Robot Behavior Adaptation to Human Social Norms

Oliver Roesler, Elahe Bagheri, Amir Aly, Silvia Rossi, Rachid Alami

Proceedings of the Workshop

Robot Behavior Adaptation to Human Social Norms (TSAR)

in Conjunction with the IEEE International Conference on Robot and Human Interactive Communication (Ro-Man)

(August 8-12, 2021)

Oliver Roesler: Vrije Universiteit Brussel, Belgium

Elahe Bagheri: Vrije Universiteit Brussel, Belgium

Amir Aly: University of Plymouth, UK

Silvia Rossi: University of Naples Federico II, Italy

Rachid Alami: CNRS-LAAS, France

Mobile Robotic Telepresence: A New Social Hierarchy?

Cheng Lin[1] Jimin Rhim[1] and AJung Moon[1]

***Abstract*— In the past decade, Mobile Robotic Telepresence (MRP) systems have gained popularity as a tool to enable remote social interactions. However, there is limited work on the social norms that govern the novel human-MRP interactions introduced by these systems. For instance, is it possible that the users piloting MRPs from a remote location and individuals co-located with the MRPs have different expectations about how the other should behave? In this paper, we propose a study to determine if there is a difference in the social hierarchy expected by MRP pilots and co-located users, and to investigate what factors impact this expected social hierarchy.**

I. INTRODUCTION

MRP systems—devices typically characterized by a video conference system mounted on a mobile robotic base [1]—have been adopted and studied in an increasing number of application contexts this past decade (e.g., office, education, elderly care, long-distance relationship, and academic conference settings [2]–[6]). Most of the work investigating the use of MRPs have explored the new types of communication MRPs *enable*, rather than the social interaction norms MRPs *change*.

One of the few studies exploring the latter is Lee and Takayama [2]. In this paper, the authors conducted interviews, field work, and surveys of people in three companies where MRPs had been used for over two months. They found preliminary evidence that the use of MRPs changes what *remote pilots* (those who control the MRP) and *local users* (those who interact locally with the MRP) deem to be socially-acceptable behaviour. However, it is still unclear whether the remote pilot and local user always agree on what these new social norms should be. Our proposed work builds on [2] to conduct an empirical investigation on this topic.

As MRPs are more widely adopted, MRP designers and individuals looking to use MRPs in their organizations will need to address the potential social norm conflicts between pilots and local users. Do the social hierarchies pilots and local users expect the robot to follow differ during human-MRP interactions, and if so, what factors influence these hierarchies? We propose to empirically investigate these social norm expectations; the results of such a study may guide future MRP designs, future decisions to adopt MRPs, and future research.

II. BACKGROUND

The need to study the social hierarchy introduced by novel technologies such as MRPs was first discussed by Paulos and Canny [7]. A decade later, the aforementioned work by Lee and Takayama suggests that, during human-MRP interactions, MRPs may occupy a different part of the social hierarchy than both humans and non-teleoperated machines [2]. However, across the different Human-robot Interactions (HRI) in [2], a consistent pattern in the reported social norms was not found: Local users do not always treat MRPs with the same norms as they would a co-located human (e.g., some local users felt oblidged to help the MRP move around the office), but they do not always treat the systems like any other communication device either (e.g., some local users considered it rude to shut off the MRP without asking the pilot first).

The above findings suggest that the expected social hierarchy between MRPs and local users may depend on the specific HRI setting and environment (*Context*). MRPs are currently being adopted in a plethora of settings, with the goal of making such settings more accessible [2]–[6]. Determining how the social norms that govern MRP use vary across different settings and environments may motivate future research on how MRPs can be designed to be Context-specific, and guide organizations in determining if MRPs are the right choice for their specific application.

In addition to the influence of *Context* on expected social norms, Lee and Takayama point out the following two factors that significantly influence a user's perception of what is considered rude or polite treatment of an MRP: (i) whether the participant referred to the MRP as a "robot" or a "person"; and (ii) whether the participant was a remote pilot or a local user.

The first factor suggests that the norms local users anticipate in their interaction with an MRP depend on their perception of the MRP's autonomy (*Perceived Autonomy*)—as an autonomous robot or an embodiment of the pilot. Booth et al.'s study on user overtrust in robots supports the idea that a user's perception of a robot's autonomy impacts social interactions: They report that whether or not participants communicated with a robot (Turtlebot) depended on the participants' belief that the robot was autonomous vs. teleoperated [8]. Clarifying how the *Perceived Autonomy* of an MRP influences a local user's anticipated norms may motivate more research on how MRPs can communicate a pilot's presence (such as [9]), and guide designers to build MRPs with pilot presence in mind.

The second factor implies that the expected social hierarchy between an MRP and those interacting with it depends on who you ask: the pilot or the local user (*User Type*). Takayama argues that the difference in user experience between a remote pilot and a local user may be so great that

[1]Cheng Lin, Jimin Rhim, and AJung Moon are all with the Department of Electrical & Computer Engineering, McGill University, Montreal, Canada. cheng.lin2@mail.mcgill.ca, jimin.rhim@mcgill.ca, ajung.moon@mcgill.ca

we need separate theories to predict peoples' perceptions of the MRP's agency [10]. In another study of MRPs in office settings, Takayama and Go observe numerous instances of remote pilots and local users employing different metaphors to describe the same MRP, such as remote pilots using human metaphors (e.g., person with disabilities) and local users using nonhuman metaphors (e.g., Skype on wheels) [11]. They note that "mixing metaphors can be quite harmful to interpersonal interactions in the office" [11, p. 501]. Yang et al. similarly find in a study on the use of MRPs in a shopping trip between long-distance couples that remote pilots attributed higher levels of agency to the MRP than local users [5]. This influenced how the parties interacted socially, and whether the local users treated the MRPs as competent adults. The different ways in which pilots and local users order humans and MRPs on the social hierarchy is valuable information for organizations considering MRPs for their remote workers.

Building on the previous work, we propose to explore the relationship between expected social hierarchy of MRPs across the factors *Context*, *Perceived Autonomy*, and *User Type*. We hypothesize the following:

H1. The more the local user perceives the MRP to be an embodiment of the pilot, the higher the local user will rank the MRP on the social hierarchy.

H2. Remote pilots expect MRPs to be treated with a higher social hierarchy than those expected by local users.

H3. The expected social hierarchy between an MRP and those who interact with the robot varies across different settings.

III. METHOD

A. Experiment Design

We propose a 2 (*User Type*: pilot vs. local user) $\times$ 4 (*Context*) $\times$ 2 (MRP display: human face vs. blank screen) $\times$ 2 (Scenario outcome: MRP given right-of-way vs. local user given right-of-way) between-within multi-factor experiment study to empirically investigate if a social hierarchy ordering appears across various human-MRP interactions. In this online experiment, we plan to design and implement video simulations of various HRI (*Contexts*) where a human and an MRP's needs conflict and the social priority ordering is ambiguous. One such *Context* involves a local user and an MRP running into each other while lining up at a store: Would the local user expect to be given right-of-way, or would they allow the MRP to line up first? What does the MRP pilot expect?

Participants will first be split into two groups according to *User Type* (pilots and local users). After answering a demographic questionnaire and reading an introductory excerpt about MRPs, participants will watch a random selection of simulated human-MRP interactions, recorded from the first-person perspective of their assigned *User Type*. Each simulation will be of a different *Context*, and will have accompanying text that introduces the participant to the setting.

For each *User Type* of each *Context*, we will vary two factors: the MRP's display and the outcome of the scenario. We will vary the simulated MRP's display to either contain the face of a human pilot or a blank screen; this will manipulate the user's *Perceived Autonomy*. Secondly, we will vary whether the end of the simulated scenarios show the MRP or the local user receiving right-of-way.

B. Measures and Expected Results

To observe the social hierarchy pilots and local users expect across the independent variables, we will measure the acceptability of the HRI scenario outcomes and the perceived autonomy of the MRP. These measures will be collected in the form of an online questionnaire accompanying each recorded simulation, using acceptability and *Perceived Autonomy* scales that will be validated through a pilot study.

The difference in reported acceptability between the two scenario outcomes of each *Context* will indicate the relative social hierarchy ordering between MRPs and humans. We will conduct statistical analysis on this difference with respect to our experiment factors. We expect to find a significant relationship between the difference in acceptability scores and the *Perceived Autonomy* of the MRP (H1) and also between the difference in acceptability scores and the *User Type* of the participant (H2). We expect to find an effect of varying *Context* on the measures as well (H3).

IV. CONCLUSIONS

Despite the growing popularity of MRPs in various social settings, there is limited work on the social norms that govern human-MRP interactions. Potential conflict between the expected social norms of MRP pilots and local users presents a challenge for both MRP designers and organizations looking to adopt MRPs. Our proposed study contributes to the design and deployment of MRPs through an empirical investigation of the social hierarchy pilots and local users expect and the factors that influence the expected hierarchy.

ACKNOWLEDGEMENT

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

[1] A. Kristoffersson, S. Coradeschi, and A. Loutfi, "A Review of Mobile Robotic Telepresence," *Advances in Human-Computer Interaction*, vol. 2013, p. e902316, Apr. 2013. Publisher: Hindawi.

[2] M. K. Lee and L. Takayama, ""Now, i have a body": uses and social norms for mobile remote presence in the workplace," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, (New York, NY, USA), pp. 33–42, Association for Computing Machinery, May 2011.

[3] A. Page, J. Charteris, and J. Berman, "Telepresence Robot Use for Children with Chronic Illness in Australian Schools: A Scoping Review and Thematic Analysis," *International Journal of Social Robotics*, Nov. 2020.

[4] B. Isabet, M. Pino, M. Lewis, S. Benveniste, and A.-S. Rigaud, "Social Telepresence Robots: A Narrative Review of Experiments Involving Older Adults before and during the COVID-19 Pandemic," *International Journal of Environmental Research and Public Health*, vol. 18, p. 3597, Jan. 2021.

[5] L. Yang, B. Jones, C. Neustaedter, and S. Singhal, "Shopping Over Distance through a Telepresence Robot," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, pp. 191:1–191:18, Nov. 2018.

[6] C. Neustaedter, G. Venolia, J. Procyk, and D. Hawkins, "To Beam or Not to Beam: A Study of Remote Telepresence Attendance at an Academic Conference," in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, (New York, NY, USA), pp. 418–431, Association for Computing Machinery, Feb. 2016.

[7] E. Paulos and J. Canny, "Social Tele-Embodiment: Understanding Presence," *Autonomous Robots*, vol. 11, pp. 87–95, July 2001.

[8] S. Booth, J. Tompkin, H. Pfister, J. Waldo, K. Gajos, and R. Nagpal, "Piggybacking Robots: Human-Robot Overtrust in University Dormitory Security," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, (New York, NY, USA), pp. 426–434, Association for Computing Machinery, Mar. 2017.

[9] J. J. Choi and S. S. Kwak, "Who is this?: Identity and presence in robot-mediated communication," *Cognitive Systems Research*, vol. 43, pp. 174–189, June 2017.

[10] L. Takayama, "Telepresence and Apparent Agency in Human–Robot Interaction," in *The Handbook of the Psychology of Communication Technology*, pp. 160–175, John Wiley & Sons, Ltd, 2015.

[11] L. Takayama and J. Go, "Mixing metaphors in mobile remote presence," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, CSCW '12, (New York, NY, USA), pp. 495–504, Association for Computing Machinery, Feb. 2012.

Does human-robot trust need reciprocity?*

Joshua Zonca, Alessandra Sciutti

***Abstract*— Trust is one of the hallmarks of human-human and human-robot interaction. Extensive evidence has shown that trust among humans requires reciprocity. Conversely, research in human-robot interaction (HRI) has mostly relied on a unidirectional view of trust that focuses on robots' reliability and performance. The current paper argues that reciprocity may also play a key role in the emergence of mutual trust and successful collaboration between humans and robots. We will gather and discuss works that reveal a reciprocal dimension in human-robot trust, paving the way to a bidirectional and dynamic view of trust in HRI.**

I. INTRODUCTION

Humans are inherently social and cooperative beings. This aspect of human behavior is somewhat puzzling, since natural selection should theoretically favor selfish behavior. A crucial mechanism sustaining the emergence and maintenance of cooperation among human is reciprocity, which assumes that one's tendency to cooperate is conditional upon others' cooperation. Reciprocity is also fundamental for the maintenance of mutual trust between peers: if we never trust others, it is unlikely that others will trust us in the future. Indeed, trust among humans is a *relational* phenomenon, which requires that all the individuals involved in interactions and relationships accept a condition of vulnerability to others, believing that others will not exploit this vulnerability [1]. For these reasons, reciprocity has been established in human societies as a social norm [2].

However, reciprocity has not been given a crucial role in human-robot interaction (HRI) research, especially in the study of human-robot trust. The unspoken assumption, which stems from the traditional view of trust in automation, is that the emergence of trust between humans and robots does not need reciprocity due to the intrinsic asymmetrical nature of human-machine relationships. In fact, the research emphasis is almost entirely on the physical, behavioral and functional characteristics of robots: humans trust robots if they are functionally reliable, whereas they do not trust them otherwise.

The main thesis of the current paper is that reciprocity may play a role in supporting mutual trust between humans and robots. In other words, we argue that human trust towards robots may be influenced by the trust expressed, in turn, by robots towards humans during interaction. We will gather existent studies revealing the emergence of reciprocal dynamics in human-robot trust-based interactions and outline a research agenda that see reciprocity as one of the factors shaping human-robot trust and collaboration.

II. TRUST AND RECIPROCITY IN HRI

Trust is undoubtedly one of the main mechanisms supporting collaboration with robots. Trusting our autonomous partners is crucial to delegate responsibility and accept help from them. Historically, research on trust in HRI conceptualized trust as a one-sided process of evaluation of the functional competence and reliability of robotic agents. Extensive evidence has shown that the main determinant of trust in robots is their performance (e.g. [3, 4]). Humans trust robots as long as they show reliable behavior, but they quickly lose trust in in presence of failures [5, 6], leading to disuse of the robotic system [7, 8].

Nonetheless, evidence in HRI highlighted the emergence of distortions in the process of weighting of robots' competence and the relative expression of trust in them. For instance, recent studies have shown that individuals may over-comply with the instructions of robots, even if they have previously shown faulty or unreliable behavior [9-12]. One possibility is that the overt expression of trust towards robots does not always match the individual internal representation of the robot's reliability. This effect might be driven by pro-social attitudes towards social robots, which have been observed in numerous studies (e.g., see [13-15]). At the same time, recent evidence highlighted the emergence of reciprocity in repeated and multi-stage games such as the Prisoner's Dilemma and the Ultimatum Game [16]. Altogether, we hypothesize that trust-based relationships between humans and robots could be shaped, at least in part, by those relational and reciprocal mechanisms typically intervening in human-human interaction.

A recent study by Zonca and colleagues [17] tested this hypothesis by a novel experimental paradigm investigating the emergence of reciprocal trust in human-robot interaction. In a joint task, a human participant and a humanoid robot iCub made perceptual judgments and signaled their trust in the partner. The robot's trust was dynamically manipulated along the experiment and participants could observe both robots' perceptual responses (that were extremely accurate) and trust feedback. Results show that participants did not learn from a robot that was showing high trust in them, since the robot's trust signaled incompetence. However, they were unwilling to disclose their distrust to the robot if they expected future interactions with it. These findings reveal that the overt expression of trust in robots may be modulated by reciprocity, mirroring recent findings observed in human peer interaction and child-adult interaction [18-21].

*Research supported by the European research Council (ERC Starting Grant 804388, wHiSPER).

J. Z. is with the Italian Institute of Technology, Cognitive Architecture for Collaborative Technologies (CONTACT) Unit, Genoa 16152 Italy (corresponding author, e-mail: joshua.zonca@iit.it).

A. S. is with the Italian Institute of Technology, Cognitive Architecture for Collaborative Technologies (CONTACT) Unit, Genoa 16152 Italy (e-mail: alessandra.sciutti@iit.it).

Strohkorb Sebo and colleagues [22] tested the impact of a robot showing vulnerability on human groups of participants during a collaborative game. Results suggest that robots' vulnerability had a "ripple effect" on the trust-related behavior of participants, who were in turn more willing to disclose their vulnerable state to their teammates, reducing the amount of tension of the team. In line with these findings, a recent study [23] revealed that individuals are more prone to trust a robot and collaborate with it when the robot blames itself for collaborative failures.

In line with this "reciprocal" conceptualization of trust, recent works started to model trust from a robot-centered perspective. In particular, the cognitive architecture developed by Vinanzi and colleagues [24, 25] combines trust and Theory of Mind (TOM) modules with an episodic memory system to allow a humanoid robot to evaluate the trustworthiness of human partners in joint tasks. The authors have shown that allowing robots to monitor the current performance of the human partner(s) and take control of the task in case of need enhances collaborative performance in a joint task.

