What Could a Body Tell a Social Robot that It Does Not Know?

Dennis Küster and Arvid Kappas
School of Humanities and Social Sciences, Jacobs University Bremen, Campus Ring 1, 28759 Bremen, Germany

Keywords: Social Robots, Psychophysiology, Emotion, Measurement, Taxonomy.

Abstract: Humans are extremely efficient in interacting with each other. They not only follow goals to exchange information, but modulate the interaction based on nonverbal cues, knowledge about situational context, and person information in real time. What comes so easy to humans poses a formidable challenge for artificial systems, such as social robots. Providing such systems with sophisticated sensor data that includes expressive behavior and physiological changes of their interaction partner holds much promise, but there is also reason to be skeptical. We will discuss issues of specificity and stability of responses with view to different levels of context.

1 INTRODUCTION

Early models of human communication were driven by the notion that information is encoded, sent as a message, and decoded on the receiver side. In other words, all the action is happening in the message. If it is well encoded, everything the receiver needs to know is contained and a successful decoding completes a felicitous communication episode (Rosengren, 2000). This of course conjures up the notion of a complete encapsulation in packages of speech that efficiently transport information from one person to another. We now know that this is not how human communication works. Not only, are verbal messages augmented by information transported via multiple nonverbal channels, but also the personal context of all participating individuals, as well as the situational context, play an important role in connecting interactants. When it comes to the communication of emotion, one also has to consider that a considerable amount of information is transmitted outside of conscious awareness, for example via mirror and feedback processes that are difficult to describe and assess. It is this complexity that makes it often enough difficult for humans to communicate successfully, but the challenge of creating artificial systems that succeed is at times daunting.

Consider three examples:

Jill: “Could you pass the butter?” – John passes the butter (1)

Jill: “It is cold in here” – John closes the window (2)

John: “I am sorry, I forgot our anniversary” – Jill is silent (3)

The first example is straight-forward; a demand is articulated. It is relatively easy to grasp what is intended, and a particular act, passing the butter, would appear to be the appropriate response. This is relatively easy to model and artificial service systems would not have difficulties in dealing with requests like this. The second example is somewhat more complicated. Jill simply utters a statement. However, based on speech act theory (e.g., Searle, 1976), we can assume that any statement can imply a variety of things – for example, in a particular situational context it might become clear that Jill is actually uncomfortable because of the low temperature – even if she did not state this explicitly. There is, depending on the relationship between Jill and John, the implicit message that Jill is not well, but that John has the power to change this state via closing the window. By not stating the request explicitly there is much information conveyed regarding the relationship of the two. This scenario is more complicated for an artificial system to deal with. However, what if there were signs that Jill was indeed not well? She might shiver. If an artificial system would have access to the shivering then a) it could already react before something was said, or b) the sentence could be interpreted in the context of a physiological/behavioral piece of information. The

Küster D. and Kappas A..
What Could a Body Tell a Social Robot that It Does Not Know?.
DOI: 10.5220/0004892503580367
In *Proceedings of the International Conference on Physiological Computing Systems* (OASIS-2014), pages 358-367
ISBN: 978-989-758-006-2

third example is the most challenging. The transcript does not allow much inference. Jill does not say anything. This could have many reasons. If we had access to her expressive behavior, perhaps we could sense whether Jill is very upset. We assume that she is upset – even if this was nowhere indicated. Why? Because as we interact with others, or observe the interaction between others we engage in something akin to mind reading (at least this is how some researchers, e.g., Baron-Cohen et al., 2000 refer to it). There are multiple ways of drawing inferences regarding the emotions, beliefs, and intents of others. In some cases, we use observable behavior (this would fit with the second example). In other situations, we simulate in our head how we would feel if we were in this situation and extrapolate how someone else might feel, particularly, if there is a lack of (supposedly) reliable nonverbal signs. There are many different concepts how we should or could understand these empathic processes (Batson, 2009), and it is fair to say that some of these scenarios will remain a challenge for artificial systems for a long time to come. It is important to note that many humans have difficulties with these situations as well. Children must learn these skills and some adults have difficulties because they are not good in any “theory of mind” tasks throughout their life.

