arXiv:submit/3408521 [cs.RO] 9 Oct 2020

Explainable Representations of the Social State: A Model for Social Human-Robot Interactions.

Daniel Hernandez Garcia, Yanchao Yu, Weronika Sieińska,
Jose L. Part, Nancie Gunson, Oliver Lemon, Christian Dondrup
School of Mathematical and Computer Sciences
Heriot-Watt University
Edinburgh, Scotland, UK

Abstract

In this paper, we propose a minimum set of concepts and signals needed to track the *social state* during Human-Robot Interaction. We look into the problem of complex continuous interactions in a social context with multiple humans and robots, and discuss the creation of an explainable and tractable representation/model of their social interaction. We discuss these representations according to their representational and communicational properties, and organise them into four cognitive domains (scene-understanding, behaviour-profiling, mental-state, and dialogue-grounding).

Understanding the world around us, and the intricate interactions that can take place in it, is a complex topic that has attracted interest from a large number of disciplines spanning philosophy, psychology, sociology, cognitive science, neuroscience, artificial intelligence, computer vision and robotics. While progress has been made (across all fields) and many theories, systems and architectures partially reproducing some parts of these cognitive skills have been developed (Kotseruba and Tsotsos 2018), the full mechanisms of cognition that explain these abilities in humans are still not completely understood (Frith and Frith 2012). Therefore, how to accurately obtain and interpret a representation of the social world remains a current problem for developing Socially Assistive Robots (SARs) (Feil-Seifer and Mataric 2005). This problem is generally approached by the decomposition of the cognitive processes involved and a simplification of the interaction tasks.

In the last few years we can identify major advances in many of these topics, such as environment modelling (Rosinol et al. 2020b), human activity recognition (Kong and Fu 2018), speech recognition and speaker identification (Kanda et al. 2020), conversational agents (Cercas Curry et al. 2018), natural language understanding (Vanzo, Bastianelli, and Lemon 2019), language grounding (Yu, Eshghi, and Lemon 2017), interactive task learning (Chai et al. 2018), analysing interactions and behaviours (Tapus et al. 2019), emotion recognition (Egger, Ley, and Hanke 2019), personality detection (Mehta et al. 2019), inferring intentions (Bianco and Ognibene 2019), and learning human-robot interactions (Liu et al. 2018), among others. Based on the presented background, we define a minimal "social state" which enables continuous social interaction between SARs and humans by supporting reasoning and decision making during interaction with multiple agents and the environment. To this end, we *i)* group the above topics into areas of cognitive domains, presented along two dimensions (see Figure 1) as we consider their representations and interconnections; and *ii)* present a minimal set of signals/information that will be needed to define a *social state*. Maintaining such a model of the world and the social interactions is a complex task which makes decomposition and simplification a necessary abstraction. This abstraction, however, presents not just a challenge but an opportunity to not only reason about the state of the world and the agents within but also facilitate the creation of explainable decision-making engines that are not purely based on raw sensor data, as in most machine learning approaches, but on higher-level representations of the world.

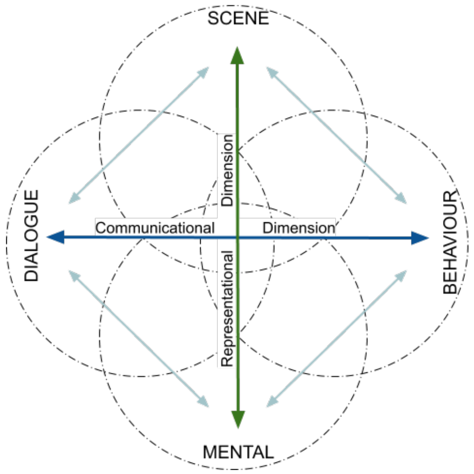

Figure 1: The space of the social-world model is grouped into 4 cognitive domains (scene-understanding, behaviour-profiling, mental-state and dialogue-grounding), organized at opposite ends across 2 functional dimensions: *communicational*, verbal (dialogue) vs non-verbal (behaviour), and *representational*, external (scene) vs internal (mental).

