Towards Explainable Proactive Robot Interactions for Groups of People in Unstructured Environments

Tamlin Love
tlove@iri.upc.edu
Institut de Robòtica i Informàtica Industrial, CSIC-UPC
Llorens i Artigas 4-6, 08028, Barcelona, Spain

Antonio Andriella
antonio.andriella@pal-robotics.com
Pal Robotics
C/ de Pujades, 77, 08005, Barcelona
Spain

Guillem Alenyà
galenya@iri.upc.edu
Institut de Robòtica i Informàtica Industrial, CSIC-UPC
Llorens i Artigas 4-6, 08028, Barcelona, Spain

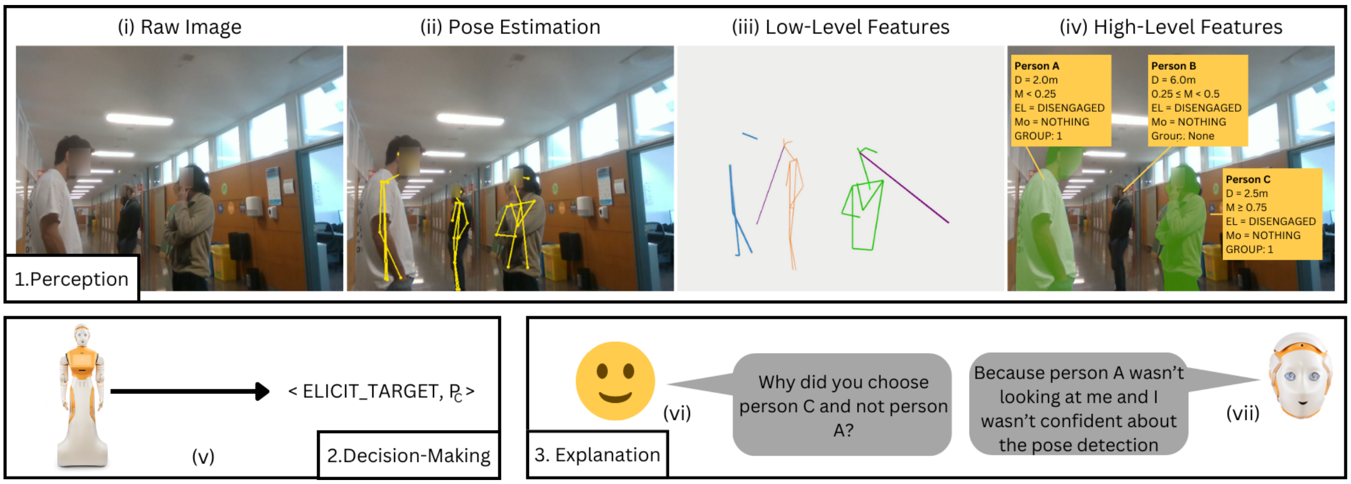

Figure 1: A depiction of our layered system. In the perception layer, the robot receives footage of the scene from the camera (i), estimates the pose of each individual in the scene (ii), and uses the pose to calculate 3D position, velocity and orientation (iii). These lower-level features are used to calculate higher-level features such as pairwise engagement, motion activity and group membership (iv). In the decision-making layer, the robot uses these features to decide which action to take and on whom to target the action (v). Finally, at a later point, a user can query the decision system (vi) and receive an explanation based on counterfactual search (vii).

ABSTRACT

For social robots to be able to operate in unstructured public spaces, they need to be able to gauge complex factors such as human-robot engagement and inter-person social groups, and be able to decide how and with whom to interact. Additionally, such robots should be able to explain their decisions after the fact, to improve accountability and confidence in their behavior. To address this, we present a two-layered proactive system that extracts high-level social features from low-level perceptions and uses these features to make high-level decisions regarding the initiation and maintenance of human-robot interactions. With this system outlined, the primary focus of this work is then a novel method to generate counterfactual explanations in response to a variety of contrastive queries. We provide an early proof of concept to illustrate how these explanations can be generated by leveraging the two-layer system.

HRI '24 Companion, March 11–14, 2024, Boulder, CO, USA

ACM ISBN 979-8-4007-0323-2/24/03…$15.00
https://doi.org/10.1145/3610978.3640734

CCS CONCEPTS

• **Computing methodologies** → **Reasoning about belief and knowledge**; • **Human-centered computing** → *Human computer interaction (HCI).*

KEYWORDS

Explainability, Human-Robot Interaction, Engagement, Proactive Decision-Making

ACM Reference Format:
Tamlin Love, Antonio Andriella, and Guillem Alenyà. 2024. Towards Explainable Proactive Robot Interactions for Groups of People in Unstructured Environments. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24 Companion), March 11–14, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3610978.3640734

1 INTRODUCTION

With social robots increasingly being integrated into everyday environments, such as homes [10], hospitals [4], and public spaces [3, 6], it is necessary for such a robot to interact with people in a way that is natural and accommodates social rules. For example, a service robot positioned in a public space, such as a mall or the lobby of a public building, must be able to (1) determine whether or not prospective interaction partners are amenable to engagement with the robot and (2) select an appropriate action to proactively initiate or maintain such an engagement. There have been many approaches that tackle aspects of this combined problem. For example, models such as SVMs [8, 19] and LSTMs [1] have been used to predict the intention to engage, using features such as face and body orientation, distance to the camera, velocity, etc. Kato et al. [8] and Abbate et al. [1] also go on to implement decision-making behaviours on physical robots in response to engagement predictions. However, these approaches employ so-called "black box" models which can be difficult to explain [16].

Indeed, there has recently been a large push for decision-making systems, including robots, to be explainable - meaning a user is able to find out why the system made a particular decision [15, 16]. Endowing a Human-Robot Interaction (HRI) system with the ability to explain its decisions can improve trust in the robot [17, 20] and aid in understanding and debugging its behaviour [18]. Among current approaches to detect engagement, Bi et al. [5] do use a more transparent model (gradient boosting) and perform an explainability analysis based on feature importance but do not consider multi-person environments (which would complicate the explanation process, as a dynamic number of people requires a dynamic number with which an explanation can be made) or the decision-making of a robot.

In this article, we present a two-layered proactive system that relies on perception (layer 1) and decision-making (layer 2) to allow a robot to autonomously initiate interactions in an unstructured multi-person environment, as depicted in Fig. 1 and described in Sec. 2. The primary contribution of this work is then a counterfactual explanation generation method tailored to this use case to allow for decisions to be contrastively explained *post hoc* in response to a variety of queries, detailed in Sec. 3.

2 PERCEPTION AND DECISION-MAKING

In this section, we describe the components of our two-layered system, namely (1) perception and feature-extraction, and (2) decision-making. While this section details our implementation, we note that the explanation generation method described in Sec. 3 is agnostic to the perception and decision-making components, only requiring a causal model of the features used by the decision-maker.

2.1 Perception

The role of the perception module (labelled 1 in Fig. 1) is to detect each of the people in the scene and calculate a number of high-level features including their engagement with the robot, motion and group membership. In our implementation, RGB and depth video streams are captured by an Intel® RealSense™ Depth Camera D435i and the pose of each person is estimated in real-time using OpenDR's lightweight implementation of OpenPose [13]. From the pose, a person's orientation and velocity can be calculated.

Moving on to higher-level features, we calculate the pairwise engagement between each person (and the robot) using a modification of the visual social engagement metric of Webb et al. [22]. The engagement value between two individuals P_A and P_B is defined as $S_{AB} = min(1, \frac{M_{AB}}{d_{AB}})$, where M_{AB} is the mutual gaze score between P_A and P_B and d_{AB} is the distance between them. The mutual gaze score is defined as the product of the gazes G_{AB} and G_{BA}, which we define as $G_{AB} = max(0, 1 - \frac{\theta_{AB}}{180°})$, where θ_{AB} is the angular distance between P_A's orientation vector and the vector going from P_A to P_B. Thus, $M_{AB} = 1$ when P_A is looking directly at P_B and $M_{AB} = 0$ when P_A has their back on P_B. Following the ROS4HRI standard [12], we use S_{AB}, averaged over a window of time, to determine a categorical engagement. A person's velocity vector is used to determine their motion activity, such as walking away from or towards the robot. Finally, social groups (which may consist of two or more individuals, potentially including the robot) are constructed by linking any pair of individuals whose engagement value S_{AB} is above a given threshold.

Confidence measures for the engagement level, motion and group membership can be calculated from the variance within a sliding window, and confidence in the pose estimation can be retrieved from the pose estimator.

2.2 Decision-Making

The role of the decision-making module (labelled 2 in Fig. 1) is to determine what action the robot should take for a given observation. In this scenario, the robot's decision is a tuple $\langle A, T \rangle$, where A is the action the robot takes and T is the target of the action. The possible values of A are *NOTHING*, *WAIT* (which is used exclusively while the robot is waiting for another action to finish executing), *ELICIT_TARGET* (which is used to get the attention of a specific individual), *ELICIT_GENERAL* (which is used to attract attention with no specific target), *MAINTAIN* (which maintains an existing interaction) and *RECAPTURE* (which attempts to recapture the attention of someone who is starting to disengage from the robot). T can be any person which the robot has detected, or can be nobody if the actions *NOTHING*, *WAIT* or *ELICIT_GENERAL* have been selected.

In the interest of having a lightweight, fully transparent decision-making system, our implementation uses a simple set of rules. Firstly, if the robot is currently executing an action, the decision should be to WAIT. Otherwise, the robot's decision depends on the people it observes. If nobody is observed, the robot does NOTHING. Otherwise, if the robot is in a group with one or more people, the robot must either MAINTAIN the interaction if everyone is engaged or RECAPTURE the attention of disengaging group members. If the robot is not in a group, it must try and elicit engagement. To do so, it calculates a score for each person it detects, based on their motion, distance, mutual gaze, group membership, and the confidence measures of these variables (see Fig. 2 for the list of variables affecting the score). If nobody has a score above a threshold parameter, then the robot will ELICIT_GENERAL, otherwise it will ELICIT_TARGET on the person with the highest score.

3 EXPLANATIONS

In this section we describe out approach to generating explanations for decisions made by the robot in response to *post hoc* queries posed by a user. There are a number of ways to approach the problem of generating explanations. In machine learning, LIME [14] and SHAP [9] are among the most popular, both operating by determining the importance of features in making a classification. However, through their review of literature on explainability in the social sciences, Miller [11] argues that explanations should be contextual. In particular, they identify that explanations are *contrastive* - responding to the query "Why X and not Y?" - and *selected* - only a relevant subset of causes is included in an explanation.

To address the contrastive criterion, an explainer can employ counterfactual reasoning to contrast the queried decision with a counterfactual, hypothetical decision. Some popular approaches in this regard include those of Wachter et al. [21] and Dhurandhar et al. [7], both of which construct loss functions in order to find a counterfactual which is close to the original state but which results in a different decision. In order to select the most relevant variables for an explanation, Albini et al. [2] use a graph-based search to find "critical influences" - variables for which any change would result in a different outcome.

In our implementation, to allow for more expressive counterfactual queries to be made, we allow a user to pose both *why* and *why-not* queries. Given a real decision $D_R = \langle A_R, T_R \rangle$, a user can pose the query "Why D_R and not D_H", where the hypothetical $D_H = \langle A_H, T_H \rangle$. For a counterfactual F with a resulting decision $D_F = \langle A_F, T_F \rangle$ to be valid for a given query, it should satisfy the condition in Eq. 1.

$$
\begin{aligned}
&D_F \neq D_R \text{ if neither } A_H \text{ nor } T_H \text{ are specified} \\
&D_F = \langle A_F = A_H, T_F \rangle \text{ if only } A_H \text{ is specified} \\
&D_F = \langle A_F, T_F = T_H \rangle \text{ if only } T_H \text{ is specified} \\
&D_F = \langle A_H, T_H \rangle \text{ if both } A_H \text{ and } T_H \text{ are specified}
\end{aligned} \tag{1}
$$

In order to find counterfactual explanations that contain a small number of relevant differences, we adapt the notion of critical influences from Albini et al. [2] to respond to the aforementioned queries, to allow for both categorical and continuous (albeit discretised) variables to be included in an explanation, and to allow for causal relationships between variables to be respected. For categorical variables, we adapt the definition of a critical influence in Definition 1.

Definition 1. *Let the observation V_R be an assignment of variables $v \in V$, for which the robot made a decision D_R. Suppose a given categorical variable $x \in V$ took the value $x_R \in X$ for V_R, where X is the set of possible values for x. Given a contrastive query D_H, we say x is a critical influence on D_R if, $\forall x' \in X \setminus \{x_R\}$, an intervention setting $x = x'$, without intervening on any other variables in V, results in a decision that satisfies Eq. 1.*

In other words, if x is a critical influence, then any intervention on x, with no other interventions being made, will result in a decision that satisfies the user's query, and thus its true value must be an important factor in the decision made by the robot. Definition 2 allows for continuous variables to be included in explanations.

Definition 2. *Let V_R, V, D_R, x, x_R, and X retain their definitions from Definition 1, except that x is now a continuous variable with range X. Given a contrastive query D_H, we say x is a critical influence on D_R if $\exists t \in X$ which splits X into two sets $\{x' | x' \leq t\}$ and $\{x' | x' > t\}$, such that $\forall x'$ in whichever set does not include x_R, an intervention setting $x = x'$, without intervening on any other variables in V, results in a decision that satisfies Eq. 1.*

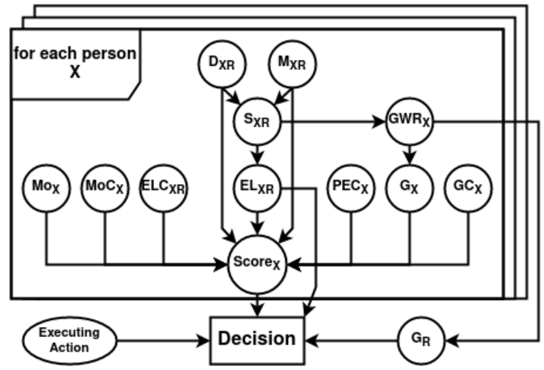

Figure 2: The causal model used by the explanation module. For each person P_X, the model considers their motion Mo_X (confidence: MoC_X), distance to the robot D_{XR}, mutual gaze with the robot M_{XR}, engagement score S_{XR}, engagement level EL_{XR} (confidence: ELC_{XR}), whether or not P_X is in a group with the robot (GWR_X), whether or not P_X is in a group with anyone (G_X; confidence: GC_X), and a confidence measure for the pose estimate (PEC_X) - which all contribute to a score used by the decision-making module. Additionally, the model considers whether or not the robot is executing an action and whether or not the robot is in a group with anyone (G_R).

To search for these critical influences, we require a means of determining the effects of an intervention on a variable, and a means of calculating in which decision a counterfactual would result. To address these requirements, we construct a graphical causal model of the relationships between variables in the observation and the decision made by the robot, depicted in Fig. 2. Such a model is desirable as it can easily be extended or reduced with a dynamic number of people in a given scenario. Given the simplicity of the decision-making module presented in Sec. 2.2, it can be reused directly to calculate the counterfactual decision once interventions have been applied.