It has often been suggested (e.g., Picard, 1999) to augment human-computer interaction with analyses of nonverbal behavior, as humans also require this information in many situations. However, theoretically it is possible that an artificial system could overcome some of the challenges in communication by including information that would not be available to the human interactant. Imagine in the third scenario, that an artificial system had access to changes in Jill’s cardiovascular system. Perhaps Jill does not say anything, maybe she does not show anything on her face, but perhaps she is, metaphorically, boiling inside. An artificial system might be in an even better position than a human to conclude that Jill is very much upset – possibly offended by the statement of John. This is the reason why the idea of augmenting human computer interaction, for example in the context of social robotics, with an analysis of nonverbal behavior and physiological responses is so intuitively seductive. Thus, in the last few years, several attempts have been made to incorporate such information.

It is the goal of this presentation to describe some of the challenges that an analysis of expression and psychophysiology entail. Initially, we will discuss some conceptual challenges, based on the current state-of-the-art in psychology. This relates particularly to the question what nonverbal behavior and changes in psychophysiological activation mean. Then, we will discuss some technical challenges, which include issues such as sensor placements and artifacts with illustrations from our own laboratory.

2 CONCEPTUAL CHALLENGES

Ideally, psychophysiological and expressive data would reliably yield unambiguous information about the emotional state of a subject across a large range of different situations. However, this is generally not the case. Even our best measures have been shown to correlate only moderately with any other indicator of emotional states (Mauss and Robinson, 2009). Why? The answer is not confined simply to technical aspects of our measurement instruments, but in part relates to more fundamental conceptual issues. To understand these conceptual challenges, we should first consider the *specificity* and *generality* of the relationships between any hypothetical set of measures (Cacioppo et al., 2000). For example, a blood glucose test at a medical examination (Cacioppo et al., 2000) will only be valid as long as certain assumptions about the context are met. Specifically, the measure of blood glucose will not be very informative about a medical condition like diabetes if the patient decided to have a quick snack just before going to the doctor. At a more abstract level, the need for constraints to be met relates to the degree of generality at which a given measure can be expected to faithfully reflect the construct that is to be measured. Within psychophysiology, this has been defined as the level of generality of psychophysiological relationships (Cacioppo and Tassinary, 1990; Cacioppo et al., 2000). For the measurement of emotions in HRI, the implication is that different individual indicators will vary in their validity across experimental contexts.

In addition to their degree of generality, or context-dependency, psychophysiological measures of emotion can vary in how specifically they are tied to emotional states, i.e., in how close they come to having a one-to-one relationship with emotions. This dimension is important because emotions are typically not the only drivers of physiological or expressive behaviors. In other words, there are typically many reasons why a physiological parameter might change at any given moment; i.e., these are instances of many-to-one relationships (Cacioppo et al., 2000). Important examples for this kind of situation are measures of electrodermal activity (EDA) and facial electromyography (EMG), both of which are frequently used indicators of emotional states. While an emotionally arousing stimulus is likely to trigger an electrodermal response, and a pleasant experience will often be accompanied by a response of facial muscles, other factors have an impact on either one of these measures as well. For example, we may smile out of politeness, or we may show an EDA response due to an unexpected noise that has nothing to do with our interaction partner at the time.

Generality of valid measurement contexts and specificity of the relationships of emotion measures can be considered jointly in terms of a 2 by 2 taxonomy. Figure 1 shows the categories of relationships that can be derived from the taxonomy elaborated by Cacioppo and colleagues (e.g., Cacioppo and Tassinary, 1990; Cacioppo et al., 2000).

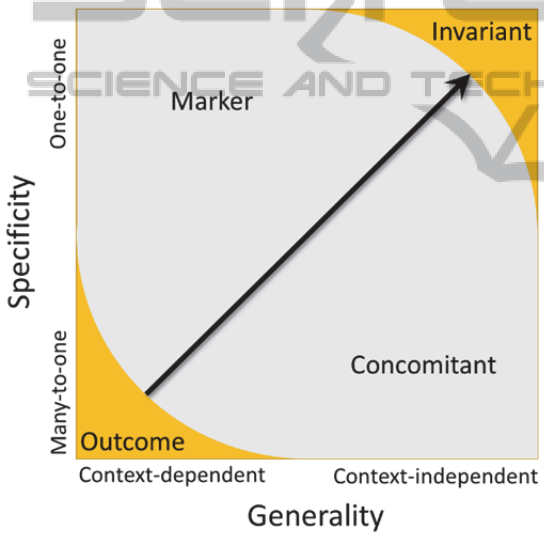

Figure 1: Taxonomy of psychophysiological relationships (adapted from Cacioppo et al., 2000).

In most cases, our measures of emotion in HRI should likely be interpreted cautiously as *outcomes* that are based on many-to one relationships in a specific type of context. This does not mean that they would not allow systematic inferences about emotions that can be useful for live HRI – but it important to understand that we are generally not dealing with specific *markers* or even context-independent, *invariant*, indicators of emotional states that could provide a readout of the true emotional state of a human as it evolves in interaction with a robot. Rather, we need to consider which measure, or set of measures, will be most appropriate for a given measurement situation, and understand which other factors might bias our indicators in this situation. As we will discuss, this also implies a need for further testing and validation of new measures that claim to measure the same psychological processes as conventional laboratory measures – but which aim to do so in a different context.

Given these challenges, what type of inferences can still be drawn from physiology and expression? The recording of facial activity may be a good example here because recent advances in technology have been paving the way for fully automated face-based affect detection (Calvo and D'Mello, 2010). However, as we have argued previously (Kappas, 2010), an overestimation of cohesion between certain facial actions and emotional states can lead to wrong conclusions about user states or action tendencies in real-world applications. In certain cases, a user might smile because she is happy. At other times, she might smile to encourage a robotic system to continue – and at yet other times, she might smile to cope with an otherwise almost painful social situation.

The relationship between certain individual measures of emotional states, e.g., a smile, and fundamental dimensions of emotions such as hedonic valence is not necessarily linear (e.g., Bradley et al., 2001; Lang et al., 1993; Larsen, Norris, and Cacioppo, 2003). However, the reasons why we need to be so cautious about interpreting individual measures of expression in HRI are primarily concerned with the social functions of emotional expressions. Many functions of emotions are social (Kappas, 2010), and the social audience at which participants may direct facial expressions need not even be physically present in an experimental context (Fridlund, 1991; Hess, Banse, and Kappas, 1995).

There are of course cases where automatically recorded data from other facial actions, such as movements associated with frowning or with disgust expressions may be able to help disambiguate the contextualized meaning of a smile – yet even this is not always sufficient. In the latter case, the context-specific implications of a particular experimental situation might be studied in advance with the aid of human judges. For example, when the intended application aims at an interaction between students of a certain age with a particular type of robot in a teaching context that focuses on a specific body of content, a common practice is to train the system with likely types of responses that may occur at certain critical moments during this interaction. While such a strategy obviously does not "solve" the underlying conceptual challenges relating to the context-sensitivity of emotional expressions, it helps the system to learn how to respond more naturally in a certain set of specific situations. In consequence, behavioral rules can be formulated that make use of physiological and expressive indicators without overestimating their contextual generality.

3 PRACTICAL CHALLENGES

Apart from some of the conceptual challenges, such as the temptation to overestimate cohesion between emotional states and their physiological and expressive indicators across different types of social contexts, social robotics has to face a number of more practical challenges. Some of these might be overcome by further technical development, and others can be addressed at least partially by strict adherence to state-of-the-art research standards and additional basic research and validation studies.

Among the technical challenges that, perhaps surprisingly, to date still have not been fully overcome, is the development of a truly comprehensive and reliable *automated system* that can distinguish as many facial Action Units (AUs, see Ekman and Friesen, 1978) as trained human coders can (Calvo and D'Mello, 2010; Valstar et al., 2011). Further, for purposes of improving affect sensing in social robots, a *live integration* of coded AUs into the researchers' own software architecture is required, whereas available commercial systems sometimes only offer this data at a pre-interpreted aggregate level of basic emotions such as "happiness" or "sadness". In other cases, basic licensing issues deny direct access to the online AU-prediction generated by commercial affect detection modules. Either limitation, while seemingly trivial, presents some very practical challenges to effectively incorporating facial affect sensing into interdisciplinary research on emotionally intelligent HRI. Ideally, all action units would be directly available to the artificial intelligence controlling the robot as well as the software architecture that records all responses made over the course of an interaction.

We already know from the psychological literature on how humans perceive emotions that even subtle differences such as the type, timing, and onset of a smile can have a significant impact on how it is perceived (Johnston et al., 2010; Krumhuber and Kappas, 2005; Krumhuber et al., 2007; Schmidt et al., 2006). In consequence, even perfectly accurate information about the presence vs. absence of an action unit as such may not be sufficient to eventually approximate a humanlike level of facial perception capabilities. Clearly, dynamics and intensities matter. While it remains an empirical question to what extent more comprehensive affect recognition systems will be able to improve the socio-emotional capabilities of a robot, work on the technological challenge to collect and process this data has to be accompanied by further empirical research on facial dynamics on the level of action units.

The need for further research on facial dynamics on the level of Action Units relates to a more general set of challenges that have to do with the transfer of extant laboratory research to the context of more applied environments. In the case of facial dynamics in HRI, the additional issue arises that the system needs to interpret the evolving context of an ongoing interaction in *real time*, and this context will typically be based on a substantial number of different sources of information about changing emotional states. In other words, multiple levels of information need to be analysed in real time – the very task that humans appear to perform so effortlessly in daily life!

In the psychological laboratory, basic research usually focuses on a small number of factors that are controlled as strictly as possible to allow inferences about their relative contributions. We have argued previously that this type of fine-grained perspective is crucial for understanding emotional interactions, for example in computer-mediated communication (Theunis et al., 2012). However, it is also clear that social robotics has to find practicable means to collect, filter, and use whatever emotion-related information is available and relevant in the applied context at hand. In HRI, the robot or artificial system has to be able to act, and interact, immediately on the basis of the available input. This changes the focus of important paradigms of laboratory research on emotions, such as the study of individual modalities (e.g., Scherer, 2003), or interindividual differences (e.g., Prkachine et al., 2009), toward a focus on parameters that may be able to help the system to make more sensible decisions about how to respond at different moments of the interaction. In many ways, this challenge to find parameters that are most useful in a number of applied situations is potentially a very fruitful approach, also for the psychological study of emotions. At the same time, however, we have to be aware that factors other than those related to affect-detection per se may turn out to have an even greater impact on the success of a social robot in an interaction. Here, practical considerations are often closely linked to conceptual issues, such as the importance of helping the robot understand the context of emotional expressions rather than affect detection per se (see Kappas, 2010).

Apart from the concrete example of automatic affect detection from facial actions, the use of physiological measures of any kind faces a number of rather basic challenges related to the physical environment of the recording situation. Certain measures, e.g., electrodermal activity (EDA), are known to be influenced by *environmental factors* such as ambient temperature or noise (Boucsein, 2012; Dawson et al., 2000). This means that sudden noises generated by an experimental task, ringing cell phones, other people in the room, or even loud movements of a social robot's motors above a certain threshold, could elicit electrodermal responses that have nothing to do with the intended behaviour of the robot. Further relevant factors include *speech*, *irregular breathing*, and gross body *movements* (Boucsein et al., 2012).

While most of the typical environmental confounding factors can generally be well controlled across experimental conditions in laboratory research, *ambulatory recording* is challenged by significantly more uncertainty regarding the source of variations in EDA (Boucsein et al., 2012). As Boucsein and colleagues (Boucsein et al., 2012) further point out, a socially engaging situation, or even a novel environment may cause a similar magnitude fluctuations in electrodermal activity as stress or fear. For example, the first interaction with a new type of robot will most probably qualify as a new and socially engaging situation for naïve participants. Therefore, ambulatory experiments involving measurement of electrodermal activity in social robotics should include additional time for familiarization of subjects with the experimental environment, the recording procedures, and the robot itself. This is particularly the case for ambulatory recording devices like Affectiva's Q-Sensor (http://www.affectiva.com/q-sensor/) that require a "warmup" period for optimal recording. Due to practical considerations, researchers are often understandably hesitant to devote several minutes of valuable experiment time to seemingly unnecessary familiarization periods and resting baselines. However, in particular when psychophysiological measures are involved, this time of getting to know the experimental context can help eliminate unwanted error variance in how users initially respond to an unknown recording situation. One concrete example where this was successfully applied can be found in a recent study involving the measurement of electrodermal activity of children interacting with an iCAT (Leite et al., 2013). Here, the sensors were attached 15 minutes prior to the actual experiments, and the experimenter guided participants to the location of the interaction with the robot.

In some cases, relevant environmental factors are relatively method specific, and some of these may already be well known to computer scientists. E.g., the quality and type of lighting can have substantial impact on most facial affect recognition systems. Likewise, tracking of more than one human user at a time can pose a considerable technical challenge for the reliable recording of facial action units. For other measures, such as the recording of electrocardiographic (ECG) data, or facial electromyography (EMG), impedances between the skin of the subjects and the recording electrodes can play an important role, and that even in cases where traditional wired sensors are used (cf., Fridlund and Cacioppo, 1986; Cacioppo et al., 2007).

For this reason, the best choice of recording instruments depends not only on the type of research questions asked, but also on the *physical constraints of the recording situation*, as well as ongoing developments for both sensors and software. For example, currently available sensors for the recording of electrodermal activity, like the aforementioned Affectiva Q-Sensor, have been focusing on the advantages of a convenient placement near the wrist of participants. However, this placement may not be an optimal measurement location for the assessment of emotional sweating (van Dooren et al., 2012; see also Payne et al., 2013), nor is it recommended by the current official guidelines because this site may reflect more thermoregulatory rather than emotionally relevant electrodermal phenomena (Boucsein et al., 2012). However, either further empirical research might establish that this recording site can nevertheless generate *enough* emotionally significant data despite being not optimal (Kappas et al., 2013), or additional technical developments might make it possible to perform reliable wireless recordings from a different site.

A final, but important, set of practical challenges relates to the impact of the *psychological* rather than physical recording environment. From a social psychological perspective, the presence vs. absence of a human experimenter in the context of an ongoing social interaction between a human participant and a robotic partner is a potentially very interesting variable. In HRI, there are often practical reasons why researchers may decide to keep an experimenter nearby, e.g., in studies with younger children involving potentially very expensive or technically challenging systems. However, the physical presence or absence of an experimenter may be sufficient to fundamentally change the social context of an experiment, in particular in those cases where having an experimenter present might appear to be a practical requirement. In fact, social psychological research of the last decades has repeatedly demonstrated that even a merely implicit presence of other people can have on key emotional behaviors such as smiling (Fridlund, 1991; Hess et al., 1995; Manstead et al., 1999; Küster, 2008). For example, Fridlund (1991) measured facial responses to funny videos and found that even to believe that a friend was watching the same videos elsewhere was already sufficient to increase smiling in comparison to a truly solitary viewing condition. In comparison to such subtle effects, the actual physical presence of another person can hardly be expected to fail to have a significant effect on the psychological recording situation, no matter how justifiable the physical presence of a "silent experimenter" may appear to be.

A particularly relevant and widely used experimental technique in the study of HRI is based upon the use of a Wizard of Oz (WoZ; e.g., Dautenhahn, 2007; Riek, 2012) paradigm, i.e., a human puppeteer who controls some or all of the behavioural responses of the robot. This is a socially complex situation because the participant is interacting with a "robot" that is at least partially controlled by another human who is usually seated outside of direct view, or in a separate room. The puppeteer remains anonymous and as such invisible to the human subject, and this presents a certain safeguard against sociality effects that are tied to the immediate physical presence of an experimenter. However, the WoZ paradigm can nevertheless vary in the level at which the wizard is implicitly present in the situation, and there are potentially numerous ways in which a confederate, i.e., the wizard, may influence or "prime" responses of the participant in rather automatic ways with little or no conscious awareness (see Bargh et al., 1996; Kuhlen and Brennan, 2013). For this reason, it is important to control the psychological recording environment as well as possible across all participants taking part in an experiment. In particular, the wizard(s) should receive systematic training to standardize responses as well as minimize learning effects and fatigue which may otherwise create undesirable systematic variance in how the social context of the experiments is perceived by the very first vs. later participants. This may of course be less of an issue in WoZ studies that focus on rapid prototyping (see, e.g., Dautenhahn, 2007) yet it becomes more critical as soon as differences in user evaluations are to be tested systematically, or when physiological or expressive measures are to be tested and trained to be used as parameters. As Fridlund (1991, 1994) and others have shown, socially relevant expressions appear to be surprisingly vulnerable to even very subtle variations of the social environment. However, in a recent review on Wizard of Oz studies in HRI (Riek, 2012), only a small minority of 5.4% of studies reported any pre-experimental training of wizards, and only 24.1% reported an iterative use of WoZ. This suggests that in particular the control of seemingly minor social factors may require more systematic attention.

Further challenges related to the psychological context of an experiment with physiological measures relate to the subject's *awareness* of being measured, and the impact of preceding tasks. First, it is of course not surprising that a feeling of being observed is likely to bias results, for example when effects of social desirability are considered (Paulhus, 1991; 2002). However, this is also an example of the *context-dependency* of psychophysiological relationships that we discussed in the section on conceptual challenges above. Thus, children, for example, can be expected to respond differently to observation than adults, and adult students will likely respond differently from other specific groups such as teachers, or elderly people. Importantly, from the perspective of physiological measures, we can further not assume that, e.g., differences observed between age groups on the level of self-reported emotions will translate one-to-one into the same type of differences in the physiological domain. For example, if we were to ask a few children and a few adult students about their emotional experience in a pre-test, we might find that the children perceive certain aspects of the robot's emotional capabilities more positively than our adult sample. However, it might be that the psychological context of the experiment as such, rather than the robot with its limited response repertoire, would have been much more exciting for the children than for the adults. This generally elevated level of excitement, or arousal, may show up in physiological or expressive measures – but it can depend on the specific measure in question to what extent this is the case. However, while simplified pretests of experiments can be very useful, e.g., for the training of a confederate or WoZ, we have to be cautious about predictions derived on the basis of a different type of sample than the one that is finally used. Thus, not only may people of different age groups respond differently to the social environment of an experiment – but they might also express themselves in different ways. Physiological and expressive measures may be of particular use in explaining some of these potential differences. However, they themselves require careful testing in an experimental environment that should, physically as well as psychologically, be matched as closely as possible to the final experimental design.

4 SOLUTIONS AND NEW CHALLENGES

Until recently, our discussion of practical challenges for the use of psychophysiological measures in HRI research would have had to begin with a discussion of the very basic problems associated with moving amplifiers and cables out of the laboratory into locations that allow a certain degree of freedom of movement to participants without loss of signal quality. With the advent of a number of wireless lightweight recording systems for both expressive as well as bodily signals, many of these issues appear to have been reduced or eliminated. Or have they?

As might be expected, the answer depends on the specific measure and measurement context in question. If, for example, the participant can be expected to remain seated, and only smiling activity needs to be recorded for the purposes of an experiment, a number of inexpensive or freely available facial affect detection systems can be expected to provide this data reliably. As discussed above, the detection of a larger range of AUs, however, is still far from solved (Valstar et al., 2012; Chu, De la Torre, and Cohn, 2013). Further, automated facial affect detection is still challenged by individual differences in facial morphology that can dramatically influence the performance of classifiers for previously unseen individuals (Chu et al., 2013). Such differences include, for example the shape and type of eyebrows and deep wrinkles (Chu et al., 2013).

Other techniques, such as facial electromyography (EMG) use electrodes attached to the face, and can be more robust in this respect because trained human experimenters affix the electrodes at the precise recording site appropriate for an individual subject. Further strengths of Facial EMG are a high temporal resolution and sensitivity to even very subtle intensity changes of activation (van Boxtel, 2010). Due to the technical and practical limitations involved, however, facial EMG has been a strictly laboratory based measure until just a few years ago. Yet facial EMG is an example where ongoing technical developments are beginning to look rather promising. Thus, meanwhile, wireless off-the-shelf solutions have been produced for the recording of facial EMG (e.g., BioNomadix, www.biopac.com). Further, at a prototype level, head-mounted measurement devices no longer require a physical attachment of electrodes to the face. In an initial validation study, strong correlations were observed between such a device and traditional measurement at typical recording sites (Rantanen et al., 2013). This might address a number of disadvantages of the use of facial EMG, such as its relative intrusiveness compared to a video recording, including the pre-treatment of the skin before electrodes can be attached. However, further empirical validation of contact-free EMG recording is still required, as well as a reduction in weight and general usability before such devices may be ready for a larger-scale use in applied contexts.

For other physiological signals, such as EDA, particularly lightweight and convenient portable sensors have already been developed. As for facial EMG, such sensors may help to overcome some of the typical practical challenges associated with using physiological recording devices "in the wild". Sensors that can be worn just as easily as a wrist-watch, for example, are likely to cause much less interference with an ongoing study. In consequence, it can be hypothesized that they will have a substantially smaller impact on levels of self-awareness of participants, and the general extent to which subjects feel observed. Likewise, as physiological recording systems are becoming more useable and less obtrusive, the range of possible applications broadens, and this may allow entirely new avenues for research. However, new sensors in this domain often still lack systematic empirical validation studies. This concerns not only the reliability of the measurements taken by these devices but also the validity of the psychological constructs being measured. In particular where new and innovative recording sites are used, the empirical question arises if the convenient new measurement location still reflects the same kind of psychological mechanisms. If it does not, then the inclusion of such data risks contributing little to the effective affect-sensing capabilities of a robot – and, in the worst case, it might even be counter-productive.

5 OUTLOOK

We have described some of the most important conceptual and practical challenges that have to be overcome to incorporate expressive and psychophysiological data in field research on HRI. On the conceptual level, we have emphasized some of the fundamental limits of generality and specificity in the relationships between bodily measures of emotion and the underlying psychological constructs. On a more practical level, we have discussed many of the most commmen problems faced by researchers who want to employ psychophysiolgical measures outside the laboratory. We have emphasized how both the physical and the psychological environment of an experiment need to be carefully considered when designing experiments such as WoZ studies. This is particularly the case when meaningful effects are to be compared between experimental conditions, and when inferences are to be drawn about the associated psychological processes involved. We now conclude our contribution by the attempt of a brief look into the future.

While the significant challenges to the use of bodily signals highlighted in this paