Affective Computing for Human-Robot Interaction Research: Four Critical Lessons for the Hitchhiker

Hatice Gunes[†,*] and Nikhil Churamani[†]

***Abstract*— Social Robotics and Human-Robot Interaction (HRI) research relies on different Affective Computing (AC) solutions for sensing, perceiving and understanding human affective behaviour during interactions. This may include utilising *off-the-shelf* affect perception models that are pre-trained on popular affect recognition benchmarks and directly applied to situated interactions. However, the conditions in situated human-robot interactions differ significantly from the training data and settings of these models. Thus, there is a need to deepen our understanding of how AC solutions can be best leveraged, *customised* and applied for situated HRI. This paper, while critiquing the existing practices, presents *four critical lessons* to be noted by the hitchhiker when applying AC for HRI research. These lessons conclude that: (i) The six basic emotions categories are not always relevant in situated interactions, (ii) Affect recognition accuracy (%) improvement as the sole goal is inappropriate for situated interactions, (iii) Affect recognition may not generalise across contexts, and (iv) Affect recognition alone is insufficient for adaptation and *personalisation*. By describing the background and the context for each lesson, and demonstrating how these lessons have been compiled from the various studies of the authors, this paper aims to enable the hitchhiker to successfully leverage AC solutions for advancing HRI research.**

I. Introduction & Background

Social Robotics has emerged as an inherently multi-disciplinary field bringing together research efforts from Affective Computing (AC), Social Signal Processing (SSP), Computer Vision (CV), Machine Learning (ML) and Human-Robot Interaction (HRI). Yet, there is a need to develop affect sensing, perception and understanding methodologies targeted specifically to facilitate social robotics applications. To avoid re-inventing the wheel, researchers within the Human-Computer Interaction (HCI), HRI and Social Robotics fields often, and rightly so, utilise available *off-the-shelf* sensing or perception tools from other domains (such as face and gesture recognition) directly for their in-house studies, datasets and evaluations. However, these practices hinder progress leading to a lack of novel and domain-specific (affect) sensing, learning and adaptation algorithms. Furthermore, it impedes measures for reproducibility [1] due to a lack of purposeful, naturalistic and publicly available (affect) models, datasets and metrics, which are vital for comparative evaluation and gathering insights to push the field forward towards real-world adoption.

Recent research discussions[1] around situated affective computing have emphasised understanding the role of AC research, especially in situated interactions, and in realising social and affective interactions with robots. It is essential to appreciate what *does not work* when undertaking situated AC research and what *lessons* we can learn from these failures. Furthermore, linking these lessons to HRI research[2], it is critical to understand how advances in affect sensing, perception and understanding influence how individuals interact with social robots. The aim of this paper, thus, is to provide the hitchhiker with a guide for leveraging AC for HRI research, based on the critical lessons learnt, both from *successes* and *failures*, grounded in and distilled from a broader set of AC research studies we have conducted under situated interaction settings. Such a guide aims to inform the HRI community, especially the hitchhikers starting their HRI research journey, *what to be aware of* when applying AC tools for HRI research. Similar recommendations have been compiled and shared as advice to aspiring experimenters on child-robot interaction in the wild [2].

This paper discusses four *critical lessons* learnt from applying AC tools for HRI research, especially for situated interactions, compiled from the various studies of the authors. For each lesson, along with the *background* understanding, a detailed account is provided of the *context* under which the lesson is learnt, with *explanations and insights* gathered, linking it to an HRI context. These lessons are:

- **Lesson 1:** The six basic emotion [3] categories (*happiness, sadness, surprise, fear, anger* and *disgust*) are not always relevant in situated interactions;
- **Lesson 2**: Affect recognition accuracy (%) improvement as the sole goal is inappropriate for situated interaction settings;
- **Lesson 3**: Affect recognition may not generalise well across contexts (e.g., user, task, etc. - using the definition of *context* in [4]);
- **Lesson 4**: Affect recognition alone is insufficient for adaptation and personalisation.

The overall pipeline (with the different stages) of AC for HRI implementations is illustrated in Fig. 1. Lesson 1 (Section II) relates to how the user data acquired during the interactions are annotated or labelled, while Lesson 2 (Section III) relates to the robot's perception of the user. Lesson 3 (Section IV) corresponds to robot learning and Lesson 4 (Section V) corresponds to the robot's adaptation and actions.

[†] Department of Computer Science and Technology, University of Cambridge, UK. [*]Corresponding author: hatice.gunes@cl.cam.ac.uk

[1]Discussions following a keynote address at the 3rd Workshop on Applied Multimodal Affect Recognition (AMAR), International Conference on Pattern Recognition (ICPR) 2022.

[2]Discussions following workshop keynote addresses at IEEE Int'l Conference on Robot & Human Interactive Communication (RO-MAN'22), and the AAAI Fall Symposium on Artificial Intelligence for Human-Robot Interaction 2022.

