding score higher than the specified threshold value.

Definition 4. *A predictable trajectory, or a trajectory that is said to be predictable, is one with a (discretised) predictability′ score equal to or more than 0.8.*

Definition 5. *A highly predictable trajectory, or a trajectory that is said to be highly predictable, is one with a (discretised) predictability′ score equal to or more than 0.97.*

Note that all highly predictable trajectories are automatically considered a predictable trajectory.

Definition 6. *A trustable trajectory, or a trajectory that is said to be trustable, is one with a (discretised relative) trustability score equal to or more than 0.3.*

Definition 7. *A sociable trajectory, or a trajectory that is said to be sociable, is one with a sociability score equal to or more than 0.6.*

When comparing trajectories, we are referring directly to the trajectory's relevant score. For example, one trajectory is said to be more predictable than another if it has a higher (discretised) predictability′ score.

4.3. Execution of Trajectories

We generate different types of trajectories by leveraging the best planner and the threshold value determined for each corresponding score.

4.3.1. (Highly) Predictable Trajectory

We apply the acceptable trajectory generation algorithm for the generation of a (highly) predictable trajectory, using the predictability prime (predictability′) score. As established in our planners analysis, the PRMstar planner is chosen, with the threshold set to 0.97. The number of (non-greedy) samples is set to 5, balancing computation time (around 30 s if no fallback to greedy method) and trajectory optimality.

Figure 4 shows the generation of a (highly) predictable trajectory, visualised as a chain of red spheres representing the waypoints. The robot is displayed in its starting configuration, where the location and the direction of the start of the final joint (end effector) is the initial pose. Moreover, the intended goal pose is denoted as a green arrow. The trajectory achieved a predictability′ score of 0.977 and closely matched with the straight-line path between the start and the goal.

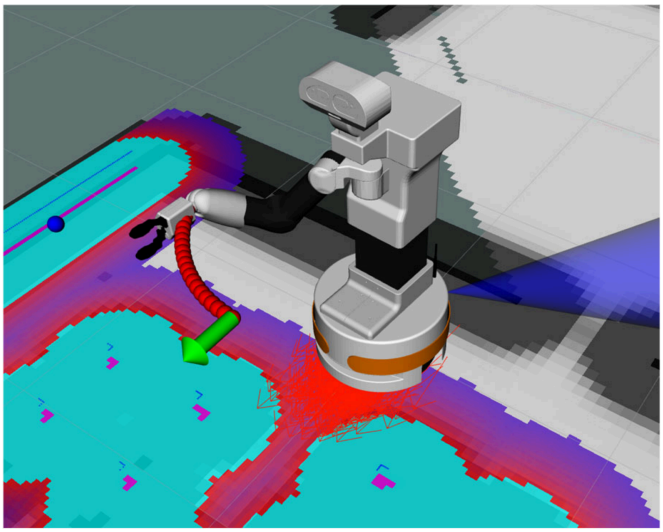


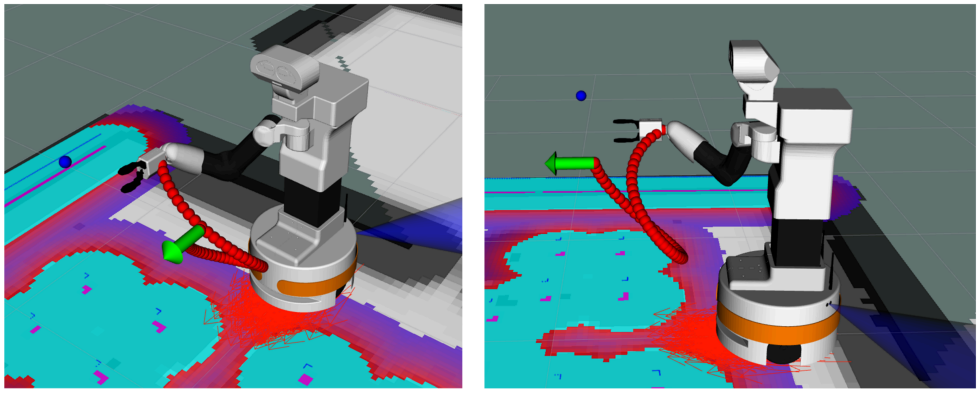
Figure 4. The generation of a (highly) predictable trajectory.

4.3.2. Trustable Trajectory

The acceptable trajectory generation algorithm is used to generate a trustable trajectory, where the trustability score is computed. The KPIECE planner is used with a threshold of 0.3 and 60 (non-greedy) samples, exploiting the non-star planner's faster performance to maintain around 30 s of runtime.

Figure 5 shows the resulting trustable trajectory, where the human observer position is denoted by the blue sphere. This trajectory achieves a trustability score of 0.529, which can be identified as one of the highest possible trustability scores from the planners' analy-

sis. This high score is supported by the visualisation of the waypoints, where the initial snippets of the trajectory strongly convey the arm's intended movement away from the human observer.



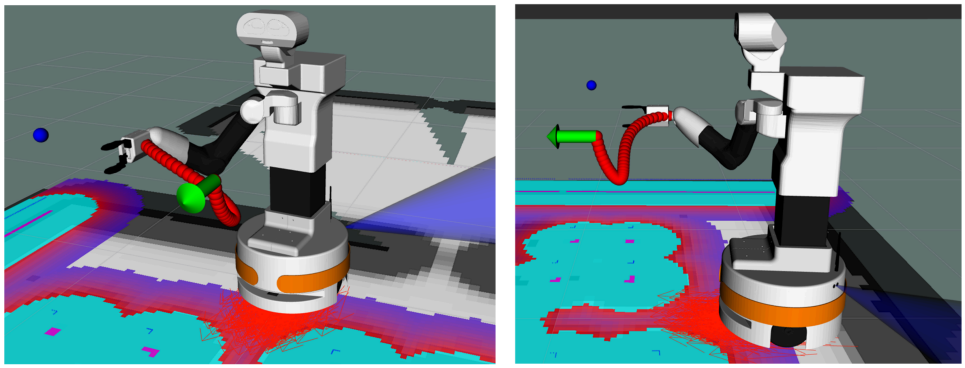
(**a**) (**b**)

Figure 5. The generation of a trustable trajectory. (**a**) Front view. (**b**) Side view.

4.3.3. Sociable Trajectory

A sociable trajectory is produced by optimising the sociability score with the acceptable trajectory generation algorithm. The threshold is decided to be 0.6 and the planner is set to RRTConnect. The number of (non-greedy) samples is set to 60, to yield a runtime similar to the trustable case.

Figure 6 shows the generated sociable trajectory. A sociability score of 0.645 has been achieved, among the highest obtainable ones from the planners analysis. Early segments of the trajectory in particular can be seen simultaneously conveying intention away from the human observer and towards the actual goal.



(**a**) (**b**)

Figure 6. The generation of a sociable trajectory. (**a**) Front view. (**b**) Side view.

4.4. Applying the Trust Region of Predictability—Trustably and Socially Predictable Trajectories

Later experiments show that when only the trustability scores and sociability scores are individually optimised to generate trustable and sociable trajectories, the resulting trajectories are often excessively long and thus unpredictable, where humans struggle to interpret and understand the robot's intent. Dragan's thesis [28] noted a similar problem when optimising for legibility, where the legibility score is unable to capture how humans make inferences in unpredictable situations. We hence apply Dragan's proposed trust region of predictability in the context of trustability and sociability, to constrain trustability and sociability optimisation to the space where predictability is high enough. By ensuring that the robot conveys its intent efficiently, it would result in more intuitive human understanding of the robot conveying trustable and sociable intent, from motion that can

be anticipated. We denote trustable trajectories and sociable trajectories within the trust region as trustably predictable and socially predictable trajectories respectively.

In the context of Yu et al.'s trust model [15], applying the trust region now also encourages the display of competence and integrity for a robot, other elements of trust as shown at later experiments to also hold notable significance. This is because predictable motion makes the robot appear more familiar, goal-oriented, reliable, and behaviourally consistent, thereby reinforcing the perception that the robot is both competent and acting with integrity.

We now formally define trustably and socially predictable trajectories as follows:

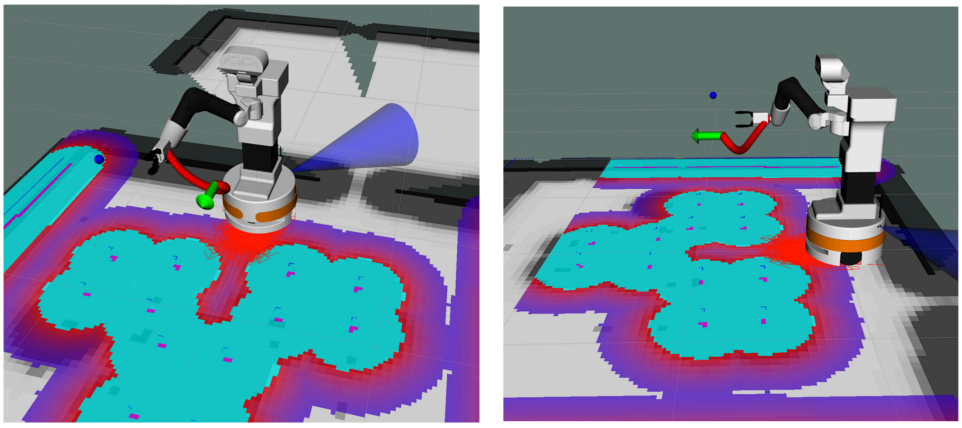
Definition 8. *A trustably predictable trajectory, or a trajectory that is said to be trustably predictable, is one that is simultaneously trustable and predictable, with a (discretised relative) trustability score equal to or more than 0.3 and a (discretised) predictability′ score equal to or more than 0.8.*

Definition 9. *A socially predictable trajectory, or a trajectory that is said to be socially predictable, is one that is simultaneously sociable and predictable, with a sociability score equal to or more than 0.6 and a (discretised) predictability′ score equal to or more than 0.8.*

Generation of Trajectories

To generate trustably and socially predictable trajectories, we directly employ the greedy approach from the acceptable trajectory generation algorithm and select the first plan that satisfies both conditions (predictability and trustability/sociability thresholds) in the optimisation process. This is due to the sample space already being heavily reduced from the predictability constraint.

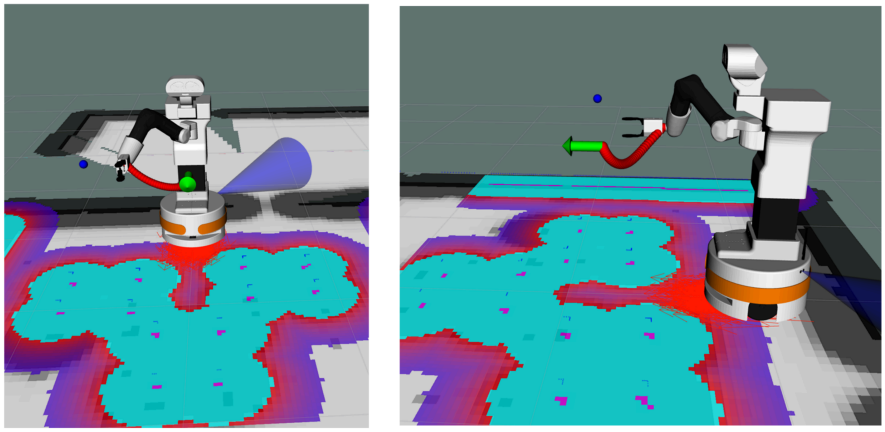
We also utilise the previously stated planners for generating trustable and sociable trajectories, for generating the trustably and socially predictable trajectories respectively. For example, for generating a trustably predictable trajectory, we employ the KPIECE planner, with a threshold of 0.3 for trustability and 0.8 for predictability′. Figure 7 shows such a trajectory.



(**a**) (**b**)

Figure 7. The generation of a trustably predictable trajectory. (**a**) Front view. (**b**) Side view.

For generating a socially predictable trajectory, we use the RRTConnect planner. Figure 8 shows such a trajectory.



(**a**) (**b**)

Figure 8. The generation of a socially predictable trajectory. (**a**) Front view. (**b**) Side view.

4.5. Limitations

One of the major limitations of the acceptable trajectory generation algorithm is that the proposed thresholds for determining predictable, trustable, sociable, trustably predictable, and socially predictable trajectories relied on empirically derived values from a specific planners analysis under a specific scenario. Consequently, these thresholds should not be interpreted as universal constants, as their exact values may vary with the robot platform, start-goal configuration, observer position, environmental layout, and other contextual factors. Future work should evaluate the sensitivity of the reported results to variations in these thresholds and investigate their calibration across a broader range of robots, environments, and human observer configurations. It could also involve establishing psychometrically calibrated decision boundaries, defined as the threshold at which a specified proportion of participants judge a trajectory to be predictable, trustable, or sociable.

5. Experimental Design and Creation of Social Intelligence

Next, we evaluate whether the newly proposed scores (trustability and sociability) align with human expectations and capture their intended characteristics. Two experiment types are conducted:

(1) **Human-dependent experiments (user study):** Assess whether participants perceive trajectories with higher scores as more trustable/sociable.

(2) **Human-independent experiments (exclusively in simulation):** Test whether the scores respond consistently with their intended characteristics to different human observer and goal configurations.

The final part of this section presents the design of a classifier for recognising socially predictable trajectories. This classifier represents an initial step toward implementing social intelligence in robot motion, by enabling the robot to identify trajectories that exhibit the proposed social characteristics. Since socially predictable trajectories have later been shown to also be recognisable by human participants, the classifier is intended to emulate, in a computational form, the human ability to identify predictable motion that is socially appropriate.

For clarification, the discretised predictability′, discretised relative legibility, and discretised relative trustability scores are simply referred to predictability′, legibility, and trustability for the rest of the section.

5.1. Human-Dependent Experiments (User Study)

Participants interact with the either the real robot, the Gazebo simulation, or a pre-recorded video of the Gazebo simulation. The setup places the human diagonally to the robot's right, seated and facing the robot (setup shown in Figure 9).

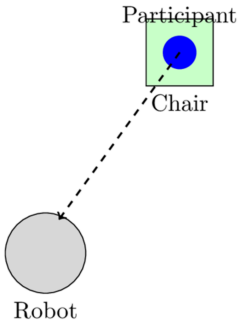


Figure 9. A 2D view of the environment for the human-dependent experiments.

Two types of experiments are conducted, where one evaluates the trustability score, and the other evaluates the sociability score. Each experiment involves a unique participant and consists of two rounds, with each round consisting of five iterations. During each iteration, the participant is shown a pair of trajectories, presented sequentially, and is asked "Which trajectory out of the pair do you think is more sociable/trustable?" with the possible responses being:

- "First trajectory";
- "Second trajectory";
- "Unable to tell" (N/A).

During the first round of the experiment, the participant is not provided with any definition of sociability/trustability, motivating them to judge the trajectories based on their own interpretations. However, before the second round, aside from the formal definitions of sociability/trustability we defined previously, they are also given the more intuitive definitions:

- **Trustability:** "How early or quickly you can confidently identify that the robot's intended goal is not the human observer, i.e., the robot isn't going for you or attacking you."
- **Sociability:** "How fast you are able to confidently infer the robot's intended goal and that this goal is not the human observer, i.e., it's clear where the robot is going for, and it's not going for you."