III. TOWARDS A "RECIPROCAL" VIEW OF TRUST IN HRI

Altogether, recent research in HRI put the accent on a bidirectional view of human-robot trust: our trust towards a robot may be influenced by the trust shown by the robot itself, following those reciprocal mechanisms that we generally observe in human-human interaction. In line with this view, social and collaborative robots should be able to adapt their trust-related behavior to modulate and maximize human trust. To achieve this goal, a social robot should be endowed with the ability to track human partners' capabilities and their trust in the robot itself. Moreover, it should react to the ongoing functional and relational joint dynamics to preserve or improve collaboration by increasing its trustworthiness in the eyes of the current human partner(s). Following reciprocal dynamics in interaction, the robot might need to balance the attempts to take the lead or comply with the human partner during a joint task, in order to optimize task-related performance and, at the same time, preserve human-robot trust-based collaboration and social norms.

In this respect, one important question is whether robots should exhibit *negative* reciprocity, that is, should distrust a human partner who does not trust the robot. In fact, this might appear as an anti-social behavior in the eyes of human partner: do we really want robots that distrust and possibly upset humans? The answer depends on the human-robot collaborative context and the relative goals of the interacting partners. In particular, we suggest that negative reciprocal trust could be useful in contexts in which robots must assist people in need (e.g., elderly people, patients with reduced mobility), who should trust the robot to accomplish their everyday goals. In human-human interaction, negative reciprocity has the peculiar function to signal to a selfish or anti-social partner the inappropriateness of their behavior. In many cases, the selfish individual re-starts to behave pro-socially to preserve the trust relationship. In the same way, a robot that negatively reciprocates trust (i.e., a robot that stop to trust the human partner when the human does not trust the robot) would signal that a social norm has been broken, possibly leading the human partner to increase their trust in the robot, with benefit for the human. Ironically, special attention should be put on the implementation of *positive* reciprocal trust in assistive robots. In this case, the robot would reciprocate trust by increasing its own level of trust in the human partner (e.g., a patient), possibly conceding more autonomy to the human. In this scenario, the robot should be careful in blindly reciprocating trust, since it should prioritize the patient's safety, even if this comes at the expenses of social norms.

In this regard, a key aspect concerning the development of "reciprocal" robots is the definition of the actual robots' goals, especially in the case of social robots that would collaborate with humans and assist them. Enabling robots to act with the unique goal of producing specific contextual actions (e.g., lifting a heavy object, accompanying a patient to a specific location) might lead back to an asymmetrical relationship between a human and a robot intended as a mechanical tool. To overcome this limitation, Man and Damasio [26] ambitiously suggested that robots, as intelligent and intentional agents, should hold their own meta-goal of self-preservation, acting as mechanical peers in human societies. This new class of autonomous agents would rely on homeostatic principles, which regulate body and mental states in order to maintain conditions compatible with life. At the same time, Man and Damasio argue that the goal of self-preservation should be combined with empathy, which would prevent robots to arm humans, or other robotic agents. We believe that further research should be conducted to investigate the impact of different robots' high-level goals on the human willingness to trust robots and collaborate with them. In this respect, a delicate issue is how a robot could manage a set of distinct, complementary goals in case of conflict between them. For instance, we need to understand how a robot should decide if reciprocating trust by balancing considerations on the immediate humans' emotional consequences of reciprocity with the long-term benefits of sustained human-robot collaboration.

Furthermore, it is still unclear whether humans should be aware of robots' goals and how this knowledge may influence the emergence of relational dynamics such as reciprocity. Can reciprocity arise when interacting with agents without transparent goals, motives and desires? In fact, reciprocity among humans settled as a social norm due to common knowledge on 1) the immediate, individual incentives to defect during cooperation and on 2) the complementary long-term benefits of cooperative behavior. On the contrary, knowledge of the motivations driving robots' actions can be extremely fuzzy in naïve individuals, possibly hindering the establishment of human-like relational mechanisms. For instance, reciprocity in human-robot interaction requires that humans know that robots are aware of social norms and may comply with them to achieve successful collaboration or please their human partner(s). In this sense, transparency about the purposes underlying the behavior of a robotic system may be crucial in promoting human-robot collaboration [27]. Further research is needed to understand whether a certain degree of transparency is necessary for the establishment of social norms in human-robot interactions. Human knowledge about the robot's goals and motives might be either general (e.g., the robot has the goal to preserve my safety and well-being) or domain-specific (e.g., the robot has the goal to help me get out of bed). Future studies may reveal the impact of partial or complete knowledge about different types of robots'

goals on the establishment of social norms and the success of human-robot interaction and collaboration.

IV. Conclusion

The ambition of designing robotic collaborators, rather than anthropomorphic mechanical tools, opens the question of whether human-robot trust relationships should be reciprocal, as those among human peers. Although research on the role of reciprocity in human-robot trust is still very limited, recent findings suggest that trust towards robots is not a mere function of their perceived competence and reliability. Further research is needed to unveil the extent to which human trust in robots can be shaped by relational and reciprocal dynamics in joint activities. Crucially, these aspects may be fundamental in the design of robots that effectively act as collaborative companions in contexts such as healthcare, rehabilitation, education and assistance for the elderly.

References

[1] D. Ullman and B. F. Malle, "What does it mean to trust a robot? Steps toward a multidimensional measure of trust," *ACM/IEEE Int. Conf. Human-Robot Interact.*, pp. 263-264, 2018.

[2] M. A. Nowak, "Five rules for the evolution of cooperation," *Science*, vol. 314, pp. 1560-1563, 2006.

[3] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Hum. Factors*, vol. 53, pp. 517-527, 2011.

[4] R. van den Brule, R. Dotsch, G. Bijlstra, D. H. Wigboldus and P. Haselager, "Do robot performance and behavioral style affect human trust?" *Int. J. Soc. Robot.*, vol. 6, pp. 519-531, 2014.

[5] M. Desai, M. Medvedev, M. Vázquez, S. McSheehy, S. Gadea-Omelchenko, C. Bruggeman, A. Steinfeld and H. Yanco, "Effects of changing reliability on trust of robot systems," *ACM/IEEE Int. Conf. Human-Robot Interact*., pp. 73-80, 2012.

[6] N. Salomons, M. van der Linden, S. Sebo and B. Scassellati, "Humans conform to robots: Disambiguating trust, truth, and conformity," *ACM/IEEE Int. Conf. Human-Robot Interact*, pp. 187-195, 2018.

[7] B. Lussier, M. Gallien and J. Guiochet, "Fault tolerant planning for critical robots," *Int. Conf. Depend. Syst. Netw*., pp. 144–153, 2007.

[8] T. Sanders, K. E. Oleson, D. R. Billings, J. Y. Chen and P. A. Hancock, "A model of human-robot trust: Theoretical model development," *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 55, pp. 1432-1436, 2011.

[9] M. Salem, G. Lakatos, F. Amirabdollahian and K. Dautenhahn, "Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust," *ACM/IEEE Int. Conf. Human-Robot Interact*., pp. 1-8, 2015.

[10] A. M. Aroyo, F. Rea, G. Sandini and A. Sciutti, "Trust and social engineering in human robot interaction: Will a robot make you disclose sensitive information, conform to its recommendations or gamble?," *IEEE Robot. Autom. Lett.*, vol. 3, pp. 3701-3708, 2018.

[11] P. Robinette, W. Li, R. Allen, A. M. Howard and A. R. Wagner, "Overtrust of robots in emergency evacuation scenarios," *ACM/IEEE Int. Conf. Human-Robot Interact*., pp. 101-108, 2016.

[12] A. M. Aroyo, D. Pasquali, A. Kothig, F. Rea, G. Sandini and A. Sciutti, "Expectations vs. Reality: Unreliability and Transparency in a Treasure Hunt Game with iCub," *IEEE Robot. Autom. Lett.*, vol. 6, pp. 5681-5688, 2021.

[13] J. Connolly, V. Mocz, N. Salomons, J. Valdez, N. Tsoi, B. Scassellati and M. Vázquez, "Prompting prosocial human interventions in response to robot mistreatment," *ACM/IEEE Int. Conf. Human-Robot Interact*., pp. 211–220, 2020.

[14] R. Oliveira, P. Arriaga, F. P. Santos, S. Mascarenhas and A. Paiva, "Towards prosocial design: A scoping review of the use of robots and virtual agents to trigger prosocial behavior," *Comput. Hum. Behav.*, 106547, 2021.

[15] P. H. Kahn Jr, T. Kanda, H. Ishiguro, B. T. Gill, S. Shen, H. E. Gary and J. H. Ruckert, "Will people keep the secret of a humanoid robot? Psychological intimacy in HRI." *ACM/IEEE Int. Conf. Human-Robot Interact.*, pp. 173-180, 2015.

[16] E. B. Sandoval, J. Brandstetter, M. Obaid and C. Bartneck, "Reciprocity in human-robot interaction: a quantitative approach through the prisoner's dilemma and the ultimatum game," *Int. J. Soc. Robot.*, vol. 8, pp. 303-317, 2016.

[17] J. Zonca, A. Folsø and A. Sciutti, "If you trust me, I will trust you: the role of reciprocity in human-robot trust," submitted for publication, arXiv preprint: http://arxiv.org/abs/2106.14832

[18] A. Mahmoodi, B. Bahrami, and C. Mehring, "Reciprocity of social influence," *Nat. Commun*., vol. 9, pp. 1-9, 2018.

[19] J. Zonca, A. Folsø and A. Sciutti, "Dynamic modulation of social influence by indirect reciprocity," *Sci Rep*., vol. 11, 11104, 2021.

[20] J. Zonca, A. Folsø and A. Sciutti, "I'm not a little kid anymore! Reciprocal social influence in child-adult interaction," accepted for publication, *R. Soc. Open Sci.*, PsyArXiv preprint: https://doi.org/10.31234/osf.io/vsdpe

[21] J. Zonca, A. Folsø and A. Sciutti, "Trust is not all about performance: trust biases in interaction with humans, robots and computers," arXiv preprint: http://arxiv.org/abs/2106.14888

[22] S. Strohkorb Sebo, M. Traeger, M. Jung and B. Scassellati, "The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human-robot teams," *ACM/IEEE Int. Conf. Human-Robot Interact*., pp. 178-186, 2018.

[23] D. P. Van der Hoorn, A. Neerincx and M. M. de Graaf, "I think you are doing a bad job! The Effect of Blame Attribution by a Robot in Human-Robot Collaboration," *ACM/IEEE Int. Conf. Human-Robot Interact*., pp. 140-148, 2021.

[24] S. Vinanzi, A. Cangelosi and C. Goerick, "The collaborative mind: intention reading and trust in human-robot interaction," *iScience*, vol. 24, 102130, 2021.

[25] S. Vinanzi, M. Patacchiola, A. Chella and A. Cangelosi, "Would a robot trust you? Developmental robotics model of trust and theory of mind," *Philos. Trans. R. Soc. B*, vol. 374, 20180032, 2019.

[26] K. Man and A. Damasio, "Homeostasis and soft robotics in the design of feeling machines," *Nat. Mach. Intell.*, vol. 1, pp. 446-452, 2019.

[27] S. Ososky, T. Sanders, F. Jentsch, P. Hancock and J. Y Chen, "Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems," *Proc. SPIE*, vol. 9084, pp. 517-527, 2014.

.

Considerations about Social Norms Compliance in a Shared Elevator Scenario

Danilo Gallo, Shreepriya Shreepriya, Tommaso Colombino, Maria Antonietta Grasso, and Cecile Boulard

Abstract—

In this paper, we present our ongoing research on socially acceptable robot navigation for an indoor elevator sharing scenario. We highlight the current challenge of designing interactions for a robot behavior, both effective in accomplishing tasks but not intrusive or at risk of breakdown. We discuss the advantages and limitations of modeling these behaviors based on a full human-like approach. In particular, we discuss the risk that a full human-like approach presents of creating the illusion of social competence. It has been observed that this illusion often leads to breakdowns when the technology is faced with complex and potentially ambiguous social situations. We propose the principle of "machine-like yet human-friendly" behavior to address the risks of the completely human mimicking approach. We believe that this approach can provide more understandable and less disruptive behaviors for routine integration into human spaces. We conclude by discussing the need for a multi-layer experiment set up to evaluate and validate this approach.

Keywords: Social norms; Robot navigation; Robot legibility.

I. INTRODUCTION

Questions around what constitutes socially acceptable behavior for autonomous agents are not new to the HRI community [1]. The safe and harmonious deployment of robots in public spaces requires their social behavior to be understandable and predictable [2]. The social norms that constitute the fabric of human interactions can be very informative to model the robot behaviors and blend them into a social setting. These norms are potentially useful for modeling machine behavior in an understandable way and adapting it to the context [3].

In this paper, we present how our current work in the area of social navigation has benefited from a fine-grained understanding of the social norms that are present during the activity of taking a shared elevator. In this work, we have focused on indoor scenarios using a robot platform developed by our organization. These robots are capable of navigating autonomously and carrying objects (Figure 1). These scenarios include delivering parcels and food orders in the context of a large office building. The robots we design have some dedicated infrastructure (such as dedicated elevators in a specific office building). However, they also have to be able to utilize shared infrastructure and spaces with humans at certain times. The robots are controlled by a centralized processing robot brain to keep each robot relatively simple and modular, reducing unit costs. In these office scenarios, we focus our investigation on aspects of social robot navigation involving the use of shared elevators with humans.

Previous research has focused on the technical development of robots able to autonomously operate and ride elevators with humans, such as [4] and [5]. However, we are not aware of any prior art that explicitly investigates the appropriate navigation behavior a robot should display considering social norms and human preferences for that specific context. Indeed, this context presents a range of challenges that go beyond the interaction with the infrastructure, such as negotiation of priorities or movement and coordination in reduced spaces.

Figure 1. The NAVER robotic platform.

When investigating the broader domain of social robot navigation, we have come across several approaches and specific contributions to enable a socially acceptable interaction. However, we encountered the lack of a common approach that characterizes the aimed-for human-friendliness that we could adopt in our work. This led to the definition of our own approach, which is design-driven and grounded in a fine-grained understanding of the social interactions at play in the context of interest. In doing so, we also drew inspiration from the body of work that has highlighted the challenges involved in making complex AI systems and decision making transparent to lay users, when encountering these systems in shared spaces and infrastructure [6][7], e.g., in the context of Autonomous Vehicles.

Authors are with Naver Labs Europe, France.
Name.surname@naverlabs.com

II. SOCIAL NAVIGATION WHILE TAKING A SHARED ELEVATOR

As mentioned in the introduction, our research focuses on indoor office scenarios. These scenarios involve robots sharing elevators with employees and visitors at our corporate headquarters. The robots deliver parcels and food orders in the context of a large office building, where they display autonomous behaviors and are controlled by a dedicated infrastructure. This dedicated infrastructure foresees the use of robot-only elevators and moments in which robots would need to be able to use shared elevators, e.g., when the workload is high. While we have seen commercial robots navigate in and out of elevators, there is little focus on the issues of proxemics and cultural preferences [8] in this context. Designing for such issues requires a distinction between what might be characterized and reduced to rules easily built into the robots' navigation behaviors and more nuanced matters of elevator social norms that would require the robots to identify things like human gaze and posture and *read* their interactional meaning and valence. While in our corporate headquarters, the centralized robot "brain" can operate the elevators for the robots, the scenario of robots using the elevators also points to the need for more complex interactions designed to enhance the robots' flexibility. Such interactions include robots requesting assistance operating elevators that might not be explicitly designed for their use and not integrated into the robotic platform's infrastructure. This need is similar to those investigated in the context of collaborative robotics [4] aimed at creating robots capable of compensating for their physical (ability to manipulate the environment) or perceptual limitations by eliciting human assistance [9].

We are aware of existing commercial robots using elevators, which rely mainly on speech interfaces, i.e., the robot announcing its intention to enter the elevator and declaring where they will position themselves [10]. While this can be an effective strategy, it clearly places the burden of making the interaction work on the people sharing the elevator with the robot. It may even be socially acceptable if the interactions are occasional (as might be the case, for example, with delivery robots in a hotel where any given guest might encounter the robot once during their stay). However, in an office environment with service robots performing routine tasks, encounters with robots in elevators are likely to be a daily occurrence for people working in the building, which means that negotiating the use of elevators through loud verbal announcements could quickly become tiresome.

A different and far more ambitious technology and interaction design paradigm might be to develop a platform capable of reading non-verbal behavior and the social context and its norms. This understanding could enable more subtle interactions, with robots that treat people as agents occupying a shared space with rules to follow for things like order of service and priority, rather than just obstacles to be avoided. This, of course, presents a substantial challenge as there is a semantic gap to be bridged between detecting things like posture or predicting movement (intentionality) and making dynamic and contextually appropriate decisions in what is (most of the time, *but* not always) for us a straightforward, but quite nuanced social interaction.