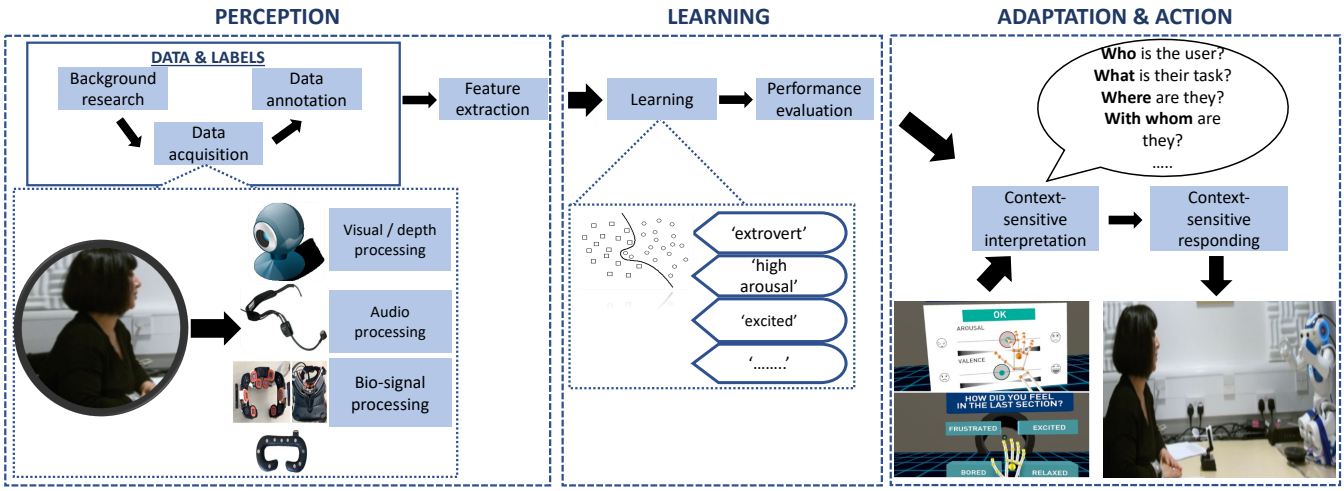

Fig. 1: The Pipeline for Affective Computing for Human-Robot Interaction implementations with marked stages of **Perception**, **Learning**, and **Adaptation & Action**.

Section VI summarises the contributions of this paper as well as reflects upon the need for a critical review of existing AC solutions of affective HRI studies.

II. Lesson 1: Six Basic Emotion Categories are Not Always Relevant

A. Background

Within the pipeline of creating an automatic affect recogniser, this lesson relates to the aspect of *affect annotations and labels* (see *Data & Labels* under Fig. 1: *Perception*). When researchers purchase or acquire commercial social robots, these robots come with black-box perception capabilities, one of which is usually proudly claimed to enable 'automatic emotion recognition' for the robot. For instance, one of the features listed for Pepper Robot is 'recognising emotions on your face'[3]. In such robotic platforms, this means the recognition of the six basic emotion categories, namely, (neutral+) happiness, sadness, surprise, fear, anger and disgust [3]. However, this works only partially and under *posed* expression settings when operating in lab environments with little-to-no variation in lighting settings [5]. Real-world interactions are much more complex, resulting in the robot struggling to accurately capture individual expressions, as demonstrated via interactive public demonstrations [6]. But even then, it is important to evaluate and understand what such categorization of user affective behaviour means for situated human-robot interactions.

B. Context

Much of AC research employs ML-based automatic affect recognition models trained and benchmarked on publicly available datasets, acquired outside of situated interaction settings. For example, most CV models undertaking the task of Facial Expression Recognition (FER) are trained on static images crawled from the internet, with cropped facial regions where the situational context information has been removed, with crowdsourced labels mostly corresponding to the aforementioned six basic emotion categories [7]. As soon as these models are embedded in robotic systems for realistic applications including tutoring and learning, assistance with rehabilitation or physical and mental health, these models cannot cope with the variation and noise in the input data that they have not encountered in their training. This results in FER often failing in situated human-robot interactions, making the robots' sensing and perception of human affect unreliable at the least [8].

Beyond automatic recognition, the six basic emotion categories are widely used in various HRI studies, even when these labels do not seem relevant for real-world contexts. One recent example investigating emotion perception using the basic emotion categories is an HRI-based rehabilitation scenario, as this is expected to improve the experience of patients [9]. In this study, a robotic arm is used to investigate whether and how it can communicate an emotional state through movements and whether people can attribute these movements to the intended emotional state. It found that *happiness* was identified well, but not *sadness* and *anger*. However, going beyond these findings, it is important to understand *'What does it mean for a robotic arm to display anger?'*. Also, *'How useful are basic emotion categories for rehabilitation robotics?'* and *'What implications does this have for HRI, in general?'*.

C. Lesson & Insights

A critical evaluation of the questions posed above requires a deeper and fundamental understanding of the situational and contextual attributes that determine human behaviour during interactions. One needs to go beyond the six basic emotion categories [10] and start exploring other affect and emotion models and instruments, while also considering how to use these contemporary models throughout the entire pipeline of study design, data acquisition, data annotation, and training and evaluation of ML models. In doing so, it is essential to start with fundamental questions, such as

[3] https://www.gwsrobotics.com/why-pepper-robot

'Which emotion or affect model is best suited to represent human behaviour and how do we decide this?' Additionally, it is also important to consider *'Whether we are taking into account situational or contextual aspects?'*.

Two contemporary instruments that can be used, instead of the six basic categories of emotions, are the Self-Assessment Manikin (SAM) [11], [12] and the Geneva Emotion Wheel (GEW) [13]. SAM is a picture-based questionnaire to independently evaluate the affect dimensions of *arousal* (activation), *valence* (pleasure) and *dominance* (sense of control), and it can be used for subjective assessment of participant/user affective responses [12]. The GEW, on the other hand, has been proposed as 'a theoretically derived and empirically tested instrument to measure emotional reactions to objects, events, and situations' [13]. The participant/user can indicate the emotion they experienced by choosing a single emotion with the corresponding intensity or a blend of multiple emotions (out of 20 emotion families). Robotics and HRI researchers have started to successfully use SAM and GEW in their works, for example, to evaluate patients' emotions induced by a robotic hand rehabilitation platform [14], to classify the expression of emotion on robots [15] and to measure perceived affect in HRI [16].

In the context of dyadic human-human interactions *vs.* human-agent interactions, Song et al. [17] report that facial reaction prediction and personality recognition performance for ML models are better for human-human interaction data. This finding indeed has implications for HRI research and brings forth further questions that, as a community, we would need to investigate. These include, but are not limited to, *'Do we display affect differently in HRI?'*, and *'Do we need different affect or emotion models for HRI that capture both qualitative and quantitative aspects of human as well as robot behaviours?'*. In order to answer these questions, a promising direction is to take a data-driven approach, similar to the pioneering study by Jam et al. [18] that aims at developing a data-driven categorical taxonomy of emotional expressions in real-world HRI.

III. Lesson 2: Affect recognition accuracy (%) improvement as the sole goal is inappropriate

A. Background

Within the pipeline of creating an automatic affect recogniser, this lesson relates to the aspect of *affect sensing* (see *Performance Evaluation* under Fig. 1). The majority of the work towards automatic affect recognition focuses on achieving results that are considered 'excellent' or 'very good' in terms of the evaluation metric used. For many researchers 'success' is then equivalent to either obtaining a recognition accuracy of $>= 75\%$ on a dataset that perhaps other researchers have not yet worked or published on, or improving the state-of-the-art (SOTA) recognition accuracy by $>= 2-3\%$ on a benchmark that others have widely reported on to be able to claim that their method is 'better' than the current SOTA results. However, benchmark datasets, even the ones that claim to be obtained *in-the-wild*, are usually stripped of context. Such datasets, for instance, contain static facial images or even videos of people without much interaction taking place. When we move away from recognising affect on such *in-the-wild* but idealised benchmark datasets to actual interaction studies with humans, we are faced with a much higher level of complexity.

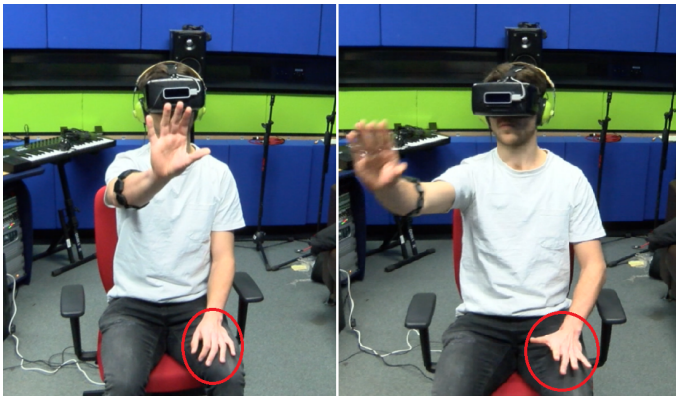

Fig. 2: Illustrating the differences in participants' *left hand* when playing the 'Memory Break' game in Virtual Reality (VR) at *Level 1 - Easy* (left) *vs. Level 3 - Hard* (right).

B. Context

To exemplify how relying on affect recognition accuracy (%) improvements may be unimportant and insufficient in HRI context, we look at *'Gamified Cognitive Training'*, as an example, as it relates to one of our study [19] undertaken in $2016-2017$. This study investigated how the affect dimensions of arousal and valence were linked to Working Memory (WM) performance of 30 participants when playing a custom video game, 'Memory Break', on Desktop *vs.* in Virtual Reality (VR), in two separate sessions, one for each interaction mode. Both game modes were designed to have three difficulty levels to evoke different levels of *arousal* while maintaining the same memory load. The WM capacity baseline of participants were measured using relevant measures while the participants self-reported their affective states and completed the Game Experience Questionnaire (GEQ) [20]. Our analyses showed an improvement in participants' WM performance when playing in VR mode, with a significant effect in those with a low WM capacity. Significantly higher levels of *valence* and *arousal* were self-reported when playing the VR version of the game.

To sense the participants' affective states, a heart-rate sensor was attached to their chest recording their heart activity and an Electromyography (EMG) armband was placed on the forearm that was used for interacting with the game environment. However, we had missed one important factor. As seen in Fig. 2, when the difficulty level of the game increased to 'hard' (Level 3), the tension was clearly observable on their hand that was resting on their lap. Post-study, we observed this to be the trend for all participants. Unfortunately, that hand did not have any sensor placed on it to measure the tension manifested, which meant we missed crucial information that could aid the recognition of participants' arousal and valence. Despite extracting features from other

Fig. 3: Predictions of (deep) ML models trained on publicly available facial expression, facial affect and facial Action Unit (AU) datasets when used on data acquired under work-like settings and tasks.

sensors and experimenting with various ML techniques for classifying arousal and valence, the recognition results did not look promising. Ultimately, accuracy (%) improvement in this context was unimportant because we were not measuring and analysing the most relevant signals. The reliability of *off-the-shelf* affect sensing solutions may also be a concern in situated interactions, potentially resulting in poor model accuracy [21]. Thus, it is important to look beyond model accuracy in situated interactions.

C. Lesson & Insights

The lesson that can be learnt from the 'Gamified Cognitive Training' study (and other relevant ones) is that *accuracy* or *accuracy improvements (%) as the sole goal or objective are inappropriate and insufficient*, especially when we are not reliably capturing and analysing the most relevant signals and cues. Expressly, undertaking human studies in situated interactions, where we aim to sense affect, requires several pilot study iterations, until we are sure about where to place the different sensors, measuring the right signals and cues related to the affect displayed, and ascertain the reliability of the *off-the-shelf* models and other tools used. In other words, we should *study the expressions and display of affect before we sense them*. This is mainly due to two reasons. Firstly, in the human-machine interaction context, humans shape machine behaviour and vice-versa [21], [22]. This often results in the emergence of new human behaviour, unforeseen in the original study or interface design. Secondly, when analysing affect and emotions, there is, at times, non-verbal and emotional *leakage*. At times, inner feelings of a person might be revealed or expressed more intensely in a certain modality or channel [23], [24] (usually the less dominant one), which might be different from the one observers focus on, for example, controlling what is being said while expressing differently through nonverbal behaviour. In light of these, in situated interactions where we aim to measure and analyse socio-emotional behaviours, we need to reflect on critical questions such as *'Are we placing the sensors in the right places?'*, *'Are we measuring the most relevant signals?'* and *'Are the tools used to sense the signals, reliable?'*.

With these aspects in mind, a possible direction for HRI research can be to adopt rich multi-modal sensing, not necessarily to improve accuracy, but also to ensure that different aspects of user affect and behaviour manifestations are captured and investigated. This is particularly important for emerging research areas that cannot simply rely on previous research findings. For instance, mental wellbeing evaluation in children via child-robot interactions [25] requires an investigation of different aspects of child multi-modal behaviour (questionnaire responses, free-from speech content, nonverbal head, face or audio behaviours and physiological reactions), going beyond what children report or say. However, it is important to note that not many social robots are equipped with high-resolution sensing or have the capabilities to enable such rich multi-modal perception. A possible solution may be to create 'hacks', for example, by 3D printing and additional sensor placement (see [26] on how a 3D printed headset is used with high-resolution cameras).

IV. Lesson 3: Affect Recognition May Not Generalise Well Across Contexts

A. Background

Openly available facial affect datasets used for training FER models generally contain displays of young and middle-aged adults. Facial affect data from sensitive user groups such as children, adolescents, and older adults are relatively less accessible due to various challenges including ethical and privacy concerns. This imbalance in data causes these models to not generalise well on other user groups such as the elderly [27] or children [28] and, in turn, results in biased algorithms for facial affect analysis and prediction. In addition to encoding demographic bias, currently available facial expression datasets are also biased towards certain affect labels such as 'neutral', 'anger' and 'happiness', compared to other affective states such as 'annoyance' [29]. Thus, models trained on most common benchmark datasets for facial affect recognition are: (i) more accurate for young and middle-aged adults; and (ii) mostly predicting affect in terms of basic emotion categories; despite the fact that this might not fit well the application context [29].

B. Context

To appreciate the challenges relating to applying *generalised* affect recognition models, it is important to consider

how these models may perform with under-represented (in traditional affect perception benchmarks) populations. In this context, the EU Horizon 2020 WorkingAge[4] project is aimed at studying and promoting healthy habits in working environments, focusing on people aged over 45. By gathering a better understanding of wellbeing at work and of factors that may inhibit or deteriorate prolonged employment, it created an integrated digital solution, the WorkingAge of Wellbeing (WAOW) Tool, to support workers' wellbeing in three types of working environments: office, teleworking, and manufacturing. Within the WAOW Tool pipeline of creating an automatic system that analyses worker psycho-social conditions, worker physical conditions and the working environment, and personalises via appropriate recommendations, customisable by the user [30], this lesson relates to the aspect of *affect sensing* and *recognition* (see Fig. 1: Perception).

As part of the WorkingAge project, we first introduced a multi-site data collection protocol for acquiring human behavioural data under simulated working conditions with three work-like tasks: the N-back task, the video conference task and the operation game. With this, we acquired the first human working facial behaviour dataset called *WorkingAge DB* [31] which was collected across four different sites in Europe. Implementing (deep) ML models (e.g., ResNet-50), trained on publicly available facial expression (e.g., RAF-DB [32]), facial affect (e.g., AffectNet [33]) and facial AU (e.g., BP4D [34]) datasets, and applying these models on facial data acquired under work-like settings and tasks, results in evaluations similar to those illustrated in Fig. 3. It can be clearly seen that such models have no knowledge about context, and provide labels such as 'surprise' and 'negative valence' when the person is focused on the task.

Having seen these results, we decided to train ML models specifically with the data acquired in work-like settings. Thus, we implemented and compared a set of (deep) ML methods using the facial data from WorkingAge DB for automatic prediction of worker periodical facial affect while also investigating how task type, recording site, gender, and feature representations affected model performance [31]. Our results showed that worker affect can be inferred from their facial behaviours using data acquired in work-like settings, and models pre-trained on naturalistic datasets are useful for prediction but are insufficient on their own. Context, specifically the task type and task setting, influenced the affect recognition performance.

C. Lesson & Insights

HCI and HRI studies are prone to adopting *off-the-shelf* affect recognition toolkits, that are pre-trained on publicly available benchmark datasets, as means to an end, for the quick modelling of user affective behaviour. For instance, in the HRI context, Mathur et al. [35] investigated how to model user empathy elicited by a robot storyteller. For this, they employed an open-source off-the-shelf toolkit (OpenFace 2.2.0 [36]) that is widely used by various researchers within the AC, HRI and HCI communities. OpenFace enables the extraction of eye gaze directions, the presence (and intensity) of 17 facial AUs, facial landmarks and head pose coordinates, amongst other features. However, as we learnt from the WorkingAge study [31], for facial affect recognition in specific contexts such as work-like settings, we cannot simply rely on generic off-the-shelf toolkits. Such models are ignorant of context and will not generalise well to real-world settings where many factors (such as ethnic or cultural background, gender, age, and the task, amongst others) influence human expressivity and nonverbal behaviour.