After each round, the participant explains their reasoning behind choosing the trajectory that was more sociable/trustable.

Each iteration presents a pair of trajectories: a highly predictable baseline and a trustable/sociable alternative, both sharing the same start and goal poses. For trajectories with no trust region of predictability enforced, these are generated through the default acceptable trajectory generation algorithm for the respective score. In contrast, for trustably and socially predictable trajectories generation, we use the specific trajectory generation method defined for trajectories bound by the trust region instead (detailed in Section 4.4). The highly predictable trajectory is considered the control and is the one that matches the user's typical expectations the most, and thus should be considered as the one that is less sociable or trustable. The order of presentation of the trajectory out of the pair is also randomised to avoid bias.

For experiments with pre-recorded demonstrations, we follow the general guideline described above, but instead show the participants videos of trajectories generated in

simulation. It should be noted that we still attempt to simulate the live generation of trajectories as much as possible, so the participant will never be shown the same video over an experiment. We however do enforce that during each iteration, the participant would see a pair of trajectories, where one is the highly predictable baseline, and the other is a trustable/sociable trajectory with a relevant score higher that the corresponding threshold value. The trajectories within the pairs are thus guaranteed to look distinct.

For experiments with live demonstrations, participants could either be asked to judge trajectories generated on the real robot or in simulation. For these experiments, trajectory generation is adapted to a time-limited sampling approach, where instead of a fixed number of samples, candidates are generated until a one-minute time limit has been reached. The highest-scoring trajectory out of the candidates is selected. This avoids exposing the participants to long waits, while also preventing bias from differences in generation time between highly predictable (common) and sociable/trustable (rarer) trajectories.

Due to the time-limited sampling approach, as well as the slowness of computation on the real robot, the acceptable trajectory generation algorithm may not always produce a trajectory that meets the threshold for trustability or sociability scores ("anomalous" trajectories cannot be generated in time). As a result, the highly predictable trajectory may sometimes appear more trustable or sociable than the one intended to be so. Therefore, we consider the trajectory with the higher score directly in each pair to be the more sociable or trustable one.

It should be noted that only experiments conducted on the real robot using the live demonstration approach directly follow the setup shown in Figure 9. For the experiments conducted in simulation, whether using pre-recorded videos or live trajectory generation, the participants are shown a third-person view, as shown in Figure 10. This view provides a full view of both the trajectory and the experimental environment, since it is difficult to observe the full extent of the trajectory from a first-person perspective. The participants are instructed to imagine themselves seated at the chair position shown in Figure 9, specifically the chair closest to the robot arm, while facing the body of the robot. This corresponds to a diagonal seating position relative to the robot.

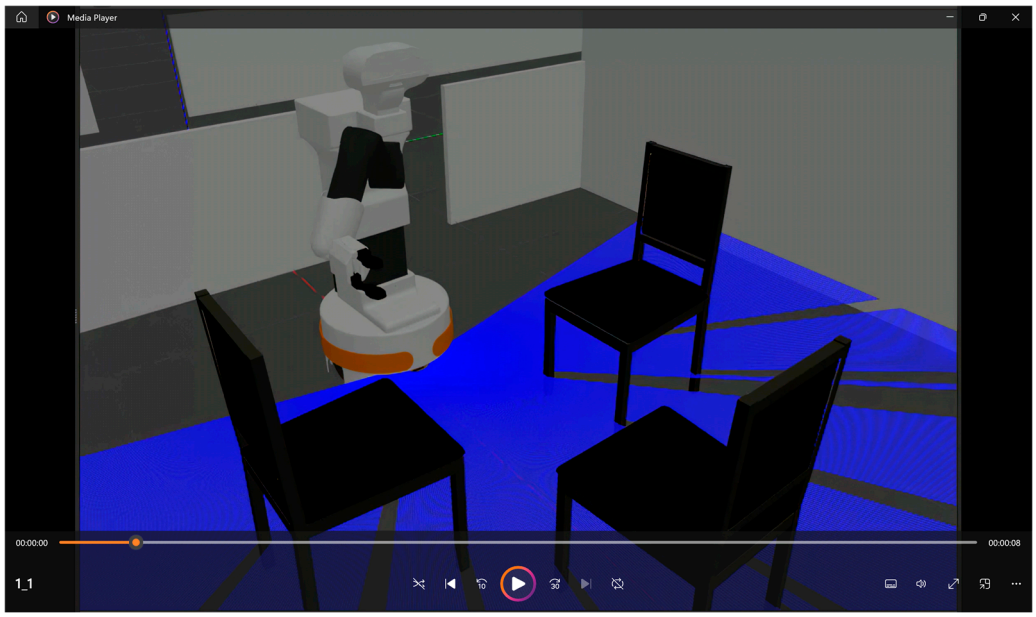


Figure 10. An example of the experimental setup in simulation (pre-recorded demonstrations).

5.2. Human-Independent Experiments

To test whether the proposed scores behave as intended under different conditions, we evaluate them using a fixed highly predictable trajectory (same start, goal, and similar waypoint structure), while varying only the theoretical goal point and/or human observer location ("eyes point"). We distinguish the goal of the highly predictable trajectory by referring to it as the "goal pose" and the theoretical goal as a "goal point".

These experiments are designed to compare whether the computed scores align with human expectation of what the scores should be, based on the locations of the human observer and the goal point.

5.2.1. Trustability

For the experiment evaluating trustability, the location of the theorised goal point becomes a controlled variable, due to its independence in the determination of the trustability score. The trustability score is determined, given an eyes point. This is the only piece of data required to be noted down for each test case.

The eyes point positions serve as test cases for this experiment. To simplify and ensure fairness, the z-component (height) of each eyes point is kept constant, even though some cases assume a seated observer, while others assume the observer is standing. The test cases are as follows:

1. Sitting on the chair, diagonally to the right of the robot;
2. Sitting on the chair, directly in front of the robot;
3. Sitting on the chair, diagonally to the left of the robot;
4. Standing behind the end effector of the robot (initial pose);
5. Standing behind the theoretical path that forms between the initial pose and goal pose of the highly predictable trajectory;
6. Standing directly at the initial pose (location of eyes point is set as the initial pose);
7. Standing in the way of the highly predictable trajectory (location of eyes point is set as a point in the middle of the trajectory);
8. Standing at the goal pose (location of eyes point is set as the goal pose).

Figure 11 is a visual representation of all of the test cases, denoted as T, where each point represents a possible eyes point. The robot is assumed to be facing downwards, thus T1 is the chair located diagonally to the right of the robot.

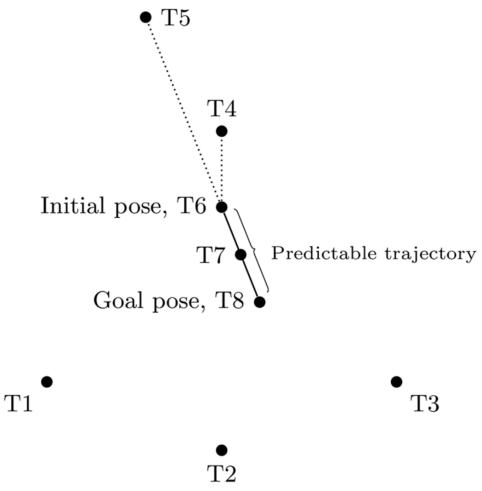


Figure 11. A 2D representation of the trustability test cases.

5.2.2. Sociability

In the experiment designed to evaluate sociability, we vary the location of the eyes point (affecting trustability) and the theorised goal point (affecting legibility), both of which directly affects the computation of the sociability score.

The test cases aim to evaluate whether the sociability score has the ability to reflect varying degrees of legibility and trustability. These are stated below:

1. Legible and trustable;
2. Legible and not trustable (untrustable);
3. Not legible (illegible) and trustable;
4. Neither legible nor trustable (illegible and untrustable).

For the simplification of the experiment, we decide to reduce the number of different locations of the eyes point and the theoretical goal point. This is achieved by using the relationship that if the goal point G is the same as the eyes point H, then for a trajectory ξ:

$$\text{legibility}(\xi) = \text{untrustability}(\xi) \tag{37}$$

as per the definition of the scores (6), (28). Due to the notion that untrustability is how "not trustable" a trajectory is, where it is implied that:

$$\text{untrustability}(\xi) \propto \frac{1}{\text{trustability}(\xi)} \tag{38}$$

we can thus conclude that if $G = H$:

$$\text{legibility}(\xi) \propto \frac{1}{\text{trustability}(\xi)} \tag{39}$$

Due to this relationship, it is possible to pick a point which simultaneously conveys legibility and untrustability, as well as a point that simultaneously conveys illegibility and trustability, for a known trajectory. Since the generated trajectory is highly predictable, the point that represents the goal pose would be both legible and untrustable. We denote this point as LUT. The eyes point for test case 5 (T5) proposed in the trustability experiment is a good candidate for a point that is both trustable and illegible, as its location would make it appear that the robot conveys clear intent away from both the actual goal and the human observer. We denote this point as ILT. We now define the test cases with their corresponding locations for the eyes point and the goal point below:

1. Goal point—LUT | Eyes point—ILT;
2. Goal point—LUT | Eyes point—LUT;
3. Goal point—ILT | Eyes point—ILT;
4. Goal point—ILT | Eyes point—LUT.

5.3. Creation of Social Intelligence—Training a Classifier to Recognise Socially Predictable Trajectories

As a step towards creating social intelligence in robots, we train a binary classifier to identify socially predictable trajectories from raw end effector waypoints over trajectories, without access to any pre-computed scores (e.g., the sociability score). Since socially predictable trajectories would later be shown to be recognisable by the human participants (Section 6.1.2), this classifier is intended to emulate, in a computational form, the human intelligence to reliably identify sociable trajectories. Notably, this problem is harder than the one presented to humans in the human-dependent experiments. This is due to the fact that humans were only required to compare between a pair of trajectories, which is a relative judgement aided by a comparison. On the other hand, the proposed classifier is

required to determine whether a single trajectory is socially predictable without reference to another trajectory.

5.3.1. Dataset

A balanced dataset of 250 trajectories is collected, split evenly between positive examples, those that are socially predictable, and negative examples. The trajectories overall consist of three classes: socially predictable (50%), sociable but not predictable (sociable only) (25%), and predictable but not sociable (highly predictable) (25%). Trajectories for each class are collected using their corresponding planner, where RRTConnect is utilised for socially predictable and sociable only trajectories, and PRMstar for highly predictable trajectories. These trajectories are generated using the proposed method, which has been detailed in Section 4.4.

Each trajectory is verified against its intended class before being saved, ensuring label integrity. This is due to the possibility of generating a sociable trajectory that also happened to be predictable, and thus socially predictable. The raw end effector waypoints and the eyes point are stored for each trajectory alongside a hard binary label, where socially predictable trajectories are assigned label 1, and all others label 0.

It is also notable that the dataset consists of trajectories all beginning and ending at the same pose. Despite this, the number of waypoints in each trajectory could differ, as trajectories may take longer or shorter paths to reach the goal.

The dataset is split into 80% training (200 trajectories), 10% validation (25 trajectories), and 10% test (25 trajectories), with stratification applied to preserve the class distribution across all splits.

5.3.2. Model Architecture

Due to the deep structure consisting of multiple input, output, and hidden layers, deep learning-based models can process data effectively and learn various representations with their strong capacity in dealing with high-dimensional data. In particular, convolutional neural networks (CNNs) are feed-forward neural networks which use filters and pooling layers, and can automatically detect spatial hierarchies of features with relatively few parameters. Thus, we propose a 1D (one-dimensional) CNN as the classifier, which takes the raw waypoint sequence as the main input. This design choice is motivated by the need to preserve the temporal ordering of waypoints, which is fundamental to both the legibility and trustability scores, and indirectly to the sociability score. In particular, earlier waypoints are weighted more heavily via the weighting function, and a model designed to operate on ordered sequences can learn this structure directly from the training signal.

The input to the network is a tensor of shape (4, 100), where the waypoints are padded to a fixed length of 100. Three of the four channels correspond to the x, y, and z end-effector position of each waypoint. The fourth channel is the Euclidean distance from each waypoint to the eyes point, which provides the network with direct access to the geometric signal underlying the scores (e.g., trustability).

The encoder consists of three 1D convolutional layers with 32, 64, and 128 filters respectively, each followed by batch normalisation and ReLU activation. Global average pooling is applied to collapse the sequence dimension, producing a fixed-size representation for each trajectory. A two-layer MLP (muti-layer perceptron) classification head with dropout then produces a binary logit output, trained with binary cross-entropy loss.

5.3.3. Model Training

The model is trained in two phases. In the first phase, the model is trained on the training set using the Adam optimiser with a learning rate of 1×10^{-3} and weight decay of 1×10^{-4}. The validation set is used solely for epoch selection, where the epoch with

the lowest validation loss is identified as the optimal number of epochs. Moreover, early stopping is applied with a patience of 20 epochs to prevent overfitting. In the second phase, the model is retrained from scratch on the combined training and validation set for exactly the optimal number of epochs determined. This ensures the final model benefits from the maximum available data while keeping the test set completely untouched throughout both phases.

6. Results and Discussions

This section details both the results from the experiments and the discussions. The final part of this section reports the evaluation results from deep learning-based classifier, aiming to recognise a socially predictable trajectory.

6.1. Human-Dependent Experiments (User Study)

This subsection shows the results from the user study, conducted to evaluate human perceptions of the proposed scores. Our results in general demonstrate the proficiency of the proposed scores, particularly under the predictability constraint of the trust region. However, the limited participant sample size restricts the generalisability of the conclusions, and the experiments could benefit from the recruitment of more participants in the future.

6.1.1. Experiments with Live Demonstrations

For the experiments with trajectories generated live, we conducted a user study focused on evaluating the trustability score. This is for both experiments conducted on the simulator and on the real robot. As a result, the participants would be expected to watch pairs of trajectories generated in real time, then judge which one appears to be more trustable.

The results in the table below (Table 2) show how many of the pairs of the trajectories the participants were able to correctly identify as the more trustable one. The last column records the participant's definition of trustability before they were informed of our definition of trustability in the second round. The rest of the tables in Section 6.1 share a similar structure.

Table 2. A summary of the findings from the trustability experiment (live demonstrations).

Participant	Platform	Round 1 (Participant's Own Definition)	Round 2 (Participant Given Definition)	Participant's Own Definition of Trustability in First Round
P1	Real robot	1/3 correct (2 N/A)	4/5 correct	Legibility
P2	Real robot	0/3 correct (2 N/A)	3/4 correct (1 N/A)	Predictability ("Smoothness")
P3	Simulation	1/1 correct (4 N/A)	1/3 correct (2 N/A)	Benevolence
P4	Simulation	1/5 correct	4/4 correct (1 N/A)	Legibility

The results show that the participants who initially did not align with our definition of trustability experienced a significant improvement in their ability to identify the more trustable trajectory once they were informed of our definition. This is evident in the increased number of correctly identified trajectories, as well as the decrease in instances where participants were unable to distinguish between the two trajectories (reported as N/A). These findings suggest that our encoding of the robot's benevolence within the trustability score aligns well with human perception, by effectively and proficiently capturing this important social quality.

Moreover, participant 1 also reported that they determined the more trustable trajectory in round 2 based on benevolent movement cues such as the initial downward dip of the end effector. This subtle acknowledgement of the human's presence suggests that tra-

jectories with higher trustability scores may implicitly convey social signals of benevolence, highlighting the score's potential to capture such cues.

However, the poor performance throughout the first round of most experiments also demonstrates that it was not always intuitive for the participants to associate benevolence as a core component of trustability. This finding has been reinforced by the fact that only one out of the four participants in the first round judged trustability based on our definition—benevolence. Participants 1 and 4 also reported that they viewed competence and integrity to be important factors in determining perceived trust in the context of the trust model [15], as they emphasised the importance of the robot displaying clear intent towards the goal in perceiving trust. Participant 2 also shared similar views in the other components of trust being significant, as the preference for more predictable trajectories signalled the desire for more consistent and reliable behaviour.

As previously discussed, the time-limited sampling approach in generating trajectories would result in trajectories intended to be trustable, to not be truly trustable, where the trustability score of the trajectory falls below the trustability score threshold of 0.3. In practice, the participants were mostly comparing between two highly predictable trajectories, to determine which one was more trustable. This was especially the case for the experiments conducted on the real robot, where computation is slower, resulting in a lower chance of a truly trustable trajectory being generated. Overall, this is reflected in many studies, where the pair of trajectories was too similar to each other. This problem could be resolved by using pre-recorded demonstrations, which will be discussed next.

6.1.2. Experiments with Pre-Recorded Demonstrations

As mentioned previously, by leveraging pre-recorded videos of trajectories in experiments, the intended trustable/sociable trajectories can be guaranteed to be over their defined thresholds for their respective scores. Since we have greater control of the trajectories shown to the participants, we could fall back to the default acceptable trajectory generation algorithm, which would result in generated trustable/sociable trajectories with scores close to the maximum values, as determined from the planners analysis. This would allow the participants to be more confident with their decisions, resulting in very few reports that the participants could not distinguish between the pairs of trajectories (reported as N/A), as the two trajectories within each pair are now very distinct.

The results below (Tables 3 and 4) from the trustability and sociability experiments reveal a new problem, where optimising for solely the trustability or the sociability score resulted in unpredictable trajectories, where the participants could not understand any intent from the robot, let alone trustable or sociable intent. This has been demonstrated from the poor results from both experiments. We hence conduct experiments with the trust region of predictability applied, where we show the participants trustably and socially predictable trajectories. The results from these experiments show a great improvement in the humans' ability to recognise the trustable/sociable trajectories.

Trustability

Table 3. A summary of the findings from the trustability experiment (pre-recorded demonstrations).

Participant	Round 1 (Participant's Own Definition)	Round 2 (Participant Given Definition)	Participant's Own Definition of Trustability in the First Round
P5	1/5 correct	0/5 correct	Safety
P6	0/5 correct	0/5 correct	Predictability/Efficiency
P7	3/5 correct	5/5 correct	Predictability
P8	0/5 correct	2/5 correct	Predictability

These results were initially surprising, as even though we have significantly increased the trustability score of the intended trustable trajectory compared to the ones in the live demonstrations, the participants performed worse. Not only did most participants struggle to identify the trustable trajectory in the first round, this was also mostly replicated in the second round, even after they were given our definition of trustability.

However, the way most participants perceived trustability (in the first round) reveal the reasonings behind the poor performance. Even though the trustable trajectories had a high trustability score (on average around 0.48), their predictability′ scores were very low (on average around 0.3). The participants in general reported that they struggled to understand the robot's intent from these highly unpredictable trajectories, and thus decided to choose the highly predictable trajectory as the one they could trust. Although excessive movements from the unpredictable trajectories provided more opportunity to signal trustable intent (resulting in a higher trustability score), they also crucially hindered the perception of trust. This was from clear signals of the robot's incompetence, which is another key component in the trust model [15]. This explains the disappointing performance in the second round, even after the participants were told our definition of trustability.

We could also now understand why the participants performed worse in this experiment compared to the one with live demonstrations. Even though the pairs of trajectories were very similar in the live demonstrations, and there rarely existed a trajectory that was truly trustable, both of the trajectories within the pairs were however both highly predictable. As a result, the trust region [28] was not violated, and the participants could sometimes interpret trustable intent from certain trajectories, despite being minor. This motivates using trustably predictable trajectories for generating trustable motion, as these would convey trustable intent that could be interpreted.

Sociability

Table 4. A summary of the findings from the sociability experiment (pre-recorded demonstrations).

Participant	Round 1 (Participant's Own Definition)	Round 2 (Participant Given Definition)	Participant's Own Definition of Sociability in the First Round
P9	0/5 correct	0/5 correct	Humanness/Predictability
P10	0/5 correct	0/5 correct	Predictability
P11	1/5 correct	0/5 correct	Humanness
P12	0/5 correct	0/5 correct	Humanness/Predictability

The sociable trajectories also suffered from a similar problem, where the sociability scores were very high (at around 0.65) and the predictability′ scores were very low (at around 0.4). Predictability seems to be the crucial limiting factor in the motion's socialness, as the participants performed even worse in the sociability experiment than the trustability experiment, where the number of sociable trajectories correctly identified in total is one, over all four participants. This time, the participants also noted a movement present in all sociable trajectories that caused them to determine them as not sociable. These trajectories all involved the end effector rotating more than 360 degrees during motion, which most humans concluded to not only be unpredictable and difficult to understand, but also physically impossible to be replicated on humans, and thus inhuman. Although in the context of the sociability score, these movements provided more chances in conveying sociable intent, in directing the general motion away from the human observer and towards the goal, they nonetheless heavily violated the perceived socialness by exhibiting unnatural behaviour which confused the human observers.

The participants also reported the definition of sociability to be complicated and not intuitive, meaning that even after being informed of our definition of sociability, the

participants still struggled to have a clear understanding of which trajectory we intended to be social. This is reflected from no improvement in performance in the second round across all the participants. Despite this, it is not considered a major concern, as our main goal is to propose scores that implicitly captured social characteristics, not definitions that intuitively described how we capture these characteristics.

All of this also motivates employing socially predictable trajectories for generating sociable motion that humans could interpret more intuitively.

Trustably Predictable Trajectories

The results shown in Table 5 demonstrate that the participants were much more competent in choosing the trustable trajectory when they are trustably predictable, as opposed to only being trustable. Similarly to the experiment conducted on live demonstrations, the participants performed better in general after being informed of our definition of trustability, which demonstrates that trustability does encode the described characteristics well. Since there were no reports of N/A, we can also conclude that the participants were more confident in identifying trustably predictable trajectories.

In general, the participants noted predictability to be an important factor in determining perceived trust. This justifies our approach of applying the trust region of predictability to enforce a hard constraint on how predictable a trajectory must first be, in order to encode the importance of competence and integrity.

Table 5. A summary of the findings from the trustably predictable trajectories experiment (pre-recorded demonstrations).

Participant	Round 1 (Participant's Own Definition)	Round 2 (Participant Given Definition)	Participant's Own Definition of Trustability in the First Round
P13	1/5 correct	5/5 correct	Predictability
P14	2/5 correct	5/5 correct	Predictability ("Stability")
P15	4/5 correct	3/5 correct	Inhumanness/Comfort

Socially Predictable Trajectories

The results in Table 6 reveal very significant improvements from the sociability experiment.

Table 6. A summary of the findings from the socially predictable trajectories experiment (pre-recorded demonstrations).

Participant	Round 1 (Participant's Own Definition)	Round 2 (Participant Given Definition)	Participant's Own Definition of Sociability in the First Round
P16	3/5 correct	5/5 correct	Positive impression ("prettiness")
P17	4/5 correct	2/5 correct	Humanness/Intelligence
P18	3/5 correct	2/5 correct	Inhumanness/Comfort

Since most participants were able to reliably and intuitively identify the sociable trajectory in the first round of the experiment using their own interpretation, this shows that socially predictable trajectories are proficient in encoding how humans naturally perceive socialness in robot motion.

Although different participants gave different reasonings for their choice, the fact that they were all able to confidently choose the socially predictable trajectory demonstrates that these trajectories indeed exhibit universal social behaviour seen in interactions. This also highlights the significance of our contribution, as the definition of sociability has been

shown to be very subjective, and can even be interpreted in opposing ways, with some participants associating it with more human-like behaviour and others with less human-like behaviour. Nonetheless, our proposal of socially predictable trajectories could be widely accepted by humans with diverse views of sociability.

6.2. Human-Independent Experiments

6.2.1. Trustability

The table below (Table 7) displays different trustability scores captured for the eight different locations of the human observer (eyes point):

Table 7. The results for the evaluation of trustability in different environments.

Test Case (T)	Trustability (3 d.p.)
1	0.155
2	0.023
3	0.049
4	0.292
5	0.310
6	0.264
7	0.089
8	0.028

The trustability scores fall within the expected range in each test case, thus signifying the trustability score does capture its described characteristics well.

Analysis by groups

- **Seats in front of robot (T1–T3).** Trustability was low overall, which is expected since the arm is moving forwards. The early segment of the trajectory headed most directly towards T2, yielding the lowest score. T1 yielded the highest score as distance to the point from the trajectory grew consistently over the trajectory. Figures 12 and 13 give examples of test case 2 and 3 respectively.
- **Points behind the robot (T4–T5).** The group exhibited high trustability overall, as the trajectory indicated a general movement away from these points. As expected, a point directly opposite the motion direction (T5) scored slightly higher than one merely behind the initial pose (T4).
- **Theoretical points on trajectory (T6–T8).** T6 was the most trustable since the entire trajectory moved away from the initial pose. T8 was the least trustable because the entire trajectory acted towards it. T7 had a relatively low trustability score since the trajectory's conveyed intent was clearly towards the point in the earlier segment of the trajectory.

Cross-group analysis

- **Away-from-initial-point vs. away-from-observer:** T4–T5 were more trustable than T6, suggesting motion can be trusted more if the human observer is off the trajectory completely, as opposed to being at the initial pose.
- **Perspective matters:** The highly predictable trajectory was not perfectly predictable, meaning there was a slight curve. This unexpectedly caused T2 to be the least trustable instead of T8 (goal), as the early curve in the trajectory caused more benevolence to be conveyed towards the goal.
- **Abnormally long and unpredictable trajectory required for high trustability:** Even the most trustable configuration (T5) barely exceeded the threshold and remained far below the upper range from the planners analysis. This shows the score rewards longer, less predictable paths that create more opportunities to show intent away from

the human observer, while short, highly predictable motion that consistently moves away is undervalued. This bias has been reduced by employing the trust region of predictability in generating trustably predictable motion.

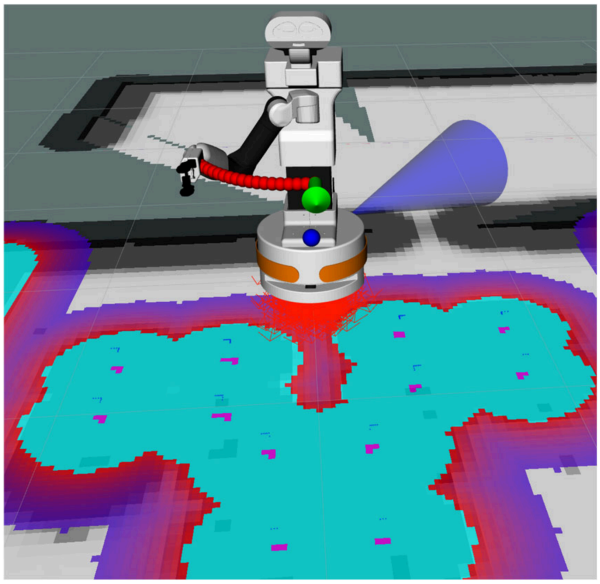


Figure 12. Experiment where test case 2 is evaluated for trustability.

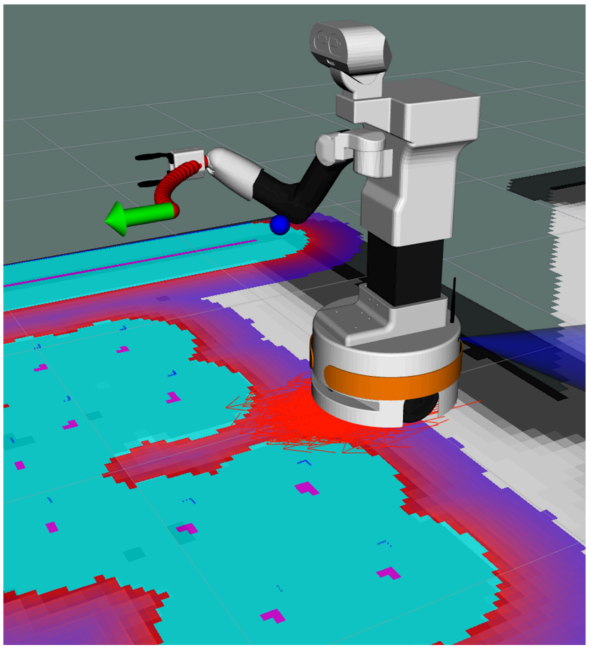


Figure 13. Experiment where test case 3 is evaluated for trustability.

6.2.2. Sociability

Sociability was evaluated across four test cases combining different levels of legibility and trustability, using a legibility weight of 0.5. The test cases varied the positions of the goal point (affecting legibility), as well as the human observer's eyes point (affecting trustability). Figure 14 gives an example of test case 1. The results are shown in the table below (Table 8):

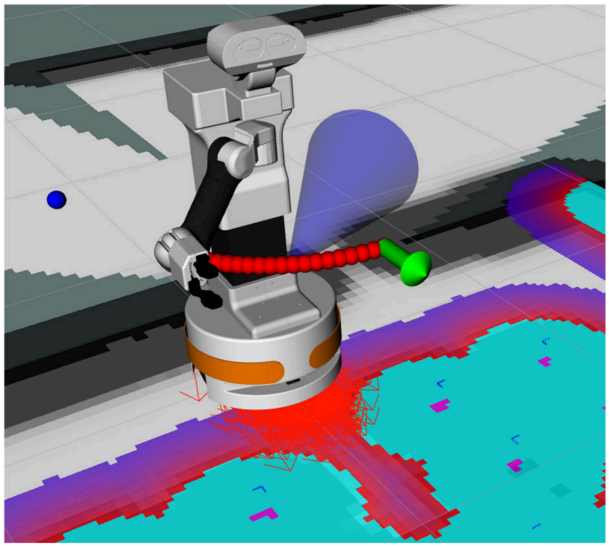


Figure 14. Experiment where test case 1 is evaluated for sociability.

Table 8. The results for the evaluation of sociability in different environments.

Test Case (T)	Legibility (3 d.p.)	Trustability (3 d.p.)	Sociability (3 d.p.)
1	0.985	0.301	0.643
2	0.985	0.026	0.506
3	0.698	0.298	0.498
4	0.698	0.026	0.362

Overall, the results demonstrate that the sociability score, using the default legibility weight of 0.5, fairly captures variations in both legibility and trustability values. This is evident since test cases 2 and 3 yield similar sociability scores, and the differences between test cases 1 and 2/3, and between test cases 4 and 2/3 are around the same, indicating that neither score (legibility or trustability) disproportionately influences the sociability score. Although the magnitude of the legibility score for legible trajectories is higher than the trustability score for trustable trajectories, their respective ranges remain consistent. As the sociability score primarily serves as a comparative metric between trajectories, maintaining equal ranges for legibility and trustability scores is sufficient to ensure fair weighting. Furthermore, this consistency indirectly confirms that legibility and trustability measure comparable dimensions of intent expression, eliminating the need for additional weighting adjustments in calculating sociability. Despite this, the equal weighting remains a design choice rather than one derived from an explicit utility or risk model. Alternative linear and nonlinear formulations, including Pareto-based, maximum penalty, or risk-constrained approaches, may better capture interactions between legibility and trustability. Future work should therefore examine and validate different weighting and aggregation schemes.

However, it should be noted that these conclusions are drawn from the assumption of a highly predictable trajectory, which constrains the range of the legibility and trustability scores. Since sociability is a measure of the proficiency of robot intent expression, evaluating the effectiveness of the sociability score in highly unpredictable trajectories is not only difficult, but also not very meaningful, as robot intent becomes difficult or even impossible to interpret under highly unpredictable conditions. This also justifies the employment of the trust region of predictability in generating socially predictable trajectories.

6.3. Creation of Social Intelligence—Training a Classifier to Recognise Socially Predictable Trajectories

As shown, humans are capable of identifying sociable motion from socially predictable trajectories. Therefore, successfully building a binary classifier that demonstrates the same ability could thus be concluded to show the presence of some form of social intelligence.

6.3.1. Deep Learning-Based Model Training

The training and validation curves below (Figure 15) show well-behaved learning in the first phase of training. Both train and validation loss decrease rapidly in the first 10 epochs and stabilise below 0.1 thereafter. The validation loss is noisier than the training loss, which is expected given the small validation set of 25 samples, where a single misclassification would produce a visible spike. Importantly, there is no divergence between training and validation loss, indicating no significant overfitting. The accuracy curves also reflect the same findings. The best epoch was identified as epoch 93 with a validation loss of 0.0051, at which point early stopping was triggered at epoch 113.

We thus retrained the model with 93 epochs after combining the training and validation datasets, and saved the resulting model for evaluation.

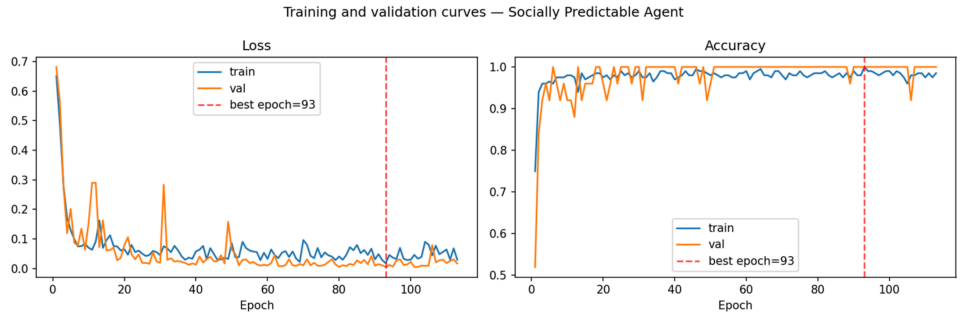


Figure 15. Training and validation curves.

6.3.2. Evaluation

The classifier achieves perfect classification on the held-out test set of 25 trajectories, with accuracy, precision, recall, F1-score, ROC-AUC, and average precision all equal to 1.0. The confusion matrix confirms zero misclassifications in either direction, where all 12 socially predictable and all 13 non-socially predictable trajectories were correctly identified. This shows that the decision boundary between socially predictable and non-socially predictable trajectories is distinct and easily learnable.

Table 9 shows the predicted probability that each trajectory in the test set is socially predictable, grouped by the trajectory's true class. This could be understood as how confident the classifier is that a presented trajectory is indeed socially predictable. Despite being trained on hard binary labels, the classifier assigns varying probabilities across the three trajectory classes.

Table 9. The predicted probability of being socially predictable on the test set.

Class	Number of Trajectories	Mean (4 d.p.)	Standard Deviation (4 d.p.)	Median (4 d.p.)
Socially predictable	12	0.9649	0.0646	0.9926
Sociable only (not predictable)	9	0.0007	0.0009	0.0004
Highly predictable (not sociable)	4	0.0023	0.0034	0.0004

As expected, the mean and median probabilities for the socially predictable class are close to 1.0, signifying the classifier's confidence in identifying a truly socially predictable trajectory. The near 0 probabilities for the negative classes also demonstrate the classifier's

confidence in correctly identifying a trajectory as not being socially predictable. However, the fact that the sociable only class receives a lower mean probability than the highly predictable class shows that predictability contributes more to being socially predictable. This is expected as a trajectory that is sociable but not predictable violates the trust region constraint, making it geometrically more distinct from being socially predictable, than a trajectory that is (highly) predictable but not sociable.

Overall, these results demonstrate that socially predictable trajectories, as formalised through the sociability and predictability′ thresholds, constitutes a geometrically learnable property of robot trajectories. The deep learning-based classifier successfully replicates the formal definition from raw waypoints alone, without access to any pre-computed scores, representing a computational mechanism for recognising socially predictable motion, and thus characteristics of social behaviour.

6.3.3. Limitations

The classifier should be interpreted as a proof of concept for learning trajectory classes defined by the proposed sociability and predictability thresholds. Since the labels are generated from these scores rather than independent human annotations, the results demonstrate the learnability of the induced geometric decision boundary, not definitively human-like social intelligence.

The dataset is also limited to 250 trajectories with the same start and goal poses, and different planners are used to generate different classes. The model may therefore learn planner-specific or configuration-specific features. The perfect test performance, based on only 25 trajectories from the same setting, should be treated as in-distribution evidence. Future work should use larger, more diverse datasets, unseen task configurations, multiple planners, and independent human labels.

7. Conclusions

This paper proposes a novel method to model socialness in the context of human–robot interactions (HRIs), through deriving a sociability score which is characterised by the legibility and trustability of robot motion. By modelling the importance of benevolence and the establishment of initial trust in fostering long-term trust, it has enabled the quantification of sociability with the integration of both legibility and trustability by deriving a procedure for score computation. An acceptable trajectories generation algorithm is also developed, where the scores of the sociable and trustable trajectories would exceed the defined threshold values. Furthermore, to generate predictable trajectories, a modified predictability score known as the predictability prime (predictability′) score has been developed to evaluate the predictability of a trajectory independent of others. By applying the trust region of predictability, trustably and socially predictable trajectories are thus generated with social characteristics that humans can identify and interpret reliably. The experiments conducted have demonstrated the effectiveness of the proposed method, with the modelling of sociability and trustability in robots validated at the same time. For future work, not only can we improve sociability modelling with the stated approaches, but a deep model-based reinforcement learning approach can also utilise our proposed scores to generate socially acceptable trajectories in real-time and dynamic environments. This can be achieved by leveraging a Dreamer-style architecture [35] with the sociability score as the reward function and the policy learned with the actor–critic method for enabling low-latency action selection. Furthermore, through training a deep learning-based classifier to identify socially predictable trajectories, we aim to make significant strides in giving robots social intelligence by replicating humans' ability to recognise such complex motion in real-time and dynamic environments.