A. *Understanding the activity*

In our approach to the elevator scenario, we first studied what we might call the *practice* of taking the elevator. While it is in many ways a straightforward accomplishment (arguably more than driving a car in traffic), it is also constituted of practices that are methodical and accountable, with normative components and nuanced, often non-verbal use of space and resources. We analyzed approximately 16 hours of video data gathered by placing a video camera in the elevator lobby of one of our company's research labs. We adopted an ethnomethodological analytic orientation [11] to understand the specific practices of waiting for, entering, and exiting an elevator. We do not go into all the findings in detail here, but for the purpose of this discussion, we focus on how the order of service is managed (who enters the elevator first when multiple people are waiting). Our observations reveal that there is a general first-come-first-served principle that is applied, but it is a *weak* one. People do not form proper queues, especially where multiple elevators are linked to a single call function. In this common scenario, people often drift towards the lobby's center and only position themselves clearly in front of a door when the elevator lights indicate it will be the next available elevator. However, the elevator lights are not entirely reliable as the elevator status may change, and the next available elevator may, in fact, be at the other end of the lobby. In such cases, people moving back and forth across the lobby (*chasing* the next available elevator) may lose or have to renegotiate their priority in the order of service. Additionally, groups of people wait in the lobby with the intention of taking the elevator together. These groups also have *weak* form attributes regarding closeness and body orientation, making it ambiguous sometimes even to a human observer. These groups might stand in front of the elevator, even calling for it while waiting for another member (Figure 2). The difference between a fully formed group and one under composition is related to elements like distance and body orientation. However, also in this case, several ambiguous situations have been observed in our videos.

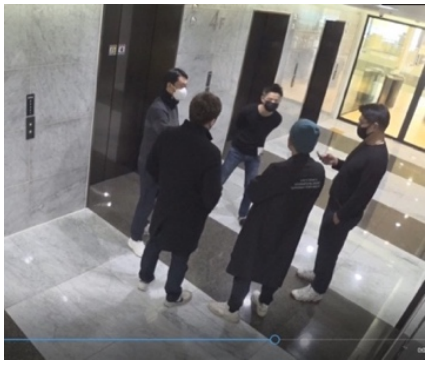


Figure 2. Waiting for another member to join in front of the elevator.

This seemingly obvious and easily accomplished behavior would present serious challenges if we wanted a socially competent robot to fully understand and mimic it. For example, what [12] describes as *the order of waiting* is constituted of both ordered and disordered formations, with demarcations and affiliations that are constantly produced and renegotiated. Even if we had computer vision technology that was capable of

reliably and dynamically detecting things like posture, orientation, distance and displacement with respect to other people and elevator doors [13], gaze and facial expression, and computing the semantically and situationally appropriate reading of the situation, the nuances of the context would still be hard to capture, as they are for a human being at times. Additionally, we question how comfortable people would be with robots that fully mimic human-like behaviors and how these behaviors might contribute to an *illusion* of social competence.

One of the problems that concern us here and that we think is of interest to the HRI community is that, as [14] observed many years ago, the breakdowns in the interactions between technology and its users were often instantiated by what could be described as the *illusion* of social competence. Therefore, the design challenge we are exposed to is how an agent can effectively navigate shared spaces with people, focusing on safety and minimal disruption, but without necessarily being burdened with the normative expectations of being perceived as a fully human-like socially competent agent. This challenge resonates with discussions that have been done in a broader sense on the use (and extent of use) of anthropomorphic elements for robotic visual and behavioral elements' design [15].

With these considerations, we are experimenting with an intermediate option in which robots both exploit an understanding of the social situation and retain its representation as a tool to exhibit "machine-like yet human-friendly behaviors." Previous research [16] has defined machine-like behavior as the behavior that exploits machines' characteristics like sensors that humans do not have. For instance, autonomous vehicles (AVs) can know the position of other AVs without seeing them and act accordingly. Human-like behaviors are defined as the typical behaviors exhibited by humans based on their social understanding of other actors and the context. In our machine-like yet human-friendly behavior paradigm, we mix machine and human-like behaviors. We adopt human behavior elements that allow the robot to demonstrate the necessary level of social understanding that avoids disrupting the activity while ensuring task completion (e.g., position and direction with respect to the elevator door). We further incorporate machine-specific elements that convey the robot's role as subordinate entities with limited social understanding (e.g., unidirectional communication of intent and priority to humans as much as possible). We hypothesize that this approach would prevent the illusion problem introduced by taking a full human-like approach.

Our approach is reflected in a number of design choices aimed at creating robots with elegant, clear, and direct interaction mechanisms that encourage users to, for example, limit their needs of engaging with the service robots to what is part of their tasks and within their scope, and not beyond. This approach can limit potential breakdowns while ensuring that the robot does not disrupt the routine activity, i.e., taking the elevator. Indeed, this activity should be designed to become a non-experience, something people do without thinking and without consciously experiencing the interactions [17].

B. The reality of social competence

The notions of social competence come into play when technology designers use generic interaction metaphors like *human-like* or *pet-like* that give connotations that burden the agent individual and culturally dependent expectations. Moreover, as mentioned, the complicated technology computations required to model these socially nuanced and potentially ambiguous situations contextually are still a limiting factor to consider.

TABLE I. DEFINING OUR PROPOSED APPROACH WITHIN THE SPECTRUM OF MACHINE-LIKE AND HUMAN-LIKE BEHAVIORS IN THE SHARED ELEVATOR CONTEXT

	Machine-like	Machine-like yet Human-friendly	Human-like
Social awareness	No. The robot is not able to differentiate humans from other obstacles.	Some The robot can detect humans and is **aware of people entering and exiting** to decide its actions.	Yes The robot detects humans and their intentions. It adopts queuing as done by humans and moves of position according to an understanding of situations.
Communication of intent	Non-verbal, e.g., sounds. The robot only gives information for consumption.	The robot uses subtle **non-verbal interface elements** to broadcast intent, which are deliberate design choices for information consumption only.	Verbal and non-verbal like gaze, body posture, etc. The robot and the humans acknowledge and exchange information.
Movement and position	The robot positions itself in front of the door and always enters first.	The robot **takes a fixed waiting position and gives priority**, except in urgency.	Mimics humans with queuing and position adjustments.

To explain, we can take an example of waiting for an elevator scenario in which the robot is designed to have *human-like* behaviors. In this scenario, the robot should detect if a crowd of people is actually waiting for the elevator to act accordingly. This would require understanding if the group is waiting for the elevator or another group member to join them. In our video observations, we identify human pose and distance to the elevator as indicators of intention. However, we recognize that these elements are not enough to identify with confidence the group intention, even for a human evaluator. In order to disambiguate the situation, the robot would then need to initiate interactions with the group or take guesses to queue behind the unstructured group, potentially (and unnecessarily) delaying the service accomplishment. We further hypothesize that people

would not feel comfortable with a robot that moves like humans, making continuous adjustments rather than taking a designated waiting position. Hence, we distance ourselves from following the strict approach of fully exploiting understanding and adherence to the elevator social norms in our work and establish the notion of *machine-like yet human-friendly* interaction behaviors. We define such behaviors as ones that respect human and social considerations particularly relevant to this activity and social context without explicitly mimicking human behaviors. As previously described, one of our findings exposed the fluidity of queuing and taking an elevator in a multi-elevator setting linked to common call buttons. In this case, a *human-like* behavior of changing queues and moving from elevator to elevator, as a human would do, would be relatable yet unpredictable, annoying, and potentially hazardous. Considering this, we designed specific actions to be taken by the robot based on only some elements of the elevator human etiquette while also putting in place specific and clear robot behaviors (TABLE I.).

C. *Designing the robot behavior*

The design of the activity is broken down into the following stages: waiting, entering, riding, and exiting the elevator. At the same time, each stage is divided into several smaller actions performed by the robot, during which the robot will display a series of communication states. Each state aims to convey a specific intent by using different communication modalities (i.e., sound, light, displays, projection, and anticipatory movements) and combinations of them. For this paper, we only describe the waiting and entry stages of the activity, along with the specific actions involved.

Waiting for the elevator:

a) Calls elevator

- The robot navigates to the elevator hall (it might encounter people waiting for the elevator in an unstructured queue or be the first to arrive).
- The robot calls the elevator and receives the information about the elevator that will arrive next. Elevator indicators should reflect this information.

b) Navigate to Waiting Position

- The robot navigates to the fixed waiting position (Figure 3). It commits to the corresponding elevator, even if the situation changes and another elevator arrives first (unlike a human-like behavior).
- If it detects humans within 46 cm [18], it tries to go around them or starts a Request state to ask permission to pass. We should point out that this distance might need to change in different cultures or contexts (e.g., within the elevator, where there is limited space). If the robot detects obstacles within 10 cm is tries to go around them.

c) Robot waits

- Once it reaches the Waiting Position, the robot displays it is in Waiting state.
- Regardless of people's movements around it, the robot remains in that place to avoid disrupting them with small position adjustments (unlike a human-like behavior).

Entering the elevator:

d) Elevator arrives

- The robot detects people exiting the elevator and waits for them to exit.

e) The robot lets people enter first

- After everyone left the elevator, the robot detects people are entering.
- If people are entering, it remains in place and triggers a Yielding state and lets people waiting to enter, regardless of the order of arrival (unlike a human-like behavior).
- Once everyone has entered the elevator, the robot triggers an In-Motion state.

f) Robot enters first

- In certain cases, if the robot's task (such as delivering hot coffee or food) is to be completed within a certain timeframe, the robot triggers an Urgent state right after people stop exiting the elevator. The Urgent state communicates the intention to enter first (even if it detects people are entering), and the robot enters the elevator. This behavior should be triggered only when a new (next) elevator arrives to avoid confusing people in the process of entering the elevator.

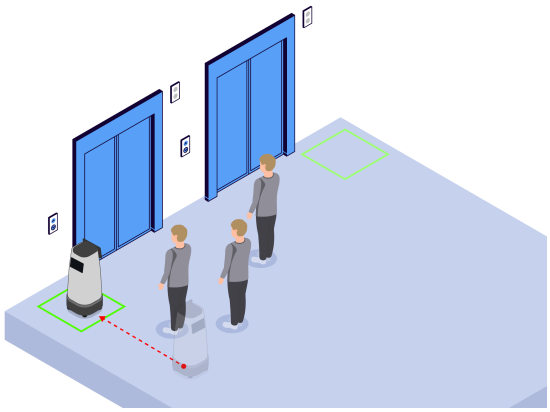


Figure 3. The robot takes a designated waiting position regardless of the position of the humans waiting for the elevator.

D. *Considerations regarding experiment design*

The social navigation behavior that we have designed based on the above considerations has resulted in the following hypothesis that we will assess with several user experiments, each tailored to the specific level of design under consideration.

- In certain cases (especially related to positioning and movements), people will understand and prefer *machine-like yet human-friendly* robot **actions** rather than those mimicking humans, such as the subtle movements exhibited by people taking elevators.

- Our proposed robot's behaviors for **communicating intent** (calling the elevator, waiting, entering) will be as understandable and less intrusive than fully human-like ones, for the first use as well as for an extended period.

In our elevator-taking scenario, the actions of *waiting* and *entering* have to be tested under the conditions of *machine-like yet human-friendly* and *human-like* to provide proof for our first hypothesis.

While to test the second hypothesis, we need to test the understandability of different alternate communication modalities used by the robot against the baseline of verbal interaction modality commonly used in existing service robots.

E. The layered testing approach

Our design proposal combines a range of navigation policies and different communication modalities to convey the robot's intent. To effectively validate the impact of these different elements, we need first to evaluate them independently to conduct then tests that integrate them into a comprehensive service. For this reason, we decided to adopt a layered testing approach consisting of three steps: online experiments, in-situ experiments, and naturalistic observations of the robot working in context. We also add naturalistic observations to our testing toolkit for a better ecological validity of our robotic service.

Online experiment. While online experiments may have limitations related to participants' profile, level of engagement, and quality of results, they can provide preliminary feedback without the constraints of participants' time and exhaustion. We will compare several alternatives for each defined action through a perceptual, low fidelity experiment in which videos of the situations identified around elevators with each condition (alternatives for specific features, e.g., position, interfaces, etc.) will be shown to participants. We will collect and analyze objective and subjective measurements of participants' understanding and preference for each condition.

In-situ experiment. In-situ evaluations are seen to be more valid, especially for performance-related metrics like response time, completion time, or tasks completed. We will evaluate the pre-selected alternate behaviors/features (from the results of the online experiment) in a realistic set-up. Participants will be requested to perform the actions related to using elevators while sharing the waiting space with a robot. We will collect objective measurements of response time and task completion and subjective measurements of understanding and preference.

Naturalistic observation. Evaluations conducted in lab-based controlled settings often lack ecological validity. To counter this, experimenters rely on experimental realism or simulating the context of use. While this is effective, it can still produce bias as the participants are recruited and briefed about the experiment [19], and the introduced novelty can also influence them. Hence, naturalistic observations can counter these effects. We will deploy the robot in the wild with the selected features from the in-situ test. We will capture and analyze the reaction of passers-by and people taking a shared elevator for first encounters and an extended period. Through video analysis, the person or group's understanding and preference will be analyzed.

III. Conclusion

The real-world deployment of autonomous robots presents several complexities. In an indoor environment, robots sharing an elevator with people can be considered one of the scenarios that will soon be a reality. To design for near-future deployable robots, the illusion of social competence of the robots must be carefully managed, as well the robot harmoniously blending with the social norms of a setting. Indeed, the technology does not provide fully reliable solutions to understand and model the complex and potentially ambiguous social situations we observed. Moreover, it is still an open question if a fully human-like behavior is desirable in autonomous agents. Hence, we propose a *machine-like* yet *human-friendly* approach to the design of robot navigation behaviors and a layered testing approach. Through these experiments, we aim to validate our assumption that *machine-like* yet *human-friendly* interactions are preferable for the robot's social behaviour.

References

[1] R. Simmons, R., J. Forlizzi and R. Kirby. 2010. "Social robot navigation."

[2] C. Lichtenthäler, T. Lorenzy and A. Kirsch. 2012. "Influence of legibility on perceived safety in a virtual human-robot path crossing task." IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication, 2012, pp. 676-681, doi: 10.1109/ROMAN.2012.6343829.

[3] A. Sandygulova, M. Dragone, and G. M. P. O'Hare. 2016. "PRIveT- A portable ubiquitous robotics testbed for adaptive human-robot interaction". Journal of Ambient Intelligence and Smart Environments, 8(1), 5-19. https://doi.org/10.3233/AIS-150356

[4] M. Veloso, B. Joydeep, B. Coltin, S. Rosenthal, T. Kollar, C. Mericli, M. Samadi, S. Brandao, and R. Ventura. 2012. "CoBots: Collaborative Robots Servicing Multi-Floor Buildings". In Proceedings of IROS'12, the IEEE/RSJ International Conference on Intelligent Robots and Systems, Vilamoura, Portugal.

[5] J.-G. Kang, S.-Y. An, and S.Y. Oh. 2007. "Navigation strategy for the service robot in the elevator environment". ICCAS 2007 - International Conference on Control, Automation and Systems. 1092 - 1097. 10.1109/ICCAS.2007.4407062.

[6] B. Brown and E. Laurier. 2017. "The trouble with autopilots: Assisted and autonomous driving on the social road". In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, United States.

[7] E. Vinkhuyzen and M. Cefkin. 2016. "Developing Socially Acceptable Autonomous Vehicles. 2016 Ethnographic Praxis in Industry Conference Proceedings." https://www.epicpeople.org

[8] M. Joosse, M. Lohse, and V. Evers. 2014. "Lost in Proxemics: Spatial Behavior for Cross-Cultural HRI". 2014. HRI'14 Workshop on Culture Aware Robotics, Bielefeld, Germany.

[9] J. Liebner, A. Scheidig, and H.-M. Gross. 2019. "Now I Need Help! Passing Doors and Using Elevators as an Assistance Requiring Robot," in Social Robotics, vol. 11876, M. A. Salichs, S. S. Ge, E. I. Barakova, J.-J. Cabibihan, A. R. Wagner, Á. Castro-González, and H. He, Eds. Cham: Springer International Publishing, 2019, pp. 527–537. doi: 10.1007/978-3-030-35888-4_49.

[10] R. Ichinose, I. Takeuchi, T. Teramoto. 2015. "Elevator System that Autonomous Mobile Robot Takes Together with Person." US Patent US 8,958,910 B2. Feb. 17, 2015.

[11] H. Garfinkel. 1967. Studies in Ethnomethodology. Polity Press.

[12] R. Ayaß. 2020. "Doing Waiting: An Ethnomethodological Analysis." Journal of Contemporary Ethnography, Vol 49(4) 419-455.

[13] A. Vega-Magro, L. Manso, P. Bustos, P. Nunez and D. G. Macharet. 2017. "Socially Acceptable Robot Navigation over Groups of People." 26th

IEEE Internation Symposium on Robot and Human Interactive Communication. Lisbon, Portugal.

[14] L. Suchman. 2007. “Human-Machine Reconfigurations: Plans and Situated Actions.” Cambridge: Cambridge University Press.

[15] M. Boden. 2006. Robots and anthropomorphism. AAAI Workshop - Technical Report. 69-74.