Thus, when analysing human affective behaviour using off-the-shelf toolkits and models, several critical questions need to be considered. These include, but are not limited to: *'How well are we taking into account the contextual aspects of the interaction?'*, and *'Are we considering person-specific aspects impacting the interacitons?'*. To address these questions, we need to focus on *personalisation* rather than generalisation, considering person-specific aspects when modelling user affective behaviour. For example, [37] presents a personalised learning companion that uses children's verbal and nonverbal affective cues to modulate their engagement levels. Facial features extracted using the off-the-shelf Affdex toolkit [38] are used for arousal prediction which, in turn, defines state space features for an Reinforcement Learning (RL)-based personalisation algorithm. More recently, in [39] we introduced and adapted the Continual Learning (CL) paradigm for Affective Robotics where a robot acquires and integrates knowledge incrementally about changing data conditions, and showed how it can be utilised in practice for adaptive HRI [40]. Furthermore, the series of LEAP-HRI[5] workshops, that we have been co-organizing since 2021, also emphasised the need to move away from generalisation and focus more on lifelong learning and personalisation, particularly when it comes to long-term HRI where novelty effect is no longer present [41].

Additionally, we also need to consider other relevant questions such as *'Are we investigating for whom the trained models work well, and why?'*, *'How do these models work for specific user groups like children and elderly?'*, and *'How to ensure that predictions from these models are not biased?'*. To date, there are many publicly available benchmarks for expression and affect recognition, however, none of these datasets have been acquired considering a *fair distribution* across the human population. Recent studies on a number of publicly available benchmark datasets such as RAF-DB [32] and CelebA [42] have shown that ML models for FER trained on such datasets are biased [43]. Despite several bias mitigation strategies [28], [43], [44] for addressing bias in the AC context (see [29] for a review), how bias in affect prediction models impacts HRI and user experience, engagement and trust, and how to achieve *fairer affective robotics* remain open research problems that need multi-disciplinary community efforts at the level of datasets, annotations, benchmarking and reproducibility.

[4] https://www.workingage.eu/

[5] https://leap-hri.github.io/

V. Lesson 4: Affect recognition alone is insufficient for adaptation and personalisation

A. Background

Within the pipeline for creating an automatic affect recogniser, this lesson relates to the aspect of *adaptation* (see Fig. 1: Adaptation & Action). Affect recognition is only one of the affective cognitive architecture modules for achieving emotionally intelligent autonomous robots that are capable of perception, learning, action, adaptation, and even anticipation [45]. One of the most common techniques for robot learning and adaptation is learning with the human-in-the-loop, or Interactive Reinforcement Learning (IRL), that focuses on sensing and incorporating user interactive (verbal, social and affective) feedback [46]–[52]. IRL with explicit feedback can be challenging as humans tend to provide more positive than negative feedback, at times ignoring the robots' mistakes. With progressing interactions, the frequency of human feedback may decrease [53]. Therefore, using implicit feedback, such as facial affect, can be more effective as the human "teacher" will be less conscious of providing the feedback and will be less likely to suffer from feedback fatigue [49]. Studies on IRL demonstrate the growing potential of sensing and utilising implicit human behavioural cues for training robots through natural interactions and shaping their behaviour in real-time. But *'is adaptation based on affect sufficient and does it always improve HRI experience?'*

B. Context

For naturalistic HRI, especially facilitating social interactions, it is imperative that robots are able to sense and adapt towards human behaviour, not only regarding individual responses as feedback on their actions but also as motivation for learning context-appropriate behaviours [52]. In this context, in [51], we explored learning socially appropriate Robo-waiter behaviours through real-time user feedback. This feedback was driven by either an implicit reward (calculated by observing participants' facial affect) or an explicit reward (incorporating their verbal responses). First, a dataset was acquired with crowd-sourced labels to learn appropriate approach behaviours for a robo-waiter based on its positioning and movement. This dataset was then used to pre-train an RL agent which, later, was extended under IRL settings to include implicit and explicit rewards, allowing for real-time adaptation from user social feedback. The approach was evaluated using a within-subjects HRI study with 21 participants with the results showing that both the explicit and implicit feedback mechanisms enabled an *adaptive robo-waiter* that was rated as more *enjoyable* and *sociable* compared to the robot implementing the pre-trained model or using a random control policy. The adaptive robo-waiter also rendered more *appropriate positioning* relative to the participants. Additionally, *adaptability* ratings showed the explicit feedback condition as the most preferred condition with the robot being rated significantly higher in terms of *understanding and adapting* to what the participant *said*. These results clearly show that for task-based interactions (as in the robo-waiter context), adaptation based on affect alone is insufficient, and we do need to take into account explicit, task-related user feedback. Combining explicit and implicit feedback to shape the reward function, although not explored in this study, has the potential to further improve user interaction experience.

C. Lesson & Insights

The creation of closed-loop affective robots that can undertake successful social interactions with humans requires that these robots keep learning in a lifelong manner and continually adapt towards user behaviour and their affective states [39]. Traditional ML approaches do not scale well to the dynamics of such real-world interactions because they assume stationarity in data conditions and distributions, but real-world contexts change continuously. Also, training data and learning objectives relevant to HRI may change rapidly.

In [39], we provide guidelines on how to utilise Continual Learning (CL) for personalised affect perception and context-appropriate behavioural learning for affective robotics. These guidelines enable CL-based *personalisation* in the context of robotic wellbeing coaching in [40], where a user study is conducted with 20 participants comparing static and scripted interactions with using affect-based adaptation without and with *continual personalisation*. The results showed that participants indicate a clear preference for a robotic coach with *continual personalisation* capabilities, with significant improvements observed in the robot's *anthropomorphism, animacy* and *likeability* ratings. Additionally, the robot is also rated as significantly better at understanding how the participants felt during the interactions.