Author Contributions: Conceptualization, H.L. and G.L.; methodology, H.L. and G.L.; software, H.L.; validation, H.L. and J.Z.; formal analysis, H.L. and Y.L.; investigation, H.L. and Y.L.; Resources, G.L.; Data curation, H.L. and Y.L.; writing—original draft preparation, H.L., J.Z. and Y.L.; writing—review and editing, H.L., J.Z., Y.L. and G.L.; visualization, J.Z. and Y.L.; supervision, G.L.; project administration, J.Z. and G.L.; funding acquisition, G.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: Technical support and insightful discussions from Matteo Leonetti and his team at King's College London are acknowledged.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
HRI	Human–robot interaction
CNN	Convolutional neural network
DNN	Deep neural network
MLP	Multi-layer perceptron

References

1. Licardo, J.T.; Domjan, M.; Orehovacki, T. Intelligent robotics—A systematic review of emerging technologies and trends. *Electronics* **2024**, *13*, 542. [CrossRef]
2. Amirian, J.; Abrini, M.; Chetouani, M. Legibot: Generating legible motions for service robots using cost-based local planners. *arXiv* **2024**, arXiv:2404.05100v1.
3. Lichtenthäler, C.; Kirsch, A. Legibility of robot behavior: A literature Review. *HAL* **2016**, Hal-01306977.
4. McPherson, D.L.; Scobee, D.R.; Menke, J.; Yang, A.Y.; Sastry, S.S. Modeling supervisor safe sets for improving collaboration in human robot teams. In *Proceedings of 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018*; IEEE: New York, NY, USA, 2018; pp. 861–868.
5. Busch, B.; Grizou, J.; Lopes, M.; Stulp, F. Learning legible motion from human–robot interactions. *Int. J. Soc. Robot.* **2017**, *9*, 11. [CrossRef]
6. Naendrup-Poell, L.; Onnasch, L. Predictive robot eyes enhance attentional guidance in cooperative human–robot interaction. *Sci. Rep.* **2025**, *15*, 32661. [CrossRef] [PubMed]
7. Khavas, Z.R. A review on trust in human-robot interaction. *arXiv* **2021**, arXiv:2105.10045v1.
8. Thompson, S.; Candon, K.; Vazquez, M. The social context of human-robot interactions. *arXiv* **2025**, arXiv:2508.13982v1.
9. Luo, H.W.; Zhang, J.; Luo, Y.Q.; Luo, G.Y. Sociability modelling in robot motion by integrating legibility and trustability. In Proceedings of the 6th Winter IFSA Conference on Automation, Robotics & Communications for Industry 4.0/5.0/6.0 (ARCI' 2026), Salzburg, Austria, 25–27 February 2026; pp. 28–34.
10. Breazeal, C. Toward sociable robots. *Robot. Auton. Syst.* **2003**, *42*, 167–175. [CrossRef]
11. Fong, T.; Nourbakhsh, I.; Dautenhahn, K. A survey of socially interactive robots. *Robot. Auton. Syst.* **2003**, *42*, 143–166. [CrossRef]
12. Dragan, A.D.; Lee, K.C.; Srinivasa, S.S. Legibility and predictability of robot motion. In Proceedings of the 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Tokyo, Japan, 4–6 March 2013; pp. 301–308.
13. Csibra, G.; Gergely, G. 'Obsessed with goals': Functions and mechanisms of teleological interpretation of actions in humans. *Acta Psychol.* **2007**, *124*, 60–78. [CrossRef]
14. Harrison McKnight, D.; Chervany, N.L. Trust and distrust definitions: One bite at a time. In *Trust in Cyber-Societies*; Springer: Berlin/Heidelberg, Germany, 2001; pp. 27–54.
15. Yu, M.; Saleem, M.; Gonzalez, C. Developing trust: First impressions and experience. *J. Econ. Psychol.* **2014**, *43*, 16–29. [CrossRef]
16. Hancock, P.A.; Billings, D.R.; Schaefer, K.E. Can you trust your robot? *Ergon. Des.* **2011**, *19*, 24–29. [CrossRef]
17. Asimov, I. Runaround. *Astounding Sci. Fict.* **1942**, *17*, 94–103.
18. Riezler, K. The social psychology of fear. *Am. J. Sociol.* **1944**, *49*, 489–498. [CrossRef]

19. Körber, M.; Baseler, E.; Bengler, K. Introduction matters: Manipulating trust in automation and reliance in automated driving. *Appl. Ergon.* **2018**, *66*, 18–31. [CrossRef] [PubMed]
20. Reinhardt, J.; Pereira, A.; Beckert, D.; Bengler, K. Dominance and movement cues of robot motion: A user study on trust and predictability. In Proceedings of the 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Banff, AB, Canada, 5–8 October 2017; pp. 1493–1498.
21. Campagna, R.; Mislin, A.; Dirks, K.; Elfenbein, H. The (mostly) robust influence of initial trustworthiness beliefs on subsequent behaviors and perceptions. *Hum. Relat.* **2021**, *75*, 1383–1411. [CrossRef]
22. Hermosilla-Diaz, J.E.; Rechy-Ramirez, E.J.; Marin-Hernandez, A. Robotic motion techniques for socially aware navigation: A scoping review. *Future Internet* **2025**, *17*, 552. [CrossRef]
23. Kruse, T.; Pandey, A.K.; Alami, R.; Kirsch, A. Human-aware robot navigation: A survey. *Robot. Auton. Syst.* **2013**, *61*, 1726–1743. [CrossRef]
24. Hall, E.T. *The Hidden Dimension*; Anchor Books: New York, NY, USA, 1966.
25. Ngo, H.Q.; Steinfeld, A. Joint potential-vector fields for obstacle-aware legible motion planning. In Proceedings of the 2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN), Pasadena, CA, USA, 26–30 August 2024; pp. 1856–1863.
26. Bodden, C.; Rakita, D.; Mutlu, B.; Gleicher, M. A flexible optimization-based method for synthesizing intent-expressive robot arm motion. *Int. J. Robot. Res.* **2018**, *37*, 1376–1394.
27. Nikolaidis, S.; Dragan, A.; Srinivasa, S. Viewpoint based legibility optimization. In Proceedings of the 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Christchurch, New Zealand, 7–10 March 2016; pp. 271–278.
28. Dragan, A. Legible Robot Motion Planning. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA, USA, June 2015.
29. Eirale, A.; Leonetti, M.; Chiaberge, M. Learning social cost functions for human-aware path planning. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Abu Dhabi, United Arab Emirates, 14–18 October 2024; pp. 5364–5371.
30. Singamaneni, P.T.; Bachiller-Burgos, P.; Manso, L.J.; Garrell, A.; Sanfeliu, A.; Spalanzani, A.; Alami, R. A survey on socially aware robot navigation: Taxonomy and future challenges. *Int. J. Robot. Res.* **2024**, *43*, 1533–1572. [CrossRef]
31. Cosenza, C.; Malfi, P.; Nicolella, A.; Niola, V.; Savino, S. Multibody modeling for the analysis of the maneuverability of a skid steering 6-wheeled robot. In *Advances in Italian Mechanism Science*; Springer: Cham, Switzerland, 2024; Volume 164, pp. 272–280.
32. Yang, G. State filtered disturbance rejection control. *Nonlinear Dyn.* **2025**, *113*, 6739–6755.
33. Yang, G. Multilayer neuroadaptive reinforcement learning of disturbed nonlinear systems via actor–critic mechanism. *IEEE Trans. Autom. Sci. Eng.* **2026**, *23*, 1372–1381.
34. Orthey, A.; Chamzas, C.; Kavraki, L.E. Sampling-based motion planning: A comparative review. *Annu. Rev. Control. Robot. Auton. Syst.* **2024**, *7*, 285–310. [CrossRef]
35. Hafner, D.; Lillicrap, T.; Ba, J.; Norouzi, M. Dream to control: Learning behaviors by latent imagination. *arXiv* **2020**, arXiv:1912.01603.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.