Each variable in the observation can then be tested to determine whether or not it is a critical influence, with the possibility of restricting explanations to only include variables relating to certain people (for example, the true target T_R and the queried target T_H). If no explanations consisting of only a single critical influence are found, longer explanations can be generated by performing an intervention, determining the new observation given the intervention, and then searching for critical influences for the new observation.

4 PROOF OF CONCEPT

We present an early proof of concept to demonstrate the generation of explanations using the method provided in Sec. 3 and the HRI system provided in Sec. 2. In this proof of concept we replace the robot with a camera positioned at the end of an office hallway in which a number of participants are positioned, with some interacting with each other and others alone. Note that for the sake of convenience, we continue to refer to the system as the "robot". For the duration of the scenario, the perception and decision-making modules (Sec. 2.1 and 2.2, respectively) are running, taking in the raw RGB-D stream and outputting decisions, logging observation variables, confidences and decisions throughout. After the observations and decisions have been gathered, they can be queried (Sec. 3). The resulting explanations draw from all levels of the system presented in Sec. 2, from low-level observations (such as distance), to high-level features (such as group membership and engagement), to variables related to the decision-making process (such as the flag that indicates whether or not the robot is waiting for an action to execute), and finally to beliefs about these variables (the confidence scores).

Consider the frame depicted in Fig. 1, in which persons P_A, P_B and P_C are visible. In this moment, the robot made the decision $D_R = \langle$ELICIT_TARGET, $P_C\rangle$. In response to the simplest query "Why D_R?", the explanation module returns all explanations consisting of a single critical influence, each of which may serve as a standalone explanation for the decision made:

Explanation 1. *(i) The robot was not already executing an action, (ii-iv) none of the people detected were in a group with the robot, (v) P_A did not have a high engagement score ($S_{AR} < 0.75$), (vi) P_B did not have a high engagement score ($S_{BR} < 0.75$), (vii) P_B was further than $0.75m$ from the camera, (viii) P_C had a high mutual gaze score ($M_{CR} \geq 0.75$), (ix) P_C had a low engagement score ($S_{CR} < 0.5$), (x) P_C was further than $1.25m$ from the camera, and (xi) the pose estimation confidence for P_C was not very low ($PEC_C \geq 0.25$)*

Each of the explanations in Explanation 1 implicitly suggests interventions that would change the decision. For example, included is the somewhat obvious explanation (i) that changing the flag signifying the robot is executing an action would change the decision (in this case to $\langle$WAIT,$\varnothing\rangle$). A somewhat less obvious explanation (x) is that P_C was further than $1.25m$ from the camera, given that P_C was already the target of the robot's decision. However, if an intervention was made to bring P_C within this distance, the resulting decision would be $\langle$MAINTAIN,$P_C\rangle$, keeping the target the same but changing the action.

Given the large number of explanations produced for a simple "Why D_R?" query, which goes against the maxim that explanations should consist of only a few, selected relevant causes [11], a user may wish to make a more directed query by explicitly contrasting the decision with a hypothetical one. For example, the user may ask "Why D_R and not $D_H = \langle\varnothing, P_A\rangle$?" (i.e. "Why not pick P_A?"). Following Eq. 1, to satisfy this query, an explanation must imply interventions that result in the robot choosing any action with P_A as the target. The explanations provided in response to this query are shown in Explanation 2.

Explanation 2. *(i) P_A was not in a group with the robot, and (ii) P_A did not have a high engagement score ($S_{AR} < 0.75$)*

With a restriction on the counterfactual decision, the list of explanations in Explanation 1 is reduced to two. Both of these explanations imply interventions that would place P_A in a group with the robot, resulting in decisions $\langle$RECAPTURE,$P_A\rangle$ and $\langle$MAINTAIN,$P_A\rangle$ for (i) and (ii) respectively. Depending on the user's interest and understanding, this answer may not be useful, as the user may really be interested in why P_C was the target of the ELICIT_TARGET action in particular, and why the same action was not applied to P_A. In this case, a more specific query, "Why D_R and not $D_H = \langle$ELICIT_TARGET,$P_A\rangle$?", can be posed. In this case, no explanations consisting of only one variable are found, but a large number with two variables are identified. Restricting the explanations to only those variables relating to P_A, we arrive at the list in Explanation 3.