Although affective adaptation is a desirable capability for social robots, we need to bear in mind that it may not always work, and at times may even hinder interactions. For example, Kennedy et al. [54] investigated the effect of a social robot tutoring strategy, with and without social and adaptive behaviours, in the context of children learning about prime numbers. Their results showed *no significant learning outcome* for children interacting with a robot using social and adaptive behaviours in addition to the teaching strategy. Therefore researchers should be cautious about the specific context they have at hand when deciding to apply social and adaptive behaviours to a robot, and whether these interactions are longitudinal or one-off. Gao et al. [55] also investigated the effects of robot behaviour personalisation on user's task performance in the context of robot-supported learning. They utilised RL for personalisation, enabling a robot tutor to select verbal supportive behaviours to maximise the user's task progress and positive reactions. Their results showed that participants were more efficient at solving logic puzzles and preferred a robot that exhibits more varied behaviours compared to a robot that personalises its behaviour by converging on a specific one over time. Overall, adaptation and personalisation based on affective or social behaviours needs further investigation, to gather insights on the impact of the context and the nature of the interaction (e.g., task-based *vs.* free-flow).

TABLE I: Four *Critical Lessons* with the *Reflective Questions* the hitchhiker needs to probe when applying Affective Computing (AC) solutions for Human-Robot Interaction (HRI) research under situated interaction settings.

Lesson	Reflective Questions
Lesson 1: The six basic emotion categories are not always relevant in situated interactions.	1) How is individual affective behaviour manifested? 2) Which affect model can best represent an individual's affective states?
Lesson 2: Affect recognition accuracy (%) improvement as the sole goal is inappropriate for situated interactions.	1) Are we placing the sensors in the right places? 2) Are we measuring the most relevant signals? 3) How can we interpret model performance in view of users' interaction experiences?
Lesson 3: Affect recognition may not generalise (well) across contexts.	1) Are we taking into account contextual and person-specific aspects? 2) Are we investigating for whom the trained models work well, and why? 3) How do model predictions work for specific user groups like children and elderly? 4) Are there any strategies in place to mitigate prediction bias in models?
Lesson 4: Affect recognition alone is insufficient for adaptation and personalisation.	1) Can the users' responses be summarised only using their affective behaviour? 2) Does the user provide additional feedback that may be helpful for robot learning? 3) Is personalisation required and/or appropriate for the context of the interaction?

VI. Summary and Conclusion

This paper, reflecting upon the pitfalls and limitations of current AC solutions for situated HRI, presents *four critical lessons* for the hitchhiker starting their HRI research journey (see Table I). These lessons, compiled from and learnt through the various studies of the authors, as well as existing literature, aim to distill key challenges that need the focus and attention of the HRI community. Specifically critiquing the use of AC solutions for sensing, perceiving and understanding user affective behaviour in situated interactions, these lessons highlight the challenges and limitations of existing methodologies and how the hitchhiker needs to be cautious of utilising *off-the-shelf* solutions, designed only for a generalised application that may not be appropriate and efficient for situated interactions. For each lesson, we present a detailed account of the background and motivation for why it is relevant and provide contextual understanding of how it relates to pitfalls, in practice, when applying *off-the-shelf* AC solutions directly for situated HRI studies. Furthermore, we also summarise our learning from these experiences and highlight key considerations for the hitchhiker to bear in mind. Our proposition, with this paper, is for the hitchhiker to reflect upon and probe specific questions, pertaining to each lesson, before designing HRI studies that depend upon an affective evaluation of user behaviour. Furthermore, we also aim to encourage other affective computing and human-robot interaction researchers to contribute to the scientific community at large by critically analysing and reflecting upon *'what does not work and why?'* in HRI studies, and sharing their perspectives for everyone's benefit. Such a reflection and consideration will no doubt contribute to more successful and fruitful implementations of AC solutions in situated interaction studies, *creating a new bridge* for HRI.

Acknowledgements

Open Access: For the purpose of *open access*, the authors have applied a Creative Commons Attribution (CC BY) license to any Accepted Manuscript version arising. **Data Access:** This work involves secondary analyses of existing studies, described and cited in the text. **Funding:** This work is supported in part by the EPSRC/UKRI under grant ref. EP/R030782/1, and in part by Google under the GIG Funding Scheme.

References

[1] H. Gunes, F. Broz, C. Crawford, A. R. der Pütten, M. Strait, and L. Riek, "Reproducibility in human-robot interaction: Furthering the science of hri," *Current Robotics Reports*, 2022.

[2] R. Ros, M. Nalin, R. Wood, P. Baxter, R. Looije, Y. Demiris, T. Belpaeme, A. Giusti, and C. Pozzi, "Child-robot interaction in the wild: Advice to the aspiring experimenter," in *Proc. of Int'l. Conf. on Multimodal Interfaces*, 2011, p. 335–342.

[3] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *J. of Personality & Social Psychology*, vol. 17, no. 2, p. 124, 1971.

[4] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1743–1759, 2009.

[5] N. Rawal and R. M. Stock-Homburg, "Facial emotion expressions in human-robot interaction: A survey," *Int'l. Journal of Social Robotics*, vol. 14, no. 7, pp. 1583–1604, Jun. 2022.

[6] H. Gunes, O. Celiktutan, and E. Sariyanidi, "Live human–robot interactive public demonstrations with automatic emotion and personality prediction," *Phil. Trans. R. Soc. B*, vol. 374, no. 1771, pp. 1–8, 2019.