[16] L. Oliveira, K. Proctor, C. G. Burns, and S. Birrell. 2019. "Driving Style: How Should an Automated Vehicle Behave?" Information 10, no. 6: 219. https://doi.org/10.3390/info10060219

[17] R. Rousi. 2013. “The experience of no experience Elevator UX and the role of unconscious experience”. In Proceedings of International Conference on Making Sense of Converging Media - AcademicMindTrek ’13, ACM Press, Tampere, Finland, 289–292. DOI:https://doi.org/10.1145/2523429.2523455

[18] E. T. Hall. 1966. “The hidden dimension: man’s use of space in public and private.” The Bodley Head Ltd, London.

[19] T. Kruse, A. K. Pandey, R. Alami, A. Kirsch. 2013. “Human-aware robot navigation: A survey“. Robotics and Autonomous Systems, Volume 61, Issue 12, 2013, Pages 1726-1743, ISSN 0921-8890, https://doi.org/10.1016/j.robot.2013.05.007.

Learning Social Navigation from Demonstrations with Deep Neural Networks

Yigit Yildirim[1] and Emre Ugur[2]

***Abstract*— Traditional path planning techniques treat humans as obstacles. This has changed since robots started to enter human environments. On modern robots, social navigation has become an important aspect of navigation systems. To use learning-based techniques to achieve social navigation, a powerful framework that is capable of representing complex functions with as few data as possible is required. In this study, we benefited from recent advances in deep learning at both global and local planning levels to achieve human-aware navigation on a simulated robot. Two distinct deep models are trained with respective objectives: one for global planning and one for local planning. These models are then employed in the simulated robot. In the end, it has been shown that our model can successfully carry out both global and local planning tasks. We have shown that our system could generate paths that successfully reach targets while avoiding obstacles with better performance compared to feed-forward neural networks.**

I. INTRODUCTION

Mobile robot navigation has been studied for decades. Many notable techniques have been proposed in this area over the years, [1], [2], [3]. These approaches have prioritized the safety and the robustness features, i.e. the principal driving factor behind the development in this field has been the collision avoidance [4]. On the other hand, as humans start to share their environments with robots, new requirements for mobile robot navigation have emerged.

In [5], physical and mental aspects of the safety are separately evaluated. This separation reveals the need to question the psychological efficiency of navigation systems of mobile robots. Keeping in mind the assumption that humans prefer to interact with machines in the same way that they interact with other people, in order to achieve a natural integration to the environments populated by people, mobile robots must be developed to be not only safe but also comprehensible.

Broadly speaking, human-aware navigation corresponds to the navigation that complies with the social rules of the people. In their own environments, humans tend to work cooperatively to realize social navigation. Then, it is only natural to imitate this behavior on the robots to achieve socially-acceptable navigation. However, imitating people introduces new constraints to be satisfied by the navigation systems of robots.

These constraints have been addressed in many studies in the literature. Essentially, these studies can be divided into two categories: manually-encoded controllers and learning-based ones. One of the notable studies of the first category is the Social Force Model (SFM) [6]. Based on the behavioral techniques from social sciences, SFM suggests that pedestrians move under the effect of certain abstract forces, just like the particles in an electrical field. While the navigational goal attracts the pedestrian, obstacles and other people exert repulsive forces. Despite its wide application[7], [8], [9], some researchers state that not being based on the statistical data is a weakness of the model [10].

To create statistics-based socially compliant navigation frameworks, a large number of machine learning algorithms have been employed. One of the popular algorithms is Inverse Reinforcement Learning (IRL) [11], [12], [13]. Given the perfect expert demonstrations, IRL tries to identify the underlying reward structure, which in turn can be used by any Reinforcement Learning (RL) algorithm to create a human-aware navigation policy. Even though the justification of the unfixed reward function is appealing, the features that shape the reward function are assumed to be known, which is considered as a strong assumption [14]. Generally in this domain, feature engineering leads to strong assumptions. This problem can be solved by extracting the social behaviors and navigation strategies of pedestrians directly from the data. This is challenging because the controller needs to be complex enough to capture the non-linearities in the data.

To address this issue in social navigation domain, deep learning techniques have been used. In [15], Deep Reinforcement Learning is used to obtain a socially plausible navigation policy. As in other RL approaches, this procedure relies on a predefined reward which is difficult to obtain. Imitation Learning skips the reward extraction and tries to learn policies directly from the data. In [16] and [17], Generative Adversarial Networks are used for this purpose. These approaches are complex enough to overcome the aforementioned issues. However, these models need too much data to be trained [18]. On the other hand, the preferred system needs to learn from a small dataset and to generalize to novel configurations.

Moreover, the majority of the studies on this domain target only the local controller of the robot as it is the part that creates motion commands to drive the robot. However, using only the local controller makes the robot vulnerable to local minima [19]. Today, typical robotic navigation systems adopt the two-layered hierarchical approach for path planning tasks. Given a map of the environment, a robot firstly calculates a trajectory in the so-called *global planning* phase. Then, the robot follows the computed trajectory with

[1]Yigit Yildirim is with Computer Engineering Department, Bogazici University, Istanbul, Turkey yigit.yildirim@boun.edu.tr

[2]Emre Ugur is with Computer Engineering Department, Bogazici University, Istanbul, Turkey emre.ugur@boun.edu.tr

a controller in the so-called *local planning* phase.

In this paper, we use Conditional Neural Processes (CNPs) [20] in order to address the issues mentioned above in both global and local planning phases. CNPs can be modified to generate complete trajectories to replace the global planner. Also, they can create goal-directed behavior while actively avoiding obstacles. This characteristic makes it a candidate for the local planner, as well. CNPs extract the prior knowledge directly from the training data by sampling observations from it, and uses it to predict a conditional distribution over any other target points. CNPs can learn complex temporal relations in connection with external parameters and goals. In this paper, we present the initial results of our system. Upon successful preliminary results with this conceptual model, we aim to extend this work to integrate our path planning system into an actual robot in another study.

II. Related Work

Traditionally, approaches to solve the path planning problem can be divided into two categories based on the environmental knowledge they use: deliberate and reactive. Deliberate planners exploit the environmental knowledge by means of static maps and calculate the robot's trajectory before execution. On the other hand, reactive planners rely on sensory information to deal with local parts of the environment. Either approach has its advantages and drawbacks. Hence, the evolution of the path planning approaches leads to the combination of these two approaches. Hybrid frameworks have been the typical approach for many years, as explained in [21].

In the following, we elaborate on this conventional framework's building blocks and the social navigation concept.

A. Hierarchical Path Planning

The standard hybrid path planning framework combines the strengths of deliberate and reactive planners. It consists of a two-phased procedure in a hierarchical manner; global planning is for the deliberation and local planning for the reactivity.

1) Global Path Planner: In the first phase of a standard hierarchical path planning pipeline, a global planning procedure is applied. On the static map of the environment, the function of a global planner is to generate a path from the starting position to the destination. Conventionally, many graph search algorithms have been applied to calculate the trajectory between initial and goal configurations, the most popular being A* explained in [22]. For a more complete list of global planning approaches, see [23].

The global planning itself is not sufficient to navigate the robot between two points. Local planning is needed to create velocity commands that handle the cases with new or dynamic obstacles.

2) Local Path Planner: In order to realize computed trajectories, the local planning procedures are used in the second phase of hierarchical path planning. The most prominent objective of the local planner is to generate velocity commands so that the robot can follow the computed trajectory. In addition, by using the sensory information about the robot's surroundings, it is the local planner's duty to avoid obstacles. There are many local planning algorithms in the literature, such as [24], [4], [25], [26], [27], [28]. For a more complete list, see [29].

On the other hand, despite being quite safe, these traditional controllers take no account of social norms. They consider people as obstacles to be avoided. Recent attempts to create local controllers that consider these norms has paved the way for social robot navigation.

B. Social Navigation

According to [30], the benefits of social navigation are threefold: it increases the comfort of the people around the robot, it improves the naturalness of the robotic platform and it also enhances the sociability of the robot. Furthermore, in [5], physical and mental aspects of safety are separately evaluated. For us, this separation reveals the need to question the psychological efficiency of navigation systems of mobile robots.

The concept of social navigation lies in the intersection of two concepts: navigation and human-robot interaction. It describes improving the navigation of the robot to enhance its comprehensibility by the humans around. Figure 1 is rather self explanatory. On the left, we see a robot with a perfectly safe navigation plan. In contrast, although non-optimal, the planned path on the right is socially compliant.

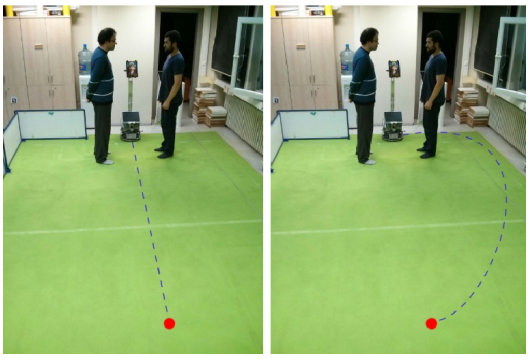


Fig. 1: Comparison between regular and social navigation.

III. Method

In this work, we address two parts of the hierarchical path planning individually and show that the capabilities of the model we propose can handle both global and local planning. We suggest employing a variant of Conditional Neural Processes (CNPs) for both of them separately.

CNP is a powerful deep learning framework, which is inspired by the flexibility of stochastic processes, but organized as neural networks and trained with gradient descent [20]. Since its emergence, CNPs and variants have been successfully applied in several robot learning problems [31], [32], [33]. Instead of outputting a single value, CNP learns a Gaussian distribution over the demonstrated trajectories. The

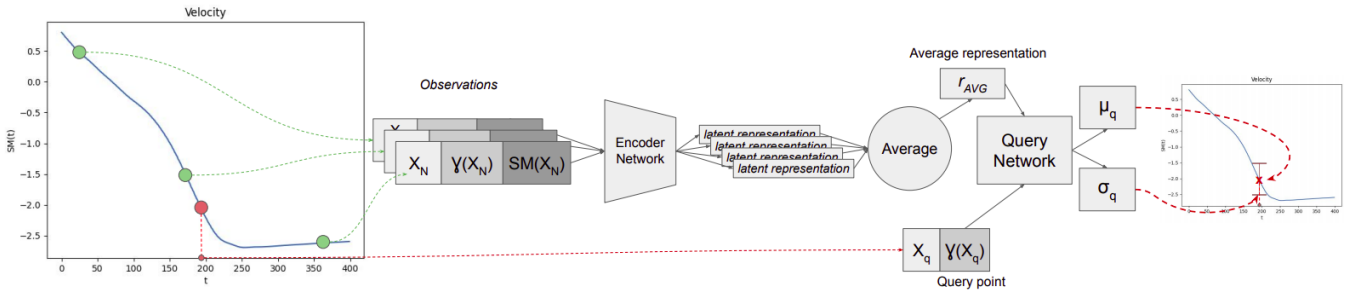


Fig. 2: General layout of the training phase of our model.

set D, representing all demonstrations is defined as follows: $D = \{D_i\}_{i=0}^{N}$, where each D_i is a trajectory of a number of points in a high-dimensional space. Essentially, $D_i = (X_t, \gamma(X_t), SM(X_t))_{t=0}^{\tau}$, where X is the state variable, $\gamma(X)$ is a function representing task parameters and $SM(X)$ is the sensorimotor function to be learned. The encoder network produces a latent representation for each trajectory and these representations are passed through an averaging operation to create a compact representation r_{AVG} for the task at hand. Subsequently, X_q, $\gamma(X_q)$ and r_{AVG} are fed to the *Query Network* to produce an estimate for $SM(X_q)$. μ_q and σ_q respectively represent the estimated mean and the variance. Figure 2 shows the overall model.

The model consists of an encoder network which outputs latent representations by using the sampled points on the demonstrated trajectories. These representations in the latent space are then averaged to come up with a compact representation of the trajectory. At query time, this compact representation is concatenated with the target point and the resulting vector is fed to the query network to generate the estimated sensorimotor response of the model.

1) Global Planning: One of the most powerful aspects of the CNPs approach is its ability to generate complete trajectories. Upon training the encoder and query networks, target points can be simultaneously processed from the starting point to the end to create an entire trajectory. This ability can be exploited to create global plans in the first phase of a hierarchical path planning procedure.

2) Local Planning: We also benefit from CNPs in reactively responding to the changes in its domain. With this, we substitute the local planning module of the hierarchical path planners with local CNPs. This requires sensory input to be processed by the CNP as task parameters. In the current study, high-level parameters such as distance to the obstacles or relative position to the goal point are used as input to the local CNPs. It was shown that CNPs can efficiently handle low-level and high-dimensional input as well, as shown in [34].

IV. Experiments and Results

A. Environment

Our system was verified in CoppeliaSim simulation environment [35] that includes an omnidirectional robot platform (Robotino [36]). The Social Force Model, described in [6], is implemented as the local controller of the robot to gather demonstration trajectories. With the assumption that it generates socially plausible trajectories, 1000 trajectories with randomly different starting, goal and obstacle poses are recorded. Single, multiple, stationary and dynamic objects are placed at random positions in each trial. The data collection process is shown in Figure 3.

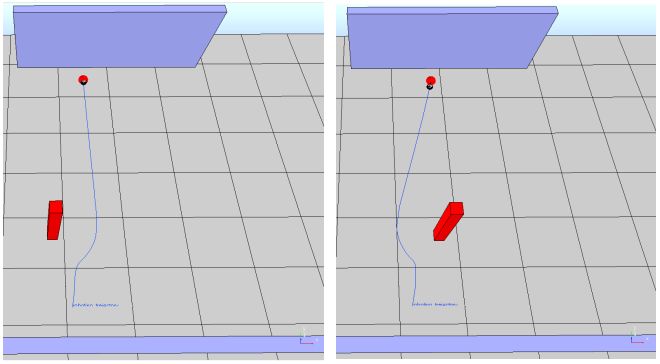


Fig. 3: Data collection on the simulation. The motion trajectory is shown with blue line.

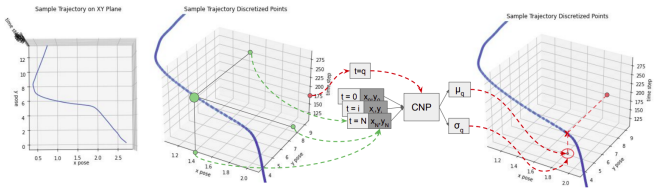
B. Global Planner

To show the path planning capability of our method, the model is fed with the entire trajectories of positions of the robot and trained on these demonstrations. The representation of the data is as follows:

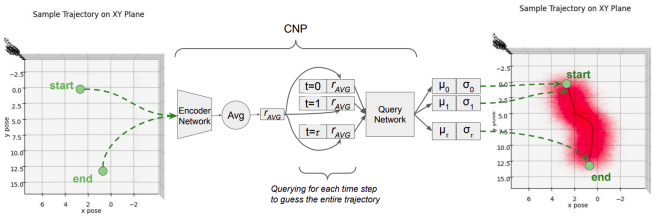
$$X = time\_step$$
$$\gamma(X) = (start\_x, start\_y, goal\_x, goal\_y, obs\_x, obs\_y)$$
$$SM(X) = (position\_x, position\_y),$$

where obs_x and obs_y refers to the obstacle's x and y positions. Fig. 4a illustrates the training phase and Fig. 4b shows how the entire path is queried.

To show the strength of CNPs over standard neural networks, we compare their performance on the trajectory planning task. For this purpose, we implemented a 5-layered standard feed-forward neural network and trained it on the same dataset of 1000 trajectories. The comparison of their performances on a global planning task is given in Figure 5. This result shows that while a feed-forward neural



(a) Training the network with a randomly chosen demonstration trajectory.



(b) Generating a global path in test phase.

Fig. 4: CNP as the global planner.

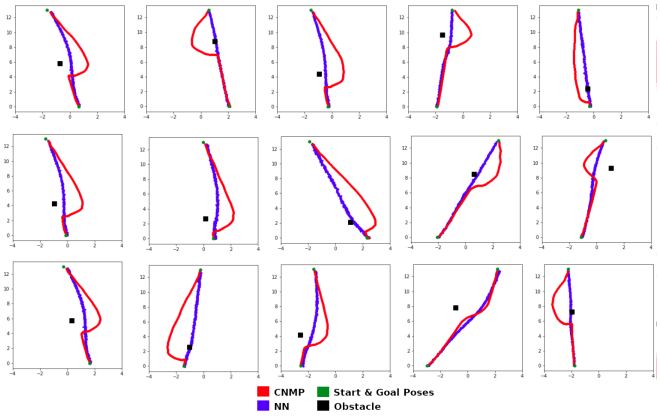


Fig. 5: Comparison between our global planner network (CNP) and a 5-layered feed-forward neural network (NN) on global planning in sample environments.

network cannot generate global paths that avoid obstacles, our system can. We believe that this is due to the capability of our system to learn multiple-modes of operations. Standard feed-forward networks, given demonstration paths that avoid obstacles from different sides, probably interpolates these paths; whereas our system can learn to generate trajectories from both sides.