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Model of the Social/World Interaction

The main focus in Human-Robot Interaction (HRI) is often on the robot response, overlooking the way humans interact and preventing the robot from adapting to general situations (Bianco and Ognibene 2020). In order to address this, we need to be able model the whole interaction taking place in a tractable form that can inform other parts of the system. That is, the social nature of an agent should spread through its cognition by creating mechanisms for constructing social representations as an interpretation of the physical world, that allows processing of diverse social situations (Rato, Mascarenhas, and Prada 2020).

Developing SARs with the capacity of performing natural and continuous interactions in a social context with multiple agents in an 'open-domain' requires the robot to show the ability to track and ascribe social meaning to its sensory information. SARs must explore the environment and understand what the environment affords, including which objects, actions, events, and scene information can be extracted from the sensory data. They must track the state of each agent, and track their conversations, determining what they are saying and to whom, where their attention is at, as well as predicting their goals, and their affective and emotional states, etc. They must also 'know' when to perform communicative actions and decide whom to respond to, and when and how to address one or more people. Developments in signal processing, behaviour analysis, multi-modal dialogue, machine learning, and robotics allow us to obtain rich sensory information from the world, but this is not enough as social signals are intrinsically ambiguous (Vinciarelli, Pantic, and Bourlard 2009), and agents must rely on the relationship between such elements (Rato, Mascarenhas, and Prada 2020). We suggest here, to broadly visualise these relationships according to the information that they contain by first categorizing then in terms of two dimensions.

Communication dimension: human actions convey or express social information, either verbally, in dialogues and conversations, or non-verbally, in gestures, behaviours, pose, etc. We also know that the non-verbal behavior of an agent is critically important as well as its verbal behaviour (Vinciarelli, Pantic, and Bourlard 2009). Hence, the social state must track both these domains to maintain conversation and interactions in their social context.

Representational dimension: social interactions are directed by what can be perceived or described from the environment as well as what can be predicted or estimated from the agent's internal belief, states, or desires. The social state must then keep both external (scene) and internal (mental) representations of the environment.

We can see for each of these dimensions, the existence of two domains, differentiated by their nature as explained above. In this way we can organized the representations (see Figure 1) into four domains (dashed circles) set across the two dimensions (solid lines). We propose this as an abstraction, as the dimensions are not to be considered as measurable spectrum but as a construct to distinguished the domains. So that the verbal (dialogue) and non-verbal (behaviour) domain can be intuitively separated on a *communicational* dimension, and the external (scene) and internal (mental) domain can be separated on the *representational* dimension. Furthermore, these domains are not excluding as they can be complementary to each other and can be used to help one another to extract meaning from/for the social context (interconnecting lines among domains), e.g. we can infer a person's intention by their gestures or conversation; dialogue can be grounded by what is known from the scene, or the person's goals; the estimation of a person's behaviour can be informed by the knowledge of the environment or the mental state attributed to them, etc.

Therefore, we propose the formation of four synergetic models for the: *1)* representation of the scene; *2)* representation of the person's behaviours; *3)* representation of the internal mental states; *4)* representation of the conversations.

Scene Understanding

The semantic understanding of a scene is important for social robots applications. Spatial perception and 3D environment understanding are key enablers for high-level task execution in the real world. State-of-the-art approaches use the *Scene Graph* paradigm, (Armeni et al. 2019) provides a hierarchical 3D model that is useful for visualisation and knowledge organisation. (Rosinol et al. 2020b) propose an efficient scene representation data structure which can capture the environment from the lowest level represented as a metric-semantic mesh up through objects and agents in the environment up to rooms and buildings. Metric-semantic understanding provides the capability to simultaneously estimate the 3D geometry of a scene (critical for robots to navigate safely and to manipulate objects) and attach a semantic label to objects and structures (providing models of the environment for a robot to understand and execute human instructions) (Rosinol et al. 2020a).

The model of the social scenario is a representation of the scene, see table 1. Where is the interaction taking place? Who is taking part in the interaction? It models objects, places, structures, and agents and their relations in a way that is physically grounded from the environment by the robot's sensory information.

Table 1: Representation of the scene.

Feature	Signal	Description
locale	metric-semantic, localisation	Where the interaction is taking place. Different rooms can require different interaction strategies.
agents	detection	List of agents in the scene (attended to or not).
objects	detection	List of (salient) objects. With attributes and affordances, etc. Tracked in the interaction.
rooms	metric-semantic	List of places (locales).
scene	scene-graph	Graph/map of the scene. Represents spatial concepts (objects, rooms, agents) and spatio-temporal relations.

Behaviour Profile

The ability to recognise and model physical human activities is a key technology to enable the development of useful HRI applications. The work of (Rossi, Ferland, and Tapus 2017) provides key themes in the context of user profiling mechanisms and behavioral adaptation from the physical, cognitive and social interaction viewpoints. (Aggarwal and Ryoo 2011) provided a classification of the various types of human activities into four different levels: gestures, actions, interactions, and group activities. State-of-the-art solutions on computer vision and deep learning allow analysing the status of each person in the scene, based on body and head pose estimation, face recognition, facial landmarks extraction and the estimation of soft biometric patterns.

The model/profiling of the behaviour is a representation of the people interacting in the scene, see table 2, and combines persistent data of the user for identification, i.e., name, id, role, and dynamic data of the user behaviour, i.e., current activity, focus of attention, location, status.

Table 2: Representation of the person's behaviours.

Feature	Signal	Description
ID	detection	ID of the person.
group ID	detection	Group they belong to.
role	behaviour analysis	People could belong to different roles depending on interaction context.
activity	behaviour analysis	Track current activity.
attention	gaze	Track focus of attention.
location	localization	Track current location.

Mental State

Mentalising, mentalisation, or theory of mind refers to our ability to read the mental states of other agents (Frith and Frith 2006). Findings in developmental psychology concerning current computational theories describing intention understanding and mental state inference from observed actions has inspired the development of architectures for social robots (Bianco and Ognibene 2020). In (Bianco and Ognibene 2019), a summary is provided on how theory of mind features have been integrated in robotic architectures for HRI. (Rabinowitz et al. 2018) designed a neural network which uses meta-learning to build such models of the agents, able to predict the behavior of multiple agents in a false-belief situation given their past and current trajectories. Developing robots with mentalising capabilities for belief understanding, proactivity, active perception and learning of human behavior will further enhance robots' capabilities and improve HRI (Bianco and Ognibene 2019).

The model of the mental state is a representation of the internal model of the robot (agent) state, see table 3. It must track the intentions and beliefs of participants of the interaction, as well as predictions of the goals, motivations and emotional states, etc.

Table 3: Representation of the agent's internal mental states.

Feature	Signal	Description
purpose	mentalising	Agent's main goal.
current target	mentalising	Goal pursued (by the agent) at present.
emotion state	behaviour analysis	Estimation of agent affective expression.
motivation	behaviour analysis	Condition of the agent, i.e, engaged, busy, waiting, etc.

Dialogue Grounding

Humans use Natural Language (NL) to enable interpersonal communication, and articulate their thoughts and intentions. Social robots deployed in diverse human settings will need to interpret and execute high-level instructions given by NL.

Research in language grounding focuses on solving the symbol grounding problem for situated robots by leveraging their interactions with the humans they are working to understand (Thomason et al. 2020). How can we talk to robots about the surrounding world? Can we enable them to interactively learn the grounded meanings needed to finish a task? How can we assist a robot in their navigation or manipulation task with language instructions? Humans and robots need to bridge the gap in their representations to build a common ground of the shared world, for social robots to be able to engage in language communication and joint tasks (Chai et al. 2017). Robots and humans will need to negotiate, using NL, the co-construction of shared representations and plans. This requires the creation of a unified and robust plan modelling and execution framework to combine dialogue actions and physical actions in the same planning domain of the human-robot social interaction (Dondrup, Papaioannou, and Lemon 2019).

The model of the dialogue state is a representation of the conversation, see table 4, tracking what has been said and by whom, the intents, the entities, the topics during interactions among multiple agents.

Table 4: Representation of a conversation dialogue turn.

Feature	Signal	Description
speaker ID	detection	ID of the speaker.
listener IDs	detection	IDs of the listeners.
intent	NLU	The intent of the speaker.
entities	NLU	Entities in the dialogue.
topic	NLU	The current topic/task.
onset	ASR	When the dialogue starts.
transcript	ASR	Transcript of the dialogue.

Memory

So far we have not think about how *memory*, or *time*, relates to the rest of the representations. One way to look at it is as a third dimension, orthogonal to the other two, so the *social state* could be view as a slices of a plane running in a *Memory* domain of current and past *social states*. For now, however, we view temporal relations as "integrated" on the representations, and consider the representations on the *social state* to correspond only to the present *state* in time, and such we don't include an explicit model for *Memory*.

Discussion

In order to create robots able to move, see, hear and communicate in a social context with multiple agents, and properly fulfil social roles and successfully execute social tasks, we need to model the human-robot interactions into an explainable and tractable representation of the social state.

Consider a social robot, tasked with interacting with patients at a clinic/hospital waiting room, in the scenario were a person comes in an approaches the robot. This immediately creates a near infinite number of decisions requiring person identification, tracking, intent beliefs, scene and behavior understanding, etc. It could be a new person, e.g. entering for the first time, and the robot would be required to welcome it, register their details and explain them a procedure to follow next. Or it could be a previously known person, e.g. coming back into the room from having been examined, asked to wait again for a follow up, here it would be unlikely that the robot is required to intervene but a social response acknowledging then back can be desirable. It could also be one of the persons waiting in the room, e.g. coming up to the robot to ask a question like *"where is there a cafeteria or a restroom, or how long till they see them?"*.

Ideally a robot would be capable of providing different responses to all this scenarios. Deep learning approaches could be used to train an agent to learn the proper social interaction strategies (Nanavati et al. 2020), (Romeo et al. 2019). Also multiple deep neural networks can be used to generate the components that will provide inputs to this systems, such as scene understanding (Zhang et al. 2017), face detection (Balaban 2019) and natural language processing (Vanzo, Bastianelli, and Lemon 2019).

These are notorious for being black-box models that are hard, if not impossible, to interpret and which require explanations. Understanding this explanations will be facilitated by having then relate to the descriptions of the *social state*. Successful human-robot social interactions will required not only that robots be able to create internal representations of the physical world, and of collaborative plans about that world, but also that they are able to communicate and negotiate about these representations in a manner that humans can understand.

For instance, in the above scenarios, the *Scene* and *Behaviour* domains would be fundamental to recognize new from known persons in the waiting room which determines different strategies to use to start an interaction with them. The *Mental* domain, to infer a person's interest and needs, is central to direct the best actions and goals the robot should pursue. The *Conversation* domain is essential to lead the both user and robots in helpful dialogues. The synergies between these representations are crucial to generate fruitful interactions, e.g. when asked *"where can I find a cafeteria?"* an advanced *Conversation* representation is need to handle such queries, only relying on *Scene* representation can the information required to answer be extracted, accurate *Mental* representations can allow answer that can satisfied the needs of the user, e.g. the person wants a drink and the robot can answer *"there's a vending machine at X"*, *Behaviour* representation can indicate progress of the interaction, e.g. the person is following the indication from the robot, etc.

The ideas presented here, therefore, constitute a first step towards building a decision-making architecture for multiparty HRI and will be used as the basis for future work on SARs in healthcare.

Acknowledgments

This work was funded by the EU H2020 programme – grant agreement no. 871245 (SPRING).

References

[Aggarwal and Ryoo 2011] Aggarwal, J., and Ryoo, M. 2011. Human activity analysis: A review. *ACM Comput. Surv.* 43(3).

[Armeni et al. 2019] Armeni, I.; He, Z.-Y.; Gwak, J.; Zamir, A. R.; Fischer, M.; Malik, J.; and Savarese, S. 2019. 3d scene graph: A structure for unified semantics, 3d space, and camera. In *Proceedings of the IEEE International Conference on Computer Vision*.

[Balaban 2019] Balaban, S. 2019. Deep learning and face recognition: the state of the art. *CoRR* abs/1902.03524.

[Bianco and Ognibene 2019] Bianco, F., and Ognibene, D. 2019. Functional advantages of an adaptive theory of mind for robotics: a review of current architectures. In *11th Computer Science and Electronic Engineering Conference, CEEC 2019, Colchester, UK, September 18-20, 2019*, 139–143. IEEE.

[Bianco and Ognibene 2020] Bianco, F., and Ognibene, D. 2020. From psychological intention recognition theories to adaptive theory of mind for robots: Computational models. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '20, 136–138. New York, NY, USA: Association for Computing Machinery.

[Cercas Curry et al. 2018] Cercas Curry, A.; Papaioannou, I.; Suglia, A.; Agarwal, S.; Shalyminov, I.; Xinnuo, X.; Dusek, O.; Eshghi, A.; Konstas, I.; Rieser, V.; and Lemon, O. 2018. Alana v2: Entertaining and informative open-domain social dialogue using ontologies and entity linking. In *1st Proceedings of Alexa Prize (Alexa Prize 2018)*.

[Chai et al. 2017] Chai, J. Y.; Fang, R.; Liu, C.; and She, L. 2017. Collaborative language grounding toward situated human-robot dialogue. *AI Magazine* 37(4):32–45.

[Chai et al. 2018] Chai, J. Y.; Gao, Q.; She, L.; Yang, S.; Saba-Sadiya, S.; and Xu, G. 2018. Language to action: Towards interactive task learning with physical agents. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 2–9. International Joint Conferences on Artificial Intelligence Organization.

[Dondrup, Papaioannou, and Lemon 2019] Dondrup, C.; Papaioannou, I.; and Lemon, O. 2019. Petri net machines for human-agent interaction.

[Egger, Ley, and Hanke 2019] Egger, M.; Ley, M.; and Hanke, S. 2019. Emotion recognition from physiological signal analysis: A review. *Electronic Notes in Theoretical Computer Science* 343:35 – 55. The proceedings of AmI, the 2018 European Conference on Ambient Intelligence.

[Feil-Seifer and Mataric 2005] Feil-Seifer, D., and Mataric, M. J. 2005. Defining socially assistive robotics. In *Proceedings of the 9th International Conference on Rehabilitation Robotics, 2005.*, 465–468.

[Frith and Frith 2006] Frith, C. D., and Frith, U. 2006. The neural basis of mentalizing. *Neuron* 50(4):531 – 534.

[Frith and Frith 2012] Frith, C. D., and Frith, U. 2012. Mechanisms of social cognition. *Annual Review of Psychology* 63(1):287–313.

[Kanda et al. 2020] Kanda, N.; Gaur, Y.; Wang, X.; Meng, Z.; Chen, Z.; Zhou, T.; and Yoshioka, T. 2020. Joint speaker counting, speech recognition, and speaker identification for overlapped speech of any number of speakers.

[Kong and Fu 2018] Kong, Y., and Fu, Y. 2018. Human action recognition and prediction: A survey. *CoRR* abs/1806.11230.

[Kotseruba and Tsotsos 2018] Kotseruba, I., and Tsotsos, J. K. 2018. 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review* 53:17–94.

[Liu et al. 2018] Liu, P.; Glas, D. F.; Kanda, T.; and Ishiguro, H. 2018. Learning proactive behavior for interactive social robots. *Autonomous Robots* 42:1067–1085.

[Mehta et al. 2019] Mehta, Y.; Majumder, N.; Gelbukh, A.; and Cambria, E. 2019. Recent trends in deep learning based personality detection. *Artificial Intelligence Review*.

[Nanavati et al. 2020] Nanavati, A.; Doering, M.; Briscic, D.; and Kanda, T. 2020. Autonomously learning one-tomany social interaction logic from human-human interaction data. *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*.

[Rabinowitz et al. 2018] Rabinowitz, N.; Perbet, F.; Song, F.; Zhang, C.; Eslami, S. M. A.; and Botvinick, M. 2018. Machine theory of mind. In Dy, J., and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 4218–4227. Stockholmsmässan, Stockholm Sweden: PMLR.

[Rato, Mascarenhas, and Prada 2020] Rato, D.; Mascarenhas, S.; and Prada, R. 2020. Towards social identity in socio-cognitive agents. *ArXiv*.

[Romeo et al. 2019] Romeo, M.; Hernández García, D.; Jones, R.; and Cangelosi, A. 2019. Deploying a deep learning agent for hri with potential ”end-users” at multiple sheltered housing sites. In *Proceedings of the 7th International Conference on Human-Agent Interaction*, HAI '19, 81–88. New York, NY, USA: Association for Computing Machinery.

[Rosinol et al. 2020a] Rosinol, A.; Abate, M.; Chang, Y.; and Carlone, L. 2020a. Kimera: an open-source library for realtime metric-semantic localization and mapping. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*.

[Rosinol et al. 2020b] Rosinol, A.; Gupta, A.; Abate, M.; Shi, J.; and Carlone, L. 2020b. 3d dynamic scene graphs: Actionable spatial perception with places, objects, and humans. *CoRR* abs/2002.06289.

[Rossi, Ferland, and Tapus 2017] Rossi, S.; Ferland, F.; and Tapus, A. 2017. User profiling and behavioral adaptation for HRI: A survey. *Pattern Recognition Letters* 99:3–12.

[Tapus et al. 2019] Tapus, A.; Bandera, A.; Vazquez-Martin, R.; and Calderita, L. V. 2019. Perceiving the person and their interactions with the others for social robotics – a review. *Pattern Recognition Letters* 118:3 – 13. Cooperative and Social Robots: Understanding Human Activities and Intentions.

[Thomason et al. 2020] Thomason, J.; Padmakumar, A.; Sinapov, J.; Walker, N.; Jiang, Y.; Yedidsion, H.; Hart, J.; Stone, P.; and Mooney, R. J. 2020. Jointly improving parsing and perception for natural language commands through human-robot dialog. *The Journal of Artificial Intelligence Research* 67:327–374.

[Vanzo, Bastianelli, and Lemon 2019] Vanzo, A.; Bastianelli, E.; and Lemon, O. 2019. Hierarchical multi-task natural language understanding for cross-domain conversational AI: HERMIT NLU. In Nakamura, S.; Gasic, M.; Zuckerman, I.; Skantze, G.; Nakano, M.; Papangelis, A.; Ultes, S.; and Yoshino, K., eds., *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, SIGdial 2019, Stockholm, Sweden, September 11-13, 2019*, 254–263. Association for Computational Linguistics.

[Vinciarelli, Pantic, and Bourlard 2009] Vinciarelli, A.; Pantic, M.; and Bourlard, H. 2009. Social signal processing: Survey of an emerging domain. *Image Vis. Comput.* 27:1743–1759.

[Yu, Eshghi, and Lemon 2017] Yu, Y.; Eshghi, A.; and Lemon, O. 2017. Learning how to learn: An adaptive dialogue agent for incrementally learning visually grounded word meanings. *Proceedings of the First Workshop on Language Grounding for Robotics*.

[Zhang et al. 2017] Zhang, Y.; Bai, M.; Kohli, P.; Izadi, S.; and Xiao, J. 2017. Deepcontext: Context-encoding neural pathways for 3d holistic scene understanding. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.