[7] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. on Affective Computing*, vol. 13, no. 3, pp. 1195–1215, 2022.

[8] M. Spezialetti, G. Placidi, and S. Rossi, "Emotion recognition for human-robot interaction: Recent advances and future perspectives," *Frontiers in Robotics and AI*, vol. 7, 2020.

[9] C. L. Viola, L. Fiorini, G. Mancioppi, J. Kim, and F. Cavallo, "Humans and robotic arm: Laban movement theory to create emotional connection," in *Proc. of IEEE Int'l Conf. on Robot and Human Interactive Communication (RO-MAN)*, 2022, pp. 566–571.

[10] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions: An emerging approach," *IEEE Trans. on Affective Computing*, vol. 12, no. 1, pp. 16–35, 2021.

[11] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *J. of Behavior Therapy & Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[12] I. Bakker, T. van der Voordt, P. Vink, and J. de Boon, "Pleasure, arousal, dominance: Mehrabian and russell revisited," *Curr. Psychol.*, vol. 33, pp. 405–421, 2014.

[13] V. Sacharin, K. Schlegel, and K. R. Scherer, *Geneva emotion wheel rating study*. Center for Person, Kommunikation, Aalborg University, NCCR Affective Sciences, Aalborg, 2012.

[14] A. Cisnal, V. Moreno-SanJuan, J. Fraile, J. Turiel, E. de-la Fuente, and G. Sánchez-Brizuela, "Assessment of the patient's emotional response with the robhand rehabilitation platform: A case series study," *J. Clin. Med.*, vol. 11, no. 4442, pp. 1–15, 2022.

[15] C. McGinn and K. Kelly, "Using the geneva emotion wheel to classify the expression of emotion on robots," in *Companion of the ACM/IEEE Int'l Conf. on Human-Robot Interaction*, 2018, p. 191–192.

[16] A. K. Coyne, A. Murtagh, and C. McGinn, "Using the geneva emotion wheel to measure perceived affect in human-robot interaction," in *Proc. of ACM/IEEE Int'l. Conf. on Human-Robot Interaction*, 2020, p. 491–498.

[17] S. Song, Z. Shao, S. Jaiswal, L. Shen, M. Valstar, and H. Gunes, "Learning person-specific cognition from facial reactions for automatic personality recognition," *IEEE Trans. on Affective Computing*, pp. 1–18, 2022.

[18] G. Saheb Jam, J. Rhim, and A. Lim, "Developing a data-driven categorical taxonomy of emotional expressions in real world human robot interactions," in *Companion of ACM/IEEE Int'l Conf. on Human-Robot Interaction*, 2021, p. 479–483.

[19] D. Gabana, L. Tokarchuk, E. Hannon, and H. Gunes, "Effects of valence and arousal on working memory performance in virtual reality gaming," in *Proc. of Int'l Conf. on Affective Computing and Intelligent Interaction (ACII)*, 2017, pp. 36–41.

[20] W. A. IJsselsteijn, Y. A. de Kort, and K. Poels, "The game experience questionnaire," *Eindhoven: Technische Universiteit Eindhoven*, vol. 46, no. 1, 2013.

[21] D. Jirak, M. Aoki, T. Yanagi, A. Takamatsu, S. Bouet, T. Yamamura, G. Sandini, and F. Rea, "Is it me or the robot? a critical evaluation of human affective state recognition in a cognitive task," *Frontiers in Neurorobotics*, vol. 16, 2022.

[22] I. Rahwan, M. Cebrian, N. Obradovich, and et al., "Machine behaviour," *Nature*, vol. 568, pp. 477–486, 2019.

[23] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," *Journal for the Study of Interpersonal Processes*, vol. 32, no. 1, pp. 88–106, 1969.

[24] P. H. Waxer, "Nonverbal cues for anxiety: An examination of emotional leakage," *Journal of Abnormal Psychology*, vol. 86, no. 3, pp. 306–314, 1977.

[25] N. I. Abbasi, M. Spitale, J. Anderson, T. Ford, P. B. Jones, and H. Gunes, "Can robots help in the evaluation of mental wellbeing in children? an empirical study," in *Proc. of IEEE Int'l Conf. on Robot & Human Interactive Communication (RO-MAN)*, 2022, pp. 1459–1466.

[26] P. Bremner, O. Çeliktutan, and H. Gunes, "Personality perception of robot avatar teleoperators in solo and dyadic tasks," *Frontiers Robotics AI*, vol. 4, p. 16, 2017.

[27] K. Ma, X. Wang, X. Yang, M. Zhang, J. M. Girard, and L.-P. Morency, "Elderreact: A multimodal dataset for recognizing emotional response in aging adults," in *Proc. Int'l. Conf. on Multimodal Interaction*, 2019, p. 349–357.

[28] A. Howard, C. Zhang, and E. Horvitz, "Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems," in *IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*, 2017, pp. 1–7.

[29] J. Cheong, S. Kalkan, and H. Gunes, "The hitchhiker's guide to bias and fairness in facial affective signal processing: Overview and techniques," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 39–49, 2021.

[30] R. M. R. de Almeida, A. G. Aberturas, Y. B. Aguado, M. Atzori, A. Barenghi, G. Borghini, C. A. C. Ortega, S. Comai, R. L. Durán, M. Fugini, H. Gunes, B. M. Lancis, G. Pelosi, V. Ronca, L. Sbattella, R. Tedesco, and T. Xu, "Decision support systems to promote health and well-being of people during their working age: The case of the workingage eu project," in *Embedded Computer Systems: Architectures, Modeling, and Simulation*, 2020, pp. 336–347.

[31] S. Song, Y. Luo, V. Ronca, G. Borghini, H. Sagha, M. A. Rick, V., and H. Gunes, "Deep learning-based assessment of facial periodic affect in work-like settings," in *European Conf. on Computer Vision (ECCV) Workshops*, 2022, pp. 1–16.

[32] S. Li and W. Deng, "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition," *IEEE Trans. on Image Processing*, vol. 28, no. 1, pp. 356–370, 2019.

[33] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," *IEEE Trans. on Affective Computing*, vol. 10, pp. 18–31, 2019.

[34] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014.

[35] L. Mathur, M. Spitale, H. Xi, J. Li, and M. J. Mataric, "Modeling user empathy elicited by a robot storyteller," in *Proc. of Int'l. Conf. on Affective Computing & Intelligent Interaction (ACII)*, 2021, pp. 1–8.

[36] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," in *Proc. of IEEE Int'l Conf. on Automatic Face Gesture Recognition (FG)*, 2018, pp. 59–66.

[37] H. W. Park, I. Grover, S. Spaulding, L. Gomez, and C. Breazeal, "A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education," in *Proc. of the AAAI Conference on Artificial Intelligence*, 2019, pp. 687–694.

[38] D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot, and R. e. Kaliouby, "AFFDEX SDK: A Cross-Platform Real-Time Multi-Face Expression Recognition Toolkit," in *Proc. of CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2016, p. 3723–3726.

[39] N. Churamani, S. Kalkan, and H. Gunes, "Continual learning for affective robotics: Why, what and how?" in *Proc. of IEEE Int'l Conf. on Robot and Human Interactive Communication (RO-MAN)*, 2020, pp. 425–431.

[40] N. Churamani, M. Axelsson, A. Caldir, and H. Gunes, "Continual Learning for Affective Robotics: A Proof of Concept for Wellbeing," in *Proc. Int'l. Conf. on Affective Computing & Intelligent Interaction Workshops & Demos (ACIIW)*, 2022.

[41] B. Irfan, A. Ramachandran, S. Spaulding, G. I. Parisi, and H. Gunes, "Lifelong learning and personalization in long-term human-robot interaction (LEAP-HRI)," in *Proc. of ACM/IEEE Int'l. Conf. on Human-Robot Interaction (HRI)*, 2022, pp. 1261–1264.

[42] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. of Int'l. Conf. on Computer Vision (ICCV)*, 2015.

[43] T. Xu, J. White, S. Kalkan, and H. Gunes, "Investigating bias and fairness in facial expression recognition," in *Proc. of European Conf. on Computer Vision (ECCV) Workshops*, 2020, pp. 506–523.

[44] N. Churamani, O. Kara, and H. Gunes, "Domain-incremental continual learning for mitigating bias in facial expression and action unit recognition," *IEEE Trans. on Affective Computing*, pp. 1–15, 2022.

[45] A. Tanevska, F. Rea, G. Sandini, and A. Sciutti, "Designing an affective cognitive architecture for human-humanoid interaction," in *Companion of ACM/IEEE Int'l. Conf. on Human-Robot Interaction*, 2018, pp. 253–254.

[46] I. Leite, G. Castellano, A. Pereira, C. Martinho, and A. Paiva, "Empathic robots for long-term interaction," *Int'l Journal of Social Robotics*, vol. 6, no. 3, pp. 329–341, 2014.

[47] G. Li, H. Dibeklioğlu, S. Whiteson, and H. Hung, "Facial feedback for reinforcement learning: a case study and offline analysis using the tamer framework," *Autonomous Agents and Multi-Agent Systems*, vol. 34, no. 22, pp. 1–29, 2020.

[48] Y. Cui, Q. Zhang, A. Allievi, P. Stone, S. Niekum, and W. Knox, "The empathic framework for task learning from implicit human feedback," in *Conference on Robot Learning*, 2020.

[49] J. Lin, Z. Ma, R. Gomez, K. Nakamura, B. He, and G. Li, "A review on interactive reinforcement learning from human social feedback," *IEEE Access*, vol. 8, pp. 120 757–120 765, 2020.

[50] S. Gillet, M. T. Parreira, M. Vázquez, and I. Leite, "Learning gaze behaviors for balancing participation in group human-robot interactions," in *Proc. of ACM/IEEE Int'l. Conf. on Human-Robot Interaction*, 2022.

[51] E. McQuillin, N. Churamani, and H. Gunes, "Learning socially appropriate robo-waiter behaviours through real-time user feedback," in *Proc. of ACM/IEEE Int'l Conf. on Human-Robot Interaction*, 2022.

[52] N. Churamani, P. Barros, H. Gunes, and S. Wermter, "Affect-driven learning of robot behaviour for collaborative human-robot interactions," *Frontiers in Robotics and AI*, vol. 9, 2022.

[53] J. MacGlashan, M. K. Ho, R. Loftin, B. Peng, G. Wang, D. L. Roberts, M. E. Taylor, and M. L. Littman, "Interactive learning from policy-dependent human feedback," in *Proc. of Int'l. Conf. on Machine Learning*, 2017, pp. 2285–2294.

[54] J. Kennedy, P. Baxter, and T. Belpaeme, "The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning," in *Proc. of ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI)*, 2015, pp. 67–74.

[55] Y. Gao, W. Barendregt, M. Obaid, and G. Castellano, "When robot personalisation does not help: Insights from a robot-supported learning study," in *Proc. IEEE Int'l Conf. on Robot & Human Interactive Communication (RO-MAN)*, 2018, pp. 705–712.