Explanation 3. *(i) P_A was not walking towards the camera and... (i-a) ...they were not ENGAGED or ENGAGING, (i-b) ...they had a low mutual gaze score ($M_{AR} < 0.75$), (i-c) ...they had a low engagement value ($S_{AR} < 0.5$). (ii) P_A had a low mutual gaze score ($M_{AR} < 0.75$) and their pose estimation confidence was very low ($PEC_A < 0.25$).*

These explanations point to interventions that would need to be made on P_A to change the decision to $\langle$ELICIT_TARGET,$P_A\rangle$, where a single intervention does not suffice. For example, explanation (ii) implies that P_A would have to be looking at the camera, but also that the confidence in their pose estimate would need to be higher.

5 CONCLUSION

In this article we have presented a two-layered system for an autonomous robot initiating interactions in unstructured, multi-person environments and an approach for explaining such decisions *post hoc* using a counterfactual search in response to a variety of contrastive queries. We have provided a proof of concept which outlines how such a system would operate in practice.

Given the early stage of this research, there are a number of limitations present, which point to directions of future research. One obvious limitation in the proof of concept is that it was performed without a robot actually executing the decisions chosen by the decision-making module. In the future, we plan to implement our system, along with each action, on the PAL Robotics ARI robot, and to conduct experiments in an "in the wild" unstructured setting.

Another limitation is the simplicity of our causal model (see Fig. 2), which does not capture the full dynamics of the real scenario. For example, the effect of factors such as distance and orientation on inter-person group membership has not been considered. Likewise, the relationships the confidence scores maintained by the decision-maker and the other variables have not been modelled. Future work would involve expanding the causal model or incorporating simulation to better capture these causal relationships.

Finally, further future work may involve expanding on the explanation module, including better search algorithms to find critical influences, refining the presentation of explanations to identify which of the options is more relevant, or incorporating a back-and-forth social explanation process as argued for by Miller [11].

6 ACKNOWLEDGMENTS

This work was supported by Horizon Europe under the MSCA grant agreement No 101072488 (TRAIL); by the "European Union NextGenerationEU/PRTR" project CHLOE-GRAPH PID2020-118649RB-I00 funded by MCIN/ AEI /10.13039/ 501100011033; A. Andriella acknowledges the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 801342 (Tecniospring INDUSTRY).

REFERENCES

[1] Gabriele Abbate, Alessandro Giusti, Viktor Schmuck, Oya Celiktutan, and Antonio Paolillo. 2023. Self-supervised prediction of the intention to interact with a service robot. *Robotics and Autonomous Systems* (2023), 104568.

[2] Emanuele Albini, Antonio Rago, Pietro Baroni, and Francesca Toni. 2020. Relation-Based Counterfactual Explanations for Bayesian Network Classifiers.. In *Proceedings of the 2020 International Joint Conferences on Artificial Intelligence*. 451–457.

[3] Antonio Andriella, Ruben Huertas-Garcia, Santiago Forgas-Coll, Carme Torras, and Guillem Alenyà. 2022. "I know how you feel": The importance of interaction style on users' acceptance in an entertainment scenario. *Interaction Studies* 23, 1 (2022), 21–57. https://doi.org/10.1075/is.21019.and

[4] Antonio Andriella, Carme Torras, Carla Abdelnour, and Guillem Alenyà. 2023. Introducing CARESSER: A framework for in situ learning robot social assistance from expert knowledge and demonstrations. *User Modeling and User-Adapted Interaction* 33, 2 (April 2023), 441–496. https://doi.org/10.1007/s11257-021-09316-5

[5] Jian Bi, Fang-chao Hu, Yu-jin Wang, Ming-nan Luo, and Miao He. 2023. A method based on interpretable machine learning for recognizing the intensity of human engagement intention. *Scientific Reports* 13, 1 (2023), 2537.

[6] Zhichao Chen, Yutaka Nakamura, and Hiroshi Ishiguro. 2023. Outperformance of Mall-Receptionist Android as Inverse Reinforcement Learning is Transitioned to Reinforcement Learning. *IEEE Robotics and Automation Letters* (2023).

[7] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. 2018. Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc.

[8] Yusuke Kato, Takayuki Kanda, and Hiroshi Ishiguro. 2015. May I help you? Design of human-like polite approaching behavior. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. 35–42.

[9] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[10] Matteo Luperto, Javier Monroy, Jennifer Renoux, Francesca Lunardini, Nicola Basilico, Maria Bulgheroni, Angelo Cangelosi, Matteo Cesari, Manuel Cid, Aladar Ianes, Javier Gonzalez-Jimenez, Anastasis Kounoudes, David Mari, Victor Prisacariu, Arso Savanovic, Simona Ferrante, and N. Alberto Borghese. 2022. Integrating Social Assistive Robots, IoT, Virtual Communities and Smart Objects to Assist at-Home Independently Living Elders: the MoveCare Project. *International Journal of Social Robotics* 15, 3 (2022), 517–545.

[11] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[12] Youssef Mohamed and Séverin Lemaignan. 2021. Ros for human-robot interaction. In *2021 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 3020–3027.

[13] Nikolaos Passalis, Stefania Pedrazzi, Robert Babuska, Wolfram Burgard, Daniel Dias, Francesco Ferro, Moncef Gabbouj, Ole Green, Alexandros Iosifidis, Erdal Kayacan, Jens Kober, Olivier Michel, Nikos Nikolaidis, Paraskevi Nousi, Roel Pieters, Maria Tzelepi, Abhinav Valada, and Anastasios Tefas. 2022. OpenDR: An Open Toolkit for Enabling High Performance, Low Footprint Deep Learning for Robotics. In *Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (to appear)*.

[14] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?". In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM.

[15] Fatai Sado, Chu Kiong Loo, Wei Shiung Liew, Matthias Kerzel, and Stefan Wermter. 2023. Explainable Goal-driven Agents and Robots-A Comprehensive Review. *Comput. Surveys* 55, 10 (2023), 1–41.

[16] Waddah Saeed and Christian Omlin. 2023. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems* 263 (2023), 110273.

[17] Rossitza Setchi, Maryam Banitalebi Dehkordi, and Juwairiya Siraj Khan. 2020. Explainable robotics in human-robot interactions. *Procedia Computer Science* 176 (2020), 3057–3066.

[18] Stefano Teso, Öznur Alkan, Wolfgang Stammer, and Elizabeth Daly. 2023. Leveraging explanations in interactive machine learning: An overview. *Frontiers in Artificial Intelligence* 6 (2023), 1066049.

[19] Dominique Vaufreydaz, Wafa Johal, and Claudine Combe. 2016. Starting engagement detection towards a companion robot using multimodal features. *Robotics and Autonomous Systems* 75 (2016), 4–16.

[20] Lennart Wachowiak, Oya Celiktutan, Andrew Coles, and Gerard Canal. 2023. A Survey of Evaluation Methods and Metrics for Explanations in Human–Robot Interaction (HRI). In *Workshop on Explainable Robotics at 2023 IEEE International Conference on Robotics and Automation (ICRA)*.

[21] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology (Harvard JOLT)* 31 (2017), 841.

[22] Nicola Webb, Manuel Giuliani, and Séverin Lemaignan. 2022. Measuring visual social engagement from proxemics and gaze. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 757–762.