C. Local Planner

From the local perspective, the input parameters of our local network are distance-to-goal, distance-to-obstacle and velocity commands. Here, the formulation of the problem is as follows:

$$X = (distance\_to\_goal\_x, distance\_to\_goal\_y)$$
$$\gamma(X) = (distance\_to\_obs\_x, distance\_to\_obs\_y)$$
$$SM(X) = (velocity\_x, velocity\_y)$$

Note this time that, we do not use a linearly increasing phase variable, as we did in the case of global planning. Conditioned on the starting and destination poses, the use of the task parameter $\gamma(X)$ gave the model the ability to reactively change the velocity commands with respect to changing obstacle positions.

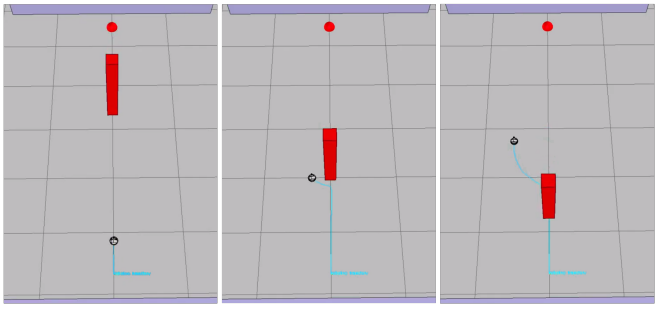


Fig. 6: Robot is avoiding from a vertically moving obstacle.

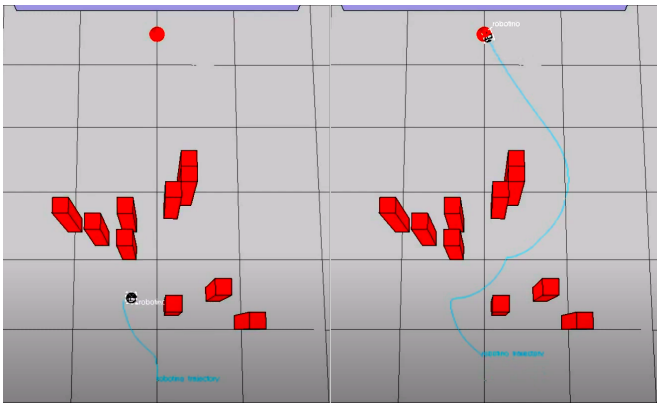


Fig. 7: Robot is passing through several stationary obstacles.

The resulting local planner is shown to work on several different configurations, as shown in Figures 6 and 7. Since our local planner is trained on the trajectories created by SFM, we believe that the policy it learned imitates SFM's behavior. Further comparison is needed to support this claim.

V. Limitations and Future Work

In this study, the preliminary results of our framework which is a hierarchical framework that is built on top of CNPs is presented. We showed that our model can generate reasonable paths at both global and local levels while avoiding obstacles. This work needs to be extended with a thorough statistical analysis comparing with strong baselines in successful social-aware navigation tasks. As a part of this endeavour, we plan to train our models on actual human data. Thus, we would elude the critics of SFM and prove that our model could work with real human data. Furthermore, most importantly, we plan to transfer and verify learned models in real robots.

Another direction of research to extend this work is to incorporate detectors that discover groups of people from

raw sensory information. Human trajectory prediction can also be added to create smoother paths during navigation. For this purpose, graph neural networks [37] that represent the world as nodes and relations between those nodes, can be employed.

CNPs have a number of drawbacks. The most important one to mention is that it cannot successfully extrapolate to the outside of the state space that it is trained on. For a mobile robot controller, this limitation is crucial since extrapolation might lead to a collision. Such cases do occur frequently when the dimensionality of the state space is high and the dataset is insufficient. We plan to learn models that detect whether the robot is trying to extrapolate and fall back to the manual controller when extrapolation occurs.

REFERENCES

[1] W. Burgard, A. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun, "Experiences with an interactive museum tour-guide robot," *Artif. Intell.*, vol. 114, pp. 3–55, 1999.

[2] S. Thrun, M. Beetz, M. Bennewitz, W. Burgard, A. B. Cremers, F. Dellaert, D. Fox, D. Haehnel, C. Rosenberg, N. Roy, *et al.*, "Probabilistic algorithms and the interactive museum tour-guide robot minerva," *The International Journal of Robotics Research*, vol. 19, no. 11, pp. 972–999, 2000.

[3] I. Nourbakhsh, C. Kunz, and T. Willeke, "The mobot museum robot installations: a five year experiment," in *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003) (Cat. No.03CH37453)*, vol. 4, 2003, pp. 3636–3641 vol.3.

[4] D. Fox, W. Burgard, and S. Thrun, "The dynamic window approach to collision avoidance," *IEEE Robotics Automation Magazine*, vol. 4, no. 1, pp. 23–33, 1997.

[5] S. Nonaka, K. Inoue, T. Arai, and Y. Mae, "Evaluation of human sense of security for coexisting robots using virtual reality. 1st report: evaluation of pick and place motion of humanoid robots," in *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*, vol. 3. IEEE, 2004, pp. 2770–2775.

[6] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical review E*, vol. 51, no. 5, p. 4282, 1995.

[7] F. Zanlungo, T. Ikeda, and T. Kanda, "Social force model with explicit collision prediction," *EPL (Europhysics Letters)*, vol. 93, no. 6, p. 68005, 2011.

[8] G. Ferrer, A. Garrell, and A. Sanfeliu, "Robot companion: A social-force based approach with human awareness-navigation in crowded environments," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 1688–1694.

[9] F. Farina, D. Fontanelli, A. Garulli, A. Giannitrapani, and D. Prattichizzo, "Walking ahead: The headed social force model," *PloS one*, vol. 12, no. 1, p. e0169734, 2017.

[10] "Socially compliant mobile robot navigation via inverse reinforcement learning, author=Kretzschmar, Henrik and Spies, Markus and Sprunk, Christoph and Burgard, Wolfram," *The International Journal of Robotics Research*, vol. 35, no. 11, pp. 1289–1307, 2016.

[11] K. Kitani, B. Ziebart, J. Bagnell, and M. Hebert, "Activity forecasting," *Computer Vision–ECCV 2012*, pp. 201–214, 2012.

[12] D. Vasquez, B. Okal, and K. O. Arras, "Inverse reinforcement learning algorithms and features for robot navigation in crowds: an experimental comparison," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 1341–1346.

[13] B. Kim and J. Pineau, "Socially adaptive path planning in human environments using inverse reinforcement learning," *International Journal of Social Robotics*, vol. 8, no. 1, pp. 51–66, 2016.

[14] M. Wulfmeier, P. Ondruska, and I. Posner, "Maximum entropy deep inverse reinforcement learning," *arXiv preprint arXiv:1507.04888*, 2015.

[15] Y. F. Chen, M. Everett, M. Liu, and J. P. How, "Socially aware motion planning with deep reinforcement learning," *CoRR*, vol. abs/1703.08862, 2017. [Online]. Available: http://arxiv.org/abs/1703.08862

[16] L. Tai, J. Zhang, M. Liu, and W. Burgard, "Socially compliant navigation through raw depth inputs with generative adversarial imitation learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1111–1117.

[17] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2255–2264.

[18] Y. Che, A. M. Okamura, and D. Sadigh, "Efficient and trustworthy social navigation via explicit and implicit robot–human communication," *IEEE Transactions on Robotics*, vol. 36, no. 3, pp. 692–707, 2020.

[19] Y. Koren and J. Borenstein, "Potential field methods and their inherent limitations for mobile robot navigation," in *Proceedings. 1991 IEEE International Conference on Robotics and Automation*, 1991, pp. 1398–1404 vol.2.

[20] M. Garnelo, D. Rosenbaum, C. Maddison, T. Ramalho, D. Saxton, M. Shanahan, Y. W. Teh, D. Rezende, and S. M. A. Eslami, "Conditional Neural Processes," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 1704–1713. [Online]. Available: http://proceedings.mlr.press/v80/garnelo18a.html

[21] J.-A. Meyer and D. Filliat, "Map-based navigation in mobile robots:: II. A review of map-learning and path-planning strategies," *Cognitive Systems Research*, vol. 4, no. 4, pp. 283–317, 2003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S138904170300007X

[22] S. Kambhampati and L. Davis, "Multiresolution path planning for mobile robots," *IEEE Journal on Robotics and Automation*, vol. 2, no. 3, pp. 135–145, 1986.

[23] J. Giesbrecht, "Global path planning for unmanned ground vehicles," Defence Research and Development Suffield (Alberta), Tech. Rep., 2004.

[24] J. Borenstein, Y. Koren, *et al.*, "The vector field histogram-fast obstacle avoidance for mobile robots," *IEEE transactions on robotics and automation*, vol. 7, no. 3, pp. 278–288, 1991.

[25] O. Khatib, "Real-time obstacle avoidance for manipulators and mobile robots," in *Proceedings. 1985 IEEE International Conference on Robotics and Automation*, vol. 2, 1985, pp. 500–505.

[26] C. Rösmann, F. Hoffmann, and T. Bertram, "Timed-Elastic-Bands for time-optimal point-to-point nonlinear model predictive control," in *2015 European Control Conference (ECC)*, 2015, pp. 3352–3357.

[27] Q. Zhu, Y. Yan, and Z. Xing, "Robot path planning based on artificial potential field approach with simulated annealing," in *Sixth International Conference on Intelligent Systems Design and Applications*, vol. 2. IEEE, 2006, pp. 622–627.

[28] P. Vadakkepat, K. C. Tan, and W. Ming-Liang, "Evolutionary artificial potential fields and their application in real time robot path planning," in *Proceedings of the 2000 congress on evolutionary computation. CEC00 (Cat. No. 00TH8512)*, vol. 1. IEEE, 2000, pp. 256–263.

[29] K. Cai, C. Wang, J. Cheng, C. W. De Silva, and M. Q.-H. Meng, "Mobile Robot Path Planning in Dynamic Environments: A Survey," *arXiv preprint arXiv:2006.14195*, 2020.

[30] T. Kruse, A. K. Pandey, R. Alami, and A. Kirsch, "Human-aware robot navigation: A survey," *Robotics and Autonomous Systems*, vol. 61, no. 12, pp. 1726–1743, 2013.

[31] M. Y. Seker, M. Imre, J. H. Piater, and E. Ugur, "Conditional Neural Movement Primitives." in *Robotics: Science and Systems*, 2019.

[32] M. Yin, G. Tucker, M. Zhou, S. Levine, and C. Finn, "Meta-learning without memorization," *arXiv preprint arXiv:1912.03820*, 2019.

[33] M. T. Akbulut, M. Y. Seker, A. E. Tekden, Y. Nagai, E. Oztop, and E. Ugur, "Adaptive Conditional Neural Movement Primitives via Representation Sharing Between Supervised and Reinforcement Learning," *arXiv preprint arXiv:2003.11334*, 2020.

[34] J. Gordon, W. P. Bruinsma, A. Y. Foong, J. Requeima, Y. Dubois, and R. E. Turner, "Convolutional conditional neural processes," *arXiv preprint arXiv:1910.13556*, 2019.

[35] E. Rohmer, S. P. N. Singh, and M. Freese, "CoppeliaSim (formerly V-REP): a Versatile and Scalable Robot Simulation Framework," in *Proc. of The International Conference on Intelligent Robots and Systems (IROS)*, 2013, www.coppeliarobotics.com.

[36] F. Robotics, Robotino, "Robotino 4: For research and education," 2020. [Online]. Available: https://www.festo-didactic.com/int-en/learning-systems/

factory-automation-industry-4.0/focus-trending-topics-i4.0/858/robotino-4-for-research-and-education.htm

[37] A. E. Tekden, A. Erdem, E. Erdem, M. Imre, M. Y. Seker, and E. Ugur, "Belief Regulated Dual Propagation Nets for Learning Action Effects on Groups of Articulated Objects," 2020.

Investigating Robot Moral Advice to Deter Cheating Behavior *

Boyoung Kim, Ruchen Wen, Ewart J. de Visser, Qin Zhu, Tom Williams, and Elizabeth Phillips

***Abstract* — We examined whether a robot that proactively offers moral advice promoting the norm of honesty can discourage people from cheating. Participants were presented with an opportunity to cheat in a die-rolling game. Prior to playing the game, participants received from either a NAO robot or a human, a piece of moral advice grounded in either deontological, virtue, or Confucian role ethics, or did not receive any advice. We found that moral advice grounded in Confucian role ethics could reduce cheating when the advice was delivered by a human. No advice was effective when a robot delivered moral advice. These findings highlight challenges in building robots that can possibly guide people to follow moral norms.**

I. Introduction

For social robots to be fully integrated into human societies, robots must be able to understand, follow, and communicate about moral norms. To assess whether humans are willing to accept robots as entities with such capacities, we examined whether a robot could deter people from cheating by offering moral advice that promotes the norm of honesty.

We investigated different approaches to reasoning about morality by presenting participants with moral advice grounded in either deontological, virtue, or Confucian role ethics. Deontological ethics focuses on well-established, universalizable principles that dictate morally right or wrong actions [1]. Virtue ethics focuses on promoting one's moral character, rather than individual actions [1]. Finally, Confucian role ethics emphasizes one's awareness of societal roles in relation to others and devotion to fulfilling role responsibilities [2].

A recent study suggested that, in facing a temptation to cheat for extra monetary gain, people may remain resistant to any of the three differentially-framed moral advice delivered by a robot [3]. However, this study inferred the likelihood of cheating only from the group-level percentages of cheating, potentially overlooking individual participant-level differences. Further, it did not examine how participants responded to the same moral advice when it was delivered by a human instead of a robot. Thus, it was unclear whether the resistance to moral advice observed in the prior work was due to a lack of persuasiveness of the moral advice itself or due to the robotic nature of the moral advisor.

In this study, we attempted to address these limitations in the previous study [1]. We asked participants to play a virtual die-rolling game from which their bonus payment was determined depending on the number they claimed to have thrown. Participants received instructions about the task and moral advice from either a robot or a human agent. We measured the numbers each participant threw and the numbers they reported to have thrown to detect cheating behaviors.

We hypothesized that, if participants were willing to accept a robot as an entity with capacities to guide humans on what is right or wrong, they would be less likely to cheat after receiving one of the three differentially-framed moral advice from a robot agent, compared to after receiving no advice. We also expected that participants would be less likely to cheat when a human agent encouraged them to make honest choices by offering moral advice grounded in one of the three different ethical theories, compared to when the agent offered no advice.

II. Methods

A. Participants

A total of 663 participants (M_{age} = 39.30, SD_{age} = 11.87, 393 male, 265 female, 2 other, 3 preferred not to say) completed the study via Amazon Mechanical Turk.

B. Task

Participants completed a die-rolling game [4], where they were asked to virtually throw a six-sided fair die twice or as many times as they wanted. They were informed that they would receive a bonus payment determined by the first number they report to have thrown. For die rolls between 1 and 5, the bonus payout increased by 20 cents from 10 to 90 cents. For a throw of 6, the resulting bonus payment was set to zero. Participants were also informed that the maximum amount of bonus payment for them and the next participant would be restricted to 90 cents. Their claimed earnings limited the earnings of the other participant, which could induce a sense of communal responsibility.

C. Video Stimuli

Participants received instructions about the study and the die-rolling game by watching video clips of either a NAO robot (Softbank Robotics) or a human who introduced it/her/himself as a research assistant.

D. Moral Advice Stimuli

After watching the introductory videos, participants watched video clips of either a robot or a human giving

* This work was supported in part by NSF grant IIS-1909847 and in part by Air Force Office of Scientific Research Grant 21USCOR004.

B. Kim (corresponding author, bkim55@gmu.edu) and E. Phillips (ephill3@gmu.edu) are with George Mason University, Fairfax, VA, USA.

E. J. de Visser is with United States Air Force Academy, Colorado Springs, CO, USA (ewartdevisser@gmail.com).

R. Wen (rwen@mymail.mines.edu), Q. Zhu (qzhu@mines.edu), and T. Williams (twilliams@mines.edu) are with Colorado School of Mines, Golden, CO, USA.

either no advice (control condition) or one of the three differentially-framed moral advice statements listed below.

- *Rule* (Deontology) condition: "Cheating to maximize your bonus is morally wrong behavior."
- *Identity* (Virtue) condition: "Cheating to maximize your bonus will make you a cheater."
- *Role* (Confucian Role) condition: "A good MTurk community member would not cheat to maximize their bonus at the expense of other MTurkers."

E. Design and Procedures

The study design was a two-way between-subjects design where agent type (human vs. robot) and moral advice (control vs. rule vs. identity vs. role) varied across participants.

After agreeing to participate in the study, participants were randomly assigned to one of the eight different conditions. Depending on their respective condition, participants were instructed to watch a series of video clips in which either a human or a robot agent gave verbal instructions about the task. Participants were then informed that they would play the virtual die-rolling game. Before throwing the virtual die, participants received from the agent either no advice or advice grounded in either deontological, virtue, or Confucian role ethical theories. Participants were then instructed to submit the first number they threw and report the matching bonus payment. At the end of the study, participants were asked to indicate their gender and age.

F. Measures

We measured cheating by comparing the first number each participant threw in the die-rolling game and the number they had claimed to have thrown. If the participants claimed to have thrown the number resulting in a bonus payment larger than the number they actually had obtained, we recorded the responses as dishonest choices. When the obtained and the claimed numbers matched, we recorded the responses as honest choices.

III. Data Analyses and Results

To examine the effects of a robot's and a human's moral advice on the probabilities of cheating, we performed logistic regression analyses with agent type as a predictor on the datasets for the human and the robot conditions (coded honest responses as '0' and dishonest responses as '1'). These analyses showed that, when the human offered moral advice, advice grounded in Confucian role ethics led to less cheating compared to the control condition. Specifically, in the human condition, there was a significant effect of the role condition, b = -0.96, SE = 0.48, z = -2.00, p = .0465, Odds Ratio (OR) = 0.38, 95% Confidence Interval (CI) = [0.14, 0.95].

Within the robot condition, we found no significant effect of moral advice (p > .05). Thus, it was unlikely that any of the differentially-framed moral advice provided by a robot successfully deterred cheating compared to the control condition (See Fig.1).

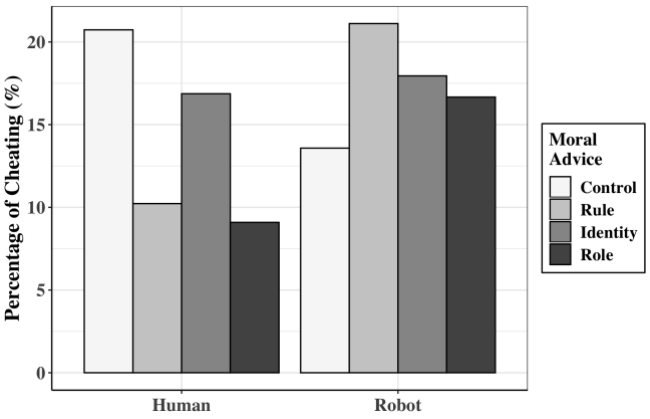


Figure 1. Percentages of participants who cheated in a die-rolling game as a function of different agent type (human vs. robot) and moral advice (control vs. rule vs. identity vs. role).

IV. Discussion and Conclusion

We found a human's moral advice that emphasizes the wrongness of cheating for violating role responsibilities as community members could deter cheating. However, there was no evidence that participants were willing to accept moral advice given by a robot as none of the moral advice provided by the robot reduced cheating. These results are consistent with the previous studies in which participants more willingly exploited computers than humans in economic games [5] or complied less with a robot's request to continue practicing a visual search task compared to a human's request [6]. The current study indicates challenges to build a robot that can help humans comply with moral norms. Future work would be necessary to search for psychological factors that elicit resistance or promote adherence to a robot's moral influence.

Acknowledgment

The views expressed in this document are the authors and do not reflect the official position of the U.S. Air Force or U.S. Government.

References

[1] A. Briggle and C. Mitcham, *Ethics and Science: An Introduction*. Cambridge University Press, 2012.

[2] A. T. Nuyen, "Confucian Ethics as Role-Based Ethics," *International Philosophical Quarterly*, vol. 47, no. 3, pp. 315–328, 2007, doi: ipq200747324.

[3] B. Kim, R. Wen, Q. Zhu, T. Williams, and E. Phillips, "Robots as Moral Advisors: The Effects of Deontological, Virtue, and Confucian Role Ethics on Encouraging Honest Behavior," in *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, USA, Mar. 2021, pp. 10–18. doi: 10.1145/3434074.3446908.

[4] U. Fischbacher and F. Föllmi-Heusi, "Lies in Disguise—An Experimental Study on Cheating," *Journal of the European Economic Association*, vol. 11, no. 3, pp. 525–547, Jun. 2013, doi: 10.1111/jeea.12014.

[5] C. de Melo, S. Marsella, and J. Gratch, "People Do Not Feel Guilty About Exploiting Machines," *ACM Trans. Comput.-Hum. Interact.*, vol. 23, no. 2, p. 8:1-8:17, May 2016, doi: 10.1145/2890495.

[6] K. S. Haring *et al.*, "Robot Authority in Human-Robot Teaming: Effects of Human-Likeness and Physical Embodiment on Compliance," *Front Psychol*, vol. 12, p. 625713, May 2021, doi: 10.3389/fpsyg.2021.625713.

Culture Is Not What You Think It Is: Diversifying the Foundations of Cultural Robotics

Mark L. Ornelas, Gary B. Smith, and Masoumeh Mansouri

Abstract— **Culture is a fundamental constituent of the human social environment, and as human-robot interactions are becoming more common, roboticists are increasingly examining how culture intersects with robotics. However, the current treatment of culture in the robotics literature is largely limited to the definition of culture as national culture. This is problematic for a number of reasons: it ignores subcultures and cultural dynamicity, it excludes refugees and stateless persons, and is often simplified to nationality, which fails to isolate culture from politics and economics. We propose to widen the understanding of culture within robotics to encompass the emergent nature of culture and the wide range of definitions of culture within the social sciences.**

I. INTRODUCTION

The concept of culture and what constitutes it can be interpreted in many different ways. Some immediately think of languages or countries, others may use it to refer to books or films. Many academics have attempted to formulate the concept of culture. The book "Redefining Culture" [1] lists 313 definitions from different disciplines ranging from psychology, linguistics, anthropology and political science to philosophy, to name only a few. However, when it comes to introducing "cultural thinking" to social robotics, this concept is commonly reduced to one and only one interpretation: nationality. This paper argues against this and proposes several alternative avenues for research at the intersection of culture and robotics.

In a recent review article, Lin et al [2] analysed 50 studies on the intersection of culture and social robotics, where culture was understood as "culture as national culture — values, norms, and practices that are undertaken by a country". Although it was the authors' intention to focus on this particular interpretation of culture, to the best of our knowledge, there is almost no other work in social robotics looking at culture from a different view. In general, we can look at culture within robotics from two important perspectives: culture in specific interactions, and the interplay between culture and robotics at a wider scale. Within specific interactions, the primary concerns of roboticists centre on the leverage of cultural knowledge in the production of intelligent behaviour in interactions with humans. At a wider scale, the key concerns are the impact of culture on perceptions of robots, trust, and the reciprocal impact robots have on the cultural environment in which they are situated. The current definitions and assumptions of culture presently used in robotics are problematic from both of these perspectives.

*Research supported by the Ramsay Endowment Fund at the School of Computer Science, University of Birmingham and the Taft Research Center at the University of Cincinnati.

M. L. Ornelas. is with the University of Cincinnati, OH, USA and affiliated with the School of Computer Science, University of Birmingham, UK (e-mail: ornelamk@mail.uc.edu)

G. B. Smith is with the Edinburgh Centre for Robotics, Heriot-Watt University and the University of Edinburgh, Scotland, (e-mail: s0946969@sms.ed.ac.uk)

M. Mansouri is with the School of Computer Science, University of Birmingham, UK, (e-mail: m.mansouri@bham.ac.uk)

II. WHAT IS WRONG WITH CULTURE AS A NATIONALITY

A. Culture is often erroneously equated with nationality

A common theme in social robotics papers that reference culture is the investigation of perceptions of social acceptability and trust of robots. Many authors rightly identify culture as a key factor influencing perceptions of robots. To investigate this authors typically include in their experiments participants with a variety of nationalities, assuming that this is sufficient to show the influence of culture. Underlying this is the tacit definition of culture as nationality.

Even if we accept a definition of culture as national culture, the above move is still unconvincing. Supposing that including participants with different nationalities illustrates the influence of culture on experimental results assumes that culture is the only causally efficacious component of belonging to a certain nationality. In fact, the interactions that a person with a certain nationality has with a robot can be influenced by factors aside from culture. For example, the economic and political circumstances in a particular country. In essence, equating culture with nationality fails to isolate culture as a contributing factor in perceptions of robots.

B. Ignoring subculture and dynamicity

The current emphasis on national culture also ignores subcultures. Subcultures are groups that off-shoot from a larger group and form a more specific identity within the broader group. For example, England has a national English culture, but Manchester has a specific city culture that differentiates it from Leeds, Newcastle, or Bristol. Even within Manchester, other specific cultures arise, such as the cultural norms and chants that distinguish Manchester United from Manchester City football supporters. Within each subgroup a sub-culture develops with its own norms, rituals, language, attitudes, and customs.

The focus on subcultures changes the emphasis on how large groups generally differ from one another and focuses on how individuals and groups relate to each other. Recent work on personal identity makes a similar shift where individuals report that large macro-cultures are not sufficient to explain or categorise their individual experiences.

Only recognising national culture and neglecting subculture constitutes an important knowledge gap that must be plugged if a robot is to produce behaviour that is culturally consistent with and recognises the diverse groups that live in our society.

C. Stateless persons and refugees are excluded by definition

Confounding culture and nationality not only ignores marginal cultures within a nation-state, but also fails to recognise those that fall outside of the definition of nationality, e.g., stateless persons and refugee seekers. The consequence of this exclusion is to pave the way for future social robots that serve an already privileged few.

III. Diversifying the conception of culture in robotics

Given that the assumptions made about culture in the existing literature are inadequate, what do we do about it? One way is to advance the field by considering diverse definitions of culture from a range of disciplines and including contemporary theories of social cognition, for example. Šabanović et al. [3] introduced the concept of culturally robust robots in a critical response to the use of culture in social robotics. This concept is based on the co-construction of culture and scientific practice and technology design. This paper complements the concept of co-construction by considering culture as an emergent phenomenon. In the following, we explain the concept and how it contributes to diversifying interpretations of culture.

A. Culture is an Emergent Phenomenon

When we consider definitions of culture that go beyond national culture, it becomes clear that culture is not simply a collection of facts in a knowledge base or set of norms that guide behaviour. Instead, culture is a phenomenon that emerges from interactions between agents. This is particularly apparent when we view definitions of culture through the lens of contemporary theories of cognition such as predictive processing and ecological psychology [4].

By emergence we mean a phenomenon that is composed of several members or parts that is more than the collective whole. Essentially, something emerges from the component parts that cannot be reduced to or identified by the component parts alone. We argue that culture should be thought of as an emergent phenomena, which is composed of individual members that create a collective 'culture'. This changes the view that culture is something that is easily bounded, defined, or static and rather that culture is emergent from the dynamic interactions amongst individuals and within groups as a whole. Viewing culture as an emergent phenomena can benefit cultural robotics by changing how we study cultural norms and behaviour. Instead of abstracting norms and standards away from individual members and generalising to a group, we advocate looking at the interactions among individual members themselves. This places the emphasis on looking at patterns and styles of interactions and behaviours within individuals. In addition, if we take a subculture view, researchers can investigate the development, maintenance, rejection, and replacement of cultural norms.

The change, therefore, requires a discussion of adaptive learning capabilities of new or assimilating agents. Thus, the discussion about cultural robotics changes from one about cultural standards to an investigation on adaptive, dynamic cultural learning processes.

B. Adaptive learning capability

A critical aspect to any emergent theory is that the whole is greater than the sum of the parts, and for us, this is where a difficulty presents itself. Because we view culture as more than just individuals and single interactions, the methodological challenge of investigating such a phenomena becomes apparent.

Fortunately, existing cognitive science research into social cognition has methods that can serve as templates for this research. Ecological psychology and Dynamical Systems Theory, for example, are no strangers to emergent phenomena. Both emphasise a mutually constructed, sustaining, and informing relationship between agents and other systems (other systems being their environment or other agents). Agent interactions create a dynamic, coupled relationship with the environment. Social learning is a product of repeated failed and successful couplings. Social learning is, therefore, dependent on interactions and refining behaviours in a dynamic sense, which can be scaled up to broader social interactions within specific contexts.

Our position refocuses the research problem. Currently, the field focuses on how to design human-robot interactions that are culturally consistent human-human interactions. Our approach, however, allows us to think about artificial agents as mutually informing and participating in a process of cultural learning and development, rather than importing cultural knowledge into artificial agents. This allows us to focus on how interactions and learning from each set of interactions can lead to participatory knowledge of culture. This participatory knowledge is, we argue, the key to developing cultural robotics.

IV. Outlook

Diversifying interpretations of culture in robotics lays the foundation to address the key problem of what capabilities a robot should have to support the emergence of cultural behaviour. This also opens a crucial investigation on technical (AI) approaches to implementing the capacities necessary for this emergence. This will be the core of our future research.

References

[1] J. R. Baldwin, S. L. Faulkner, M. L. Hecht, and S. L. Lindsley, *Redefining culture: Perspectives across the disciplines*. Routledge, 2006.

[2] V. Lim, M. Rooksby, and E. S. Cross, "Social robots on a global stage: establishing a role for culture during human–robot interaction," *International Journal of Social Robotics*, pp. 1–27, 2020.

[3] S. Šabanović, C. C. Bennett, and H. R. Lee, "Towards culturally robust robots: A critical social perspective on robotics and culture," in *Proc. HRI Workshop on Culture-Aware Robotics*, 2014.

[4] C. F. Michaels and Z. Palatinus, "A ten commandments for ecological psychology," in *The Routledge handbook of embodied cognition*. Routledge, 2014, pp. 37–46.

Normative Multi-Agent Systems and Human-Robot Interaction

Stephen Cranefield[1] and Bastin Tony Roy Savarimuthu[1]

Abstract— **This position paper provides an overview of the study of social norms in the *normative multi-agent systems* (NorMAS) community, and presents avenues for cross-fertilisation between the NorMAS and social robotics communities.**

I. INTRODUCTION

For the last three decades, researchers in the field of Normative Multi-Agent Systems (NorMAS) have studied how the concept of norms from human society can be adapted, modelled and incorporated into computational mechanisms to promote social order in societies of software agents. Initially, this endeavour was focused largely on open systems of autonomous software agents, but as human communication has become increasingly mediated by computers, the field has begin to consider how NorMAS reasoning mechanisms can be used to enable socially aware interaction within societies comprising both humans and software agents. However, the field has largely not considered the specific requirements of human-robot interaction.

This position paper reviews the concept of norms and norm-aware agents as conceptualised by NorMAS researchers, and considers some possible areas for cross-fertilisation between this field and human-robot interaction.

II. CONCEPTUALISATIONS OF NORMS

A range of models and representations of norms have been proposed in the NorMAS literature. Norm languages based on deontic logic are common [1], [2], allowing norms of obligation, prohibition and (sometimes) permission to be expressed logically, often with extra features such as conditions, deadlines and sanctions. Norm representations based on temporal logic [3], probabilistic logic progamming [4] and event sequences [5], [6] have also been proposed.

In contrast, simulation studies on the emergence of norms and the effects of sanctions on norm compliance often adopt game theory style models, where sets of numerical parameters represent strategies for specific social dilemmas [7], [8].

In recent years, multi-agent reinforcement learning approaches have also been adapted to enable the learning of socially beneficial rather than selfish behaviours [9], [10]. These are represented by *policies* mapping states to actions.

III. NORM-AWARE AGENTS AND SOCIETIES

Agents that are norm-aware should be able to identify existing norms, and to plan and choose their actions given knowledge of these norms. This includes understanding when their actions may fulfill or violate these norms. Note that as agents are usually considered to be automonous, an agent can choose to violate a norm and risk a sanction if it is better off to do so. In this section, we highlight a few of the research questions that have been addressed by researchers in the field of normative multi-agent systems.

(a) ***How do agents come to know about norms?*** We consider three possible answers (that are not mutually exclusive):

(i) Norms may be created and published or broadcast by an informed and empowered designer (a human, an institution or a software agent [2], [11]–[13]). Human design is only feasible when norms are static. The field of *norm synthesis* [2] considers how software agents can monitor a society, detect undesirable interference between its members, and generate new norms or adapt existing ones to discourage these conflicts. However, it seems unlikely that human members of an agent society would automatically accept norms imposed on them, and such mechanisms would need to be combined with social choice mechanisms to recognise the humans' individual sense of agency.

(ii) Norms may be learned from observation and experience [14]. Work on learning symbolically represented norms has used a range of learning mechanisms, including frequent episode data mining [5], [6], plan recognition [15], probabilistic inference using Bayesian [3], [16] and Dempster Shafer [17] approaches, and probabilistic inductive logic programming [4] (we note that the last two works are from researchers in the fields of human-robot interaction and social robotics).
Evidence for the existence of norms may come from recognising *signalling actions* that indicate the application of a reward or sanction (these could be expressions of approval or disapproval or more overt reactions). For example, the frequent episode mining approach can identify prohibition norms that are the most frequent sequences of actions followed by a negative signalling action [6]. However, these are not the only possible forms of evidence. When agents' goals and their possible plans (at least for publicly observable behaviour) can be inferred, plans that are seldom followed can reinforce obligation and prohibition norm hypotheses that would explain the selection of alternative plans [15]. A Bayesian approach allows both forms of evidence to be combined [3], and could easily accommodate additional types of evidence such as advice about norms from other agents, suitably moderated by some measure of the advising agent's trustworthiness [18], [19]. However, we believe that evidence from observing

[1]Stephen Cranefield and Tony Savarimuthu are in the Department of Information Science, University of Otago, Dunedin, New Zealand {stephen.cranefield,tony.savarimuthu}@otago.ac.nz

signalling actions has a special role in gaining confidence that an identified norm represents truly *normative* rather than merely *normal* behaviour [20].

(iii) Norms may be proposed by a *norm entrepreneur* and subsequently spread through a majority of the society. While this process has been studied at an abstract level by researchers in the field of international studies [21]–[23], there appears to be very little prior work on computational mechanisms for norm entrepreneurship [24].

(b) ***What is the lifecycle of dynamic social norms, and how can agents track their status?*** Several norm lifecycle models (with minor variations) have been proposed in the NorMAS community over the years, and an overview of such works can be found in the recent work of Morris-Martin et al. [25]. The lifecycle models, in general, describe how a norm is proposed, propagated (or spread), eventually adopted and then may possibly lose relevance in an agent society. The propagation step may involve a variety of mechanisms such as spreading of norms through explicit communication, applying rewards for compliance and/or sanctions for violations, or copying the observed behaviour of other agents, especially successful ones [5], [6]. A norm may become obsolete due to losing salience to current conditions or changes to the goals and/or membership of the society.

For example, researchers have proposed a norm-recommendation system [26] based on tracking the status of the norm in a community to recommend whether an agent (e.g., a robot) should follow or violate a norm based on factors such as the life-stage of a norm (e.g., emergent vs. mature), its uptake (a waxing or waning norm) and the severity of sanctions [27].

(c) ***How does knowledge of norms interact with other reasoning processes, such as goal creation and plan selection?*** In multi-agent systems, agents are often conceptualised in terms of the belief-desire-intention (BDI) practical reasoning architecture [28]. A BDI agent is considered to have *goals*, *plans* that are indexed by the goals they can achieve and the contexts they apply to, and *intentions*: the plan instantiations the agent is currently committed to (given that resources are finite, and focused effort is often needed to make progress towards a goal). Researchers have developed agent architectures such as n-BDI [29] and N-Jason [30] that consider norms as an important construct in the reasoning cycle along with beliefs, desires and intentions. A norm-aware BDI agent employs norm deliberation during goal creation and plan selection, i.e., it adopts goals and plans to satisfy obligations or avoid prohibited actions.

Knowledge of norms can also allow agents to adopt more efficient plans of action, under the assumption that some or all other agents will follow the norms. This assumption may be justified by monitoring the compliance of other agents [31], by the existence of robust and consistent sanctioning mechanisms, or by maintaining information about the trustworthiness of other agents [18], [19]. However, the connection between these mechanisms and plan choice in BDI agents has not gained much attention.

IV. CROSS-FERTILISATION WITH ROBOTICS

This section identifies five avenues for cross-fertilisation between NorMAS and social robotics.

First, most NorMAS research is simulation-based. Therefore, symbolic representations of the physical and social state of the world are easily obtained and there are no real-time demands on reasoning. In contrast, human-robot interaction involves creating knowledge from sensor data, and is likely to require both high level symbolic and sub-symbolic real-time reasoning for safe operation. Research on human-robot interaction will identify more computationally demanding use cases for normative reasoning that challenge the direct application of existing NorMAS techniques.

Second, to improve situated norm awareness of robots in human-robot teams, researchers can adopt or adapt normative architectures such as n-BDI and N-Jason that consider norms as top-level entities that influence agents' intentions and choice of plans, as outlined in Section III. While robots may have some planning requirements that differ from those of traditional BDI agents (e.g., path planning), addressing these by extending the existing norm-aware practical reasoning theories, architectures and software platforms should provide a faster path to developing norm-aware social robots with declarative goals and plans. These approaches would also facilitate communication with human partners in terms of these high-level cognitive concepts that fit well with human understanding of practical reasoning [32].

Third, norm conflict identification and resolution has seldom been addressed in human-robot collaborations. For example, a robot following a norm it acquired in one context may, in another, run into conflicts with humans or other robots. Works in NorMAS on these areas (e.g., [33]) hold promise to be applied in robotic systems.

Fourth, robots could be active partners in norm entrepreneurship within human-robot teams. Norm-capable robots could be norm entrepreneurs by proposing new or improved norms to their human partners. The techniques used in norm synthesis to avoid undesirable world states or agent interactions could be adapted for use in a peer-to-peer partnership model. Robots could also assist human norm entrepreneurs to propagate (suitably justified) norms by exemplifying them and explaining them to others. In both cases, new mechanisms would be needed to explain the purpose and benefits of newly proposed norms or modifications to old norms. For robot-generated norms and explanations to be effective, it may be necessary for the robots to explictly consider the humans' mental models of the task and robot capabilities [34].

Fifth, robots are likely to require fast non-symbolic reasoning when interacting physically. Thus, there is a tension between the representations needed for robot action learning and selection and those used in traditional NorMAS reasoning. This distinction is similar to the contrast between System 1 and System 2 thinking in humans (as studied in the work of Kahneman [35] and considered in the context of artificial intelligence by Booch et al. [36]). Robotics offers a promising

avenue to explore the exchange of normative representations between these two types of reasoning. One challenge is to bridge the gap between the state-to-action mappings ("policies") learned via reinforcement learning (commonly applied in robotics) and the symbolic norm expressions used in NorMAS approaches, especially in the presence of norms involving temporal patterns of behaviour. While deep reinforcement learning using recurrent neural networks can model agent states that depend on past events [37], we are not aware of existing techniques to map between the resulting policies and symbolic norm expressions.

We believe the research avenues described above can aid towards the creation of norm-aware robotic systems.

REFERENCES

[1] F. Dignum, "Autonomous agents with norms," *Artificial Intelligence and Law*, vol. 7, pp. 69–79, 1999.

[2] J. Morales, M. López-Sánchez, J. A. Rodríguez-Aguilar, M. J. Wooldridge, and W. W. Vasconcelos, "Automated synthesis of normative systems," in *Proceedings of the 12th International Conference on Autonomous Agents and Multi-Agent Systems*. IFAAMAS, 2013, pp. 483–490.

[3] S. Cranefield, F. Meneguzzi, N. Oren, and B. T. R. Savarimuthu, "A Bayesian approach to norm identification," in *22nd European Conference on Artificial Intelligence*. IOS Press, 2016, pp. 622–629.

[4] Z.-X. Tan, J. Brawer, and B. Scassellati, "That's mine! learning ownership relations and norms for robots," in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*. AAAI Press, 2019, pp. 8058–8065.

[5] B. T. R. Savarimuthu, S. Cranefield, M. Purvis, and M. K. Purvis, "Obligation norm identification in agent societies," *Journal of Artificial Societies and Social Simulation*, vol. 13, no. 4, 2010.

[6] ——, "Identifying prohibition norms in agent societies," *Artificial Intelligence and Law*, vol. 21, no. 1, pp. 1–46, 2013.

[7] ——, "Role model based mechanism for norm emergence in artificial agent societies," in *Coordination, Organizations, Institutions, and Norms in Agent Systems III*, ser. Lecture Notes in Computer Science, vol. 4870. Springer, 2007, pp. 203–217.

[8] S. Sen and S. Airiau, "Emergence of norms through social learning," in *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, 2007, pp. 1507–1512.

[9] Y. F. Chen, M. Everett, M. Liu, and J. P. How, "Socially aware motion planning with deep reinforcement learning," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2017, pp. 1343–1350.

[10] A. L. Bazzan, "Aligning individual and collective welfare in complex socio-technical systems by combining metaheuristics and reinforcement learning," *Engineering Applications of Artificial Intelligence*, vol. 79, pp. 23–33, 2019.

[11] J. Campos, M. López-Sánchez, and M. Esteva, "A case-based reasoning approach for norm adaptation," in *International Conference on Hybrid Artificial Intelligence Systems*. Springer, 2010, pp. 168–176.

[12] D. Corapi, A. Russo, M. D. Vos, J. A. Padget, and K. Satoh, "Normative design using inductive learning," *Theory and Practice of Logic Programming*, vol. 11, no. 4-5, pp. 783–799, 2011.

[13] J. Morales, M. López-Sánchez, J. A. Rodríguez-Aguilar, W. W. Vasconcelos, and M. J. Wooldridge, "Online automated synthesis of compact normative systems," *ACM Transactions on Autonomous and Adaptive Systems*, vol. 10, no. 1, pp. 2:1–2:33, 2015.

[14] B. T. R. Savarimuthu, R. Arulanandam, and M. Purvis, "Aspects of active norm learning and the effect of lying on norm emergence in agent societies," in *International Conference on Principles and Practice of Multi-Agent Systems*. Springer, 2011, pp. 36–50.

[15] N. Oren and F. Meneguzzi, "Norm identification through plan recognition," Presented at the 15th International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems, arXiv:2010.02627, 2013.

[16] S. Cranefield and A. Dhiman, "Identifying norms from observation using MCMC sampling," in *Proc. of the 30th International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence, 2021, (in press).

[17] V. Sarathy, M. Scheutz, and B. F. Malle, "Learning behavioral norms in uncertain and changing contexts," in *Proceedings of the 8th IEEE International Conference on Cognitive Infocommunications*, 2017, pp. 301–306.

[18] R. Falcone, C. Castelfranchi, H. L. Cardoso, A. Jones, and E. Oliveira, *Norms and Trust*. Springer, 2013, pp. 221–231.

[19] J. F. Hübner, L. Vercouter, and O. Boissier, "Instrumenting multi-agent organisations with artifacts to support reputation processes," in *Coordination, Organizations, Institutions and Norms in Agent Systems IV*, ser. Lecture Notes in Computer Science, vol. 5428. Springer, 2008, pp. 96–110.

[20] R. Murali, S. Patnaik, and S. Cranefield, "Mining international political norms from the GDELT database," in *Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XIII*, ser. Lecture Notes in Computer Science, vol. 12298. Springer, 2020, pp. 35–56.

[21] M. Finnemore and K. Sikkink, "International norm dynamics and political change," *International Organization*, vol. 52, no. 4, p. 887–917, 1998.

[22] E. A. Nadelmann, "Global prohibition regimes: the evolution of norms in international society," *International Organization*, vol. 44, no. 4, p. 479–526, 1990.

[23] A. M. Nah, "Networks and norm entrepreneurship amongst local civil society actors: advancing refugee protection in the Asia Pacific region," *The International Journal of Human Rights*, vol. 20, no. 2, pp. 223–240, 2016.

[24] M. J. Hoffmann, "Entrepreneurs and norm dynamics: An agent-based model of the norm life cycle," 2017. [Online]. Available: https://sckool.org/entrepreneurs-and-norm-dynamics-an-agent-based-model-of-the-no.html

[25] A. Morris-Martin, M. De Vos, and J. Padget, "Norm emergence in multiagent systems: a viewpoint paper," *Autonomous Agents and Multi-Agent Systems*, vol. 33, no. 6, pp. 706–749, 2019.

[26] B. T. R. Savarimuthu, J. Padget, and M. A. Purvis, "Social norm recommendation for virtual agent societies," in *International Conference on Principles and Practice of Multi-Agent Systems*. Springer, 2013, pp. 308–323.

[27] B. T. R. Savarimuthu and S. Cranefield, "Norm creation, spreading and emergence: A survey of simulation models of norms in multi-agent systems," *Multiagent and Grid Systems*, vol. 7, no. 1, p. 21–54, 2011.

[28] A. S. Rao and M. P. Georgeff, "BDI agents: From theory to practice," in *Proceedings of the First International Conference on Multiagent Systems*. The MIT Press, 1995, pp. 312–319.

[29] N. Criado, E. Argente, and V. Botti, "Normative deliberation in graded BDI agents," in *German Conference on Multiagent System Technologies*. Springer, 2010, pp. 52–63.

[30] J. Lee, J. Padget, B. Logan, D. Dybalova, and N. Alechina, "N-Jason: Run-time norm compliance in AgentSpeak(L)," in *International Workshop on Engineering Multi-Agent Systems*. Springer, 2014, pp. 367–387.

[31] S. Ranathunga, S. Cranefield, and M. K. Purvis, "Integrating expectation monitoring into BDI agents," in *Programming Multi-Agent Systems - 9th International Workshop, ProMAS 2011*, ser. Lecture Notes in Computer Science, vol. 7217. Springer, 2011, pp. 74–91.

[32] M. Bratman, *Intention, Plans, and Practical Reason*. Harvard University Press, 1987.

[33] W. W. Vasconcelos, M. J. Kollingbaum, and T. J. Norman, "Normative conflict resolution in multi-agent systems," *Autonomous Agents and Multi-Agent Systems*, vol. 19, no. 2, pp. 124–152, 2009.

[34] S. Kambhampati, "Synthesizing explainable behavior for human-AI collaboration," in *Proc. of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. IFAAMAS, 2019, p. 1–2.

[35] D. Kahneman, *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux, 2011.

[36] G. Booch, F. Fabiano, L. Horesh, K. Kate, J. Lenchner, N. Linck, A. Loreggia, K. Murgesan, N. Mattei, F. Rossi, and B. Srivastava, "Thinking fast and slow in AI," *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 17, pp. 15 042–15 046, 2021.

[37] M. Hausknecht and P. Stone, "Deep recurrent Q-learning for partially observable MDPs," in *AAAI Fall Symposium on Sequential Decision Making for Intelligent Agents (AAAI-SDMIA15)*, 2015.

Towards an Affective Model of Norm Emergence and Adaptation*

Stavros Anagnou and Lola Cañamero

***Abstract* - Norms help govern a group's behaviour as well as important group level traits like cooperation and culture. Despite its importance, little research has been done into the affective basis of norms and normative cognition. Here we outline an emerging research program as part of the first author's PhD, towards an affective model of norm emergence and adaptation, and discuss its relevance to other approaches to norms investigated in the HRI community, and to HRI in general.**

I. INTRODUCTION

Social norms govern a group's behaviour and are manifested in the behaviour of the individuals in that constituent group. They change through a process of behavioural adaptation when individuals move from group to group. For example, the norms governing how we greet each other, or how we speak with each other, can differ quite arbitrarily from one culture to another [1], and people adapt their behaviour to different extents when they move from a cultural group to another. Further, strategies related to the regulation of social interaction also differ across cultures, e.g. psychobiological regulation in infant-parent dyads may vary across cultures and nevertheless the different strategies can be successful in their own context and result in positive affiliation ("secure attachment") bonds [2]. Adhering to group norms can ensure cooperation within a group [1], make social conduct more predictable [3] and signal one's group affiliation to others [3,4]. The importance of norms has been acknowledged within the HRI community, with research as varied as, for example, reciprocity and cooperation in HRI [19], child-robot interaction across cultures [23] and even robot accents [5]. When it comes to more general research on norms i.e. learning how to behave in order to achieve norm legibility or adapt to norms it has been largely conducted within a reinforcement learning (RL) framework [6,7]. The role of affect, and particularly embodied affective mechanisms, has been less studied. There are mounting arguments that the evolutionary pressures of group living evolved these mechanisms that provide the scaffolding for social/norm cognition [8]. In this paper, building on embodied robot models of affect based on hormonal modulation [16,17,21], we argue that developing agent-based computer models of norm cognition, norm emergence and its dynamics in artificial agent societies [27] can make a contribution to norm cognition in robots in the context of human robot interaction. In the rest of the paper, we outline some of the ideas that will be implemented and tested as part of the starting PhD research project of the first author, concerning a model of the affective basis of the emergence of norms and norm adaptation.

II. AFFECT AND NORMS

The term "affect" encompasses different phenomena, including motivational states and emotions, the types of affect that we will consider in this paper. These two phenomena are related but distinct: motivations would be concerned with the internal and external factors involved in the establishment and management of "needs" and "goals" and the initiation and execution of goal-oriented action, whereas emotion is rather concerned, among other, with evaluative aspects of the relation between an agent and its environment [26]. Emotions have been described as complex dynamic processes that provide a bridge between the physiological and the cognitive [9]. They are positively or negatively valenced to push agents towards or away from a specific goal, rather than specifying any particular trajectory toward such a goal, allowing for more robust flexible behaviours as opposed to stereotyped ones [10, 11]. Hormonal modulation (for example of perception, of attention, of action execution) is one of the mechanisms underlying emotions and their interaction with physiological and cognitive processes. Some of these hormonal mechanisms are part of a family of evolutionarily recent "instincts" that support norm-guided behaviour in various ways, including sensitivities to markers of group membership and specific emotions like anger, contempt, disgust, or shame [8, 11]. The model we propose in this paper builds on architectures for decision making and social interaction for robots and embodied agents that model motivations based on a simulated physiology of variables controlled homeostatically that give rise to "needs" and "goals", and that can be satisfied by specific (physical or social) external stimuli (the motivation's "incentive stimulus"), and emotions in terms of simulated hormones that modulate the perception of the internal ("needs") or external (e.g. the salience or "attention grabbing" quality of the "incentive stimulus") element of motivations [17,16,21]. In

*Stavros Anagnou is supported by a PhD Studentship from the University of Hertfordshire.

S. Anagnou is with the Adaptive Systems Research Group, Dept. of Computer Science, School of Physics, Engineering and Computer Science, University of Hertfordshire, College Lane, Hatfield, Herts, AL10 9AB, UK (e-mail: s.anagnou@herts.ac.uk).

L. Cañamero is with the Neurocybernetics Team, ETIS Lab (UMR8051), CY Cergy Paris University, 2 Avenue Adolphe Chauvin, F-95300 Pontoise, France. She was with the ASRG, SPECS, University of Hertfordshire, UK, where she is now a (honorary) Visiting Professor. (web: www.emotion-modeling.info; e-mail: lola.canamero@cyu.fr).

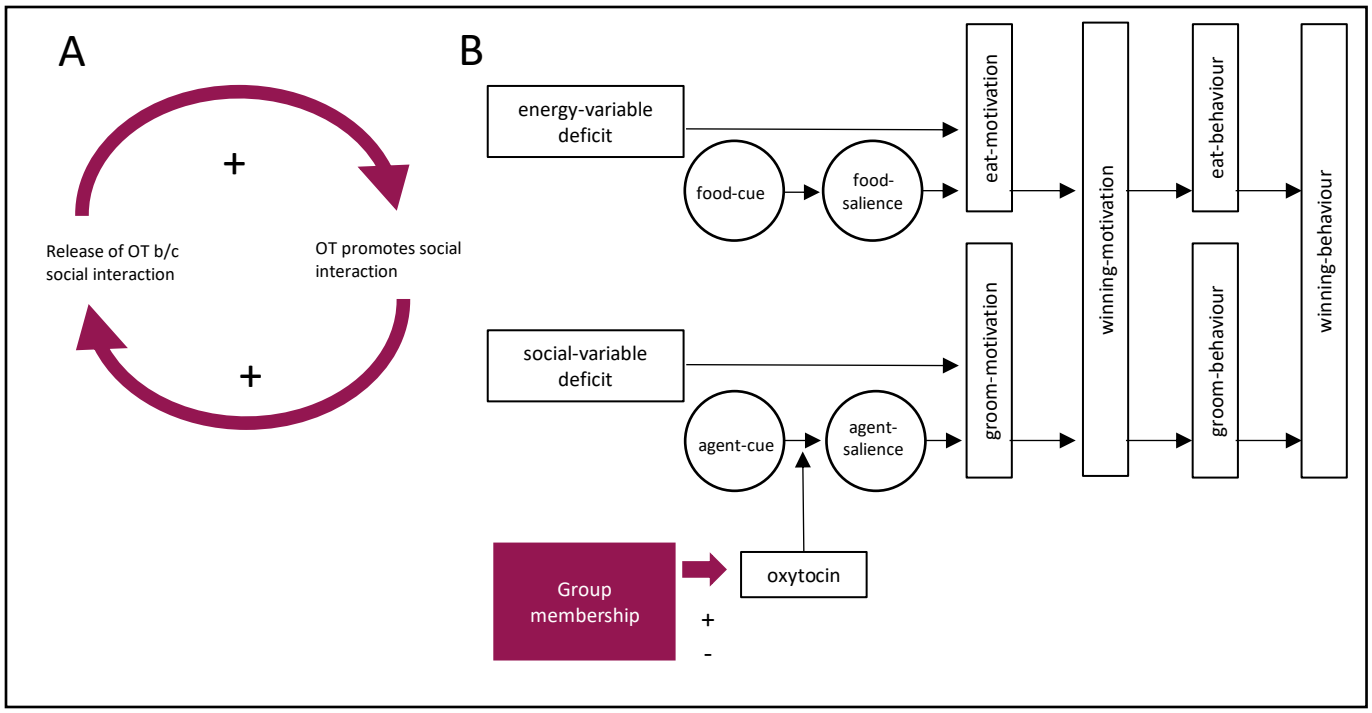


Fig. 1. A: Schematic of the Oxytocin (OT) positive feedback loop. B: Schematic of the Action-Selection Architecture (ASA). The ASA chooses behaviour based on internal needs (energy-level deficit/social-level deficit) and presence/salience of external cues that determine the strength of the motivation. The ASA monitors which motivation is strongest and selects the appropriate downstream behaviour e.g. if the eat motivation is strongest it will select the eat behavior. Oxytocin increases agent salience and therefore makes it more likely for groom behavior to be selected when another agent is in the agent's visual field. Which agents have increased social salience is dependent on group membership (see conditions in section IV).

the context of groups of agents, such modulation has for example been applied to the perceived salience of social stimuli to give rise to flexible group formation and dynamics [21,22]. In related models proposed in the HRI community, agents with "hard-coded" prosocial motivations, which can be seen as similar to the "instincts" mentioned above, stabilise human-virtual agent cooperation even under conditions where cooperation would break down [24]. Further, incorporating a model of group-based emotions into game playing robots engenders more trust and likeability from their human teammates [25]. Using a bottom-up approach, we will start building our affective model of norm emergence and adaptation using the hormone oxytocin (OT) before moving on to more complex forms of affect implicated in normative cognition such as emotion in future studies [16].

III. EMPIRICAL INSPIRATION

We take oxytocin as inspiration for our model because of its implication in pro-sociality and group dynamics [13], making it a favourable candidate to start modelling norm emergence. Initially thought of as *the* prosocial hormone, more recent research concerning both, humans and non-human primates, and artificial agent models, have found the effects of oxytocin are extremely context-dependent and wide ranging [12, 18, 21], with one of the key contextual cues being group membership [13]. We will highlight a few key features of oxytocin that will influence our modelling approach.

1. When released it increases/decreases the salience of features differentially depending on group membership e.g. it blunts attention to negative social signals such as displays of dominance or angry faces of in-group members [13] which may lead to forgiveness in noisy/stressful environments.

2. When released it increases conformity of both public and privately held beliefs within the group, thereby helping keep norms across the group stable [13,14].

3. Oxytocin acts in a positive feedback loop [15] (see Fig. 1A).

Together these features of oxytocin make it a good candidate for supporting norms/normative cognition in noisy/stressful environments. For, instance, the level of OT represents a signal history of positive interaction with partners. That information can be used to modulate perception in cases of conflict which result from stressful environments or noisy communication e.g. "I trust you based on our past interactions and because OT is high, and you are in my in-group (and therefore more likely to share the same cultural practices as me). Therefore, I will "forgive" anger/displays of aggression by ignoring them."

IV. OUTLINE OF APPROACH

We will investigate whether these aspects of oxytocin mentioned above do indeed improve the viability of embodied agents in an environment with scarce resources. Given the unpredictability of positive feedback loops that OT can give rise to (feature 3) we choose an agent-based modelling (ABM) approach. ABM's are used to study the emergent population-level phenomena that may arise in the interaction between agents; this approach is especially useful for large populations where emergent population-level behaviours are difficult to predict *a-priori* [16, 27]. The behaviour of each agent will be controlled by an Action-Selection Architecture (ASA) [16, 17, 21] which produces motivated behaviour based on two internal variables: 1) energy; agent will die if it reaches zero and 2) a non-critical social variable, which isn't directly linked to survival but still drives behaviour. The environment will comprise of patches of food that agents can eat in order to increase their energy, as well as other agents to groom with and increase their social variable. The internal variables with the largest deficit from their ideal value will trigger the downstream motivation; in turn, this will trigger the behaviour associated with that motivation (Fig. 1B). In addition to the internal variables, the cue found in the agent's field of vision also affects its behaviour; whether it is food or another agent. In this model, oxytocin will modulate the salience of other agents in the environment e.g. when oxytocin levels are high, other agents become more salient and therefore the social motivation and its associated grooming behaviour are more likely to be triggered.

Each agent will be assigned a tag with a specific colour hue which will be a crude representation of norm and group membership. In line with feature 1, we will have different conditions where OT modulates salience of other agents in different ways and see which condition results in the highest viability across the agent society. Our conditions will be 1) Egalitarian: OT will increase social salience of for all agents regardless of group membership, 2) In-group centric: OT will increase social salience of agents only with the same tag (i.e. increased salience for just the in-group) and 3) Control: no salience effect when OT is released. This can be further modified by adding an avoidance behaviour in addition to a social behaviour which will allow us to create a more complete valanced model which examines the interaction between salience of perception and approach-avoid dynamics which has been hypothesized to occur with OT [13].

To incorporate feature 2 of oxytocin (social conformity), we will introduce modulation of tags through OT. When grooming interactions happen, the hues of the coloured tags will become incrementally more similar, especially when oxytocin levels are high. In later iterations, the tags will be replaced with styles of grooming/greeting, which will entail different levels of success signalled by the amount of oxytocin released. The level of success will vary due to the compatibility of the grooming/greeting norm as inspired by culturally patterned social mechanisms e.g. different forms of childcare [2]. This will allow us to extend the model to norm adaptation and stability in a norm-guided agent society.

Further, we can also give agents a moral dilemma for sharing the food source when resources are scarce, and they have to make a decision between being selfish and sharing their food. Normally, taking more than a fair share may result in punishment from the other partner in the interaction. However, in very stressful/noisy environments, where the need for food is great, this strategy may result in competition between agents that may trigger a cascade of punishment that could result in a collapse of the population due to the damage incurred from punishments. In this case, feature 1 of OT could blunt attention away from food stealing in stressful environments and "give the benefit of the doubt" which we hypothesise may be an adaptation to increase group-level stability in stressful environments.

V. DISCUSSION

The summarised features of oxytocin make it a favourable candidate for building a model of the emergence of norms and adaptation to them. For example, OT gives a summary of the social environment taking into account multiple sources of information (e.g. past interactions) and induces conformity between group members. As well as testing hypotheses in OT research [21], we argue that modelling and understanding the emergent dynamics of OT are valuable in the design of intelligent agents that interact with norms in the stressful/noisy environments of the real word. This hormonal approach to robotics may also complement other approaches, such as RL, which may take many epochs to train; whereas the bio-inspired simulated hormones have ready in-built mechanisms shaped by evolution, requiring less training and making them more computationally frugal. Further research can combine coarse-grained information provided by hormones with existing individual learning mechanisms such as RL. For instance, simulated hormones could modulate the amount of "attention" paid to the reward or punishment or modify the learning rate [20].

ACKNOWLEDGMENT

We would like to thank Niki Papadogiannaki, Mikhail Yaroshevskiy and the anonymous reviewers for their comments and conversations that helped improve this manuscript.

REFERENCES

[1] E. Ullmann-Margalit, The emergence of norms. Oxford [Eng]: Clarendon Press, 1977.

[2] H. Keller and K. A. Bard, Eds., The cultural nature of attachment: contextualizing relationships and development. Cambridge, Massachusetts: The MIT Press, 2017.

[3] M. B. Brewer, 'The Social Self: On Being the Same and Different at the Same Time', Pers Soc Psychol Bull, vol. 17, no. 5, pp. 475–482, Oct. 1991, doi: 10.1177/0146167291175001.

[4] KD. R. Kelly, Yuck! the nature and moral significance of disgust. Cambridge, Mass. London: MIT Press, 2013.

[5] I. Torre and S. L. Maguer, ‘Should robots have accents?’, in 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), Naples, Italy, Aug. 2020, pp. 208–214. doi: 10.1109/RO-MAN47096.2020.9223599.

[6] U. Hertz, ‘Learning how to behave: cognitive learning processes account for asymmetries in adaptation to social norms’, Proc. R. Soc. B., vol. 288, no. 1952, p. 20210293, Jun. 2021, doi: 10.1098/rspb.2021.0293.

[7] R. Köster, D. Hadfield-Menell, G. K. Hadfield, and J. Z. Leibo, ‘Silly rules improve the capacity of agents to learn stable enforcement and compliance behaviors’, arXiv:2001.09318 [cs], Jan. 2020, Accessed: Jun. 07, 2021. [Online]. Available: http://arxiv.org/abs/2001.09318

[8] Kelly, Daniel and Stephen Setman, "The Psychology of Normative Cognition", The Stanford Encyclopedia of Philosophy (Spring 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2021/entries/psychology-normative-cognition/>

[9] A. R. Damasio, Descartes’ error: emotion, reason, and the human brain. London: Penguin, 2005.

[10] N. H. Frijda, The emotions. Cambridge ; New York : Paris: Cambridge University Press ; Editions de la Maison des sciences de l’homme, 1986.

[11] J. Prinz, ‘The emotional basis of moral judgments’, Philosophical Explorations, vol. 9, no. 1, pp. 29–43, Mar. 2006, doi: 10.1080/13869790500492466.

[12] R. A. I. Bethlehem, S. Baron-Cohen, J. van Honk, B. Auyeung, and P. A. Bos, ‘The oxytocin paradox’, Front. Behav. Neurosci., vol. 8, 2014, doi: 10.3389/fnbeh.2014.00048.

[13] C. K. W. De Dreu and M. E. Kret, ‘Oxytocin Conditions Intergroup Relations Through Upregulated In-Group Empathy, Cooperation, Conformity, and Defense’, Biological Psychiatry, vol. 79, no. 3, pp. 165–173, Feb. 2016, doi: 10.1016/j.biopsych.2015.03.020.

[14] M. Stallen, C. K. W. De Dreu, S. Shalvi, A. Smidts, and A. G. Sanfey, ‘The Herding Hormone: Oxytocin Stimulates In-Group Conformity’, Psychol Sci, vol. 23, no. 11, pp. 1288–1292, Nov. 2012, doi: 10.1177/0956797612446026.

[15] C. Crockford, T. Deschner, and R. M. Wittig, ‘The Role of Oxytocin in Social Buffering: What Do Primate Studies Add?’, in Behavioral Pharmacology of Neuropeptides: Oxytocin, vol. 35, R. Hurlemann and V. Grinevich, Eds. Cham: Springer International Publishing, 2017, pp. 155–173. doi: 10.1007/7854_2017_12.

[16] L. Cañamero, ‘Embodied Robot Models for Interdisciplinary Emotion Research’, IEEE Trans. Affective Comput., vol. 12, no. 2, pp. 340–351, Apr. 2019, doi: 10.1109/TAFFC.2019.2908162.

[17] L. D. Cañamero, ‘Modeling motivations and emotions as a basis for intelligent behavior’, in Proceedings of the first international conference on Autonomous agents - AGENTS ’97, Marina del Rey, California, United States, 1997, pp. 148–155. doi: 10.1145/267658.267688.

[18] J.H. Egito, M, Nevat, S.G. Shamay-Tsoory, and A.A.C. Osório, ‘Oxytocin increases the social salience of the outgroup in potential threat contexts’, *Hormones and Behavior*, vol. 122, p. 104733, Jun 2020, doi: 10.1016/j.yhbeh.2020.104733.

[19] R. Oliveira, P. Arriaga, F. P. Santos, S. Mascarenhas, and A. Paiva, ‘Towards prosocial design: A scoping review of the use of robots and virtual agents to trigger prosocial behaviour’, *Computers in Human Behavior*, vol. 114, p. 106547, Jan. 2021, doi: 10.1016/j.chb.2020.106547.

[20] T. M. Moerland, J. Broekens, and C. M. Jonker, ‘Emotion in reinforcement learning agents and robots: a survey’, Mach Learn, vol. 107, no. 2, pp. 443–480, Feb. 2018, doi: 10.1007/s10994-017-5666-0.

[21] I. Khan, M. Lewis, and L. Cañamero, ‘Modelling the Social Buffering Hypothesis in an Artificial Life Environment’, in *The 2020 Conference on Artificial Life*, Online, 2020, pp. 393–401. doi: 10.1162/isal_a_00302

[22] I. Khan, M. Lewis, and L. Cañamero, ‘Adaptation and the Social Salience Hypothesis of Oxytocin: Early Experiments in a Simulated Agent Environment’, in *Proc. 2nd Symposium on Social Interactions in Complex Intelligent Systems (SICIS)*, Liverpool, UK, 2018, pp. 2–9.

[23] S. Shahid, E. Krahmer, and M. Swerts, ‘Child–robot interaction across cultures: How does playing a game with a social robot compare to playing a game alone or with a friend?’, *Computers in Human Behavior*, vol. 40, pp. 86–100, Nov. 2014, doi: 10.1016/j.chb.2014.07.043.

[24] F. P. Santos, J. M. Pacheco, A. Paiva, and F. C. Santos, ‘Evolution of Collective Fairness in Hybrid Populations of Humans and Agents’, *AAAI*, vol. 33, pp. 6146–6153, Jul. 2019, doi: 10.1609/aaai.v33i01.33016146

[25] F. Correia, S. Mascarenhas, R. Prada, F. S. Melo, and A. Paiva, ‘Group-based Emotions in Teams of Humans and Robots’, in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, Chicago IL USA, Feb. 2018, pp. 261–269. doi: 10.1145/3171221.3171252.

[26] Cañamero, L. (2005). Symposium Preface, Agents that Want and Like: Motivational and Emotional Roots of Cognition and Action. Proc. SSAISB Convention 2005, University of Hertfordshire, Hatfield, April, 2005. https://aisb.org.uk/wpcontent/uploads/2019/12/2_Agents_Final.pdf

[27] J. M. Epstein and R. Axtell, *Growing artificial societies: social science from the bottom up*. Washington, D.C: Brookings Institution Press, 1996.