Towards an Affective Model of Norm Emergence and Adaptation

Stavros Anagnou, Lola Cañamero

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).

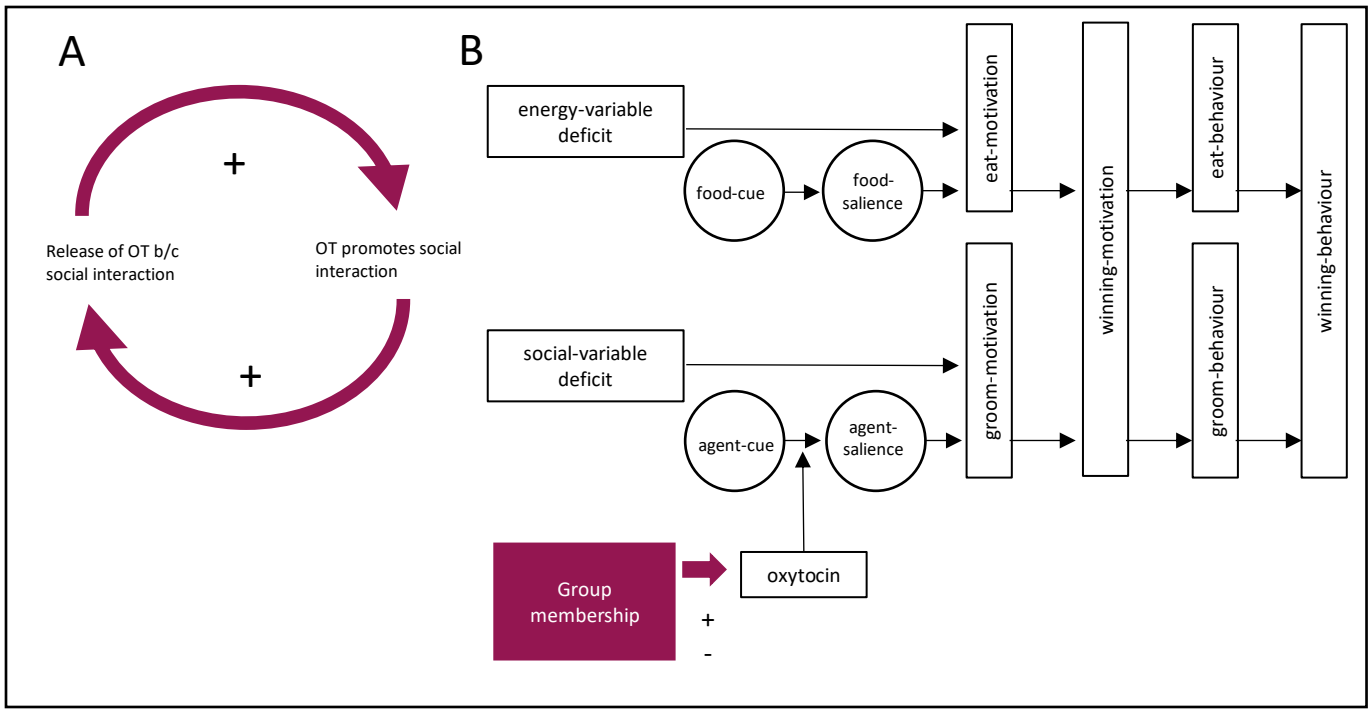

Fig. 1. A: Schematic of the Oxytocin (OT) positive feedback loop. B: Schematic of the Action-Selection Architecture (ASA). The ASA chooses behaviour based on internal needs (energy-level deficit/social-level deficit) and presence/salience of external cues that determine the strength of the motivation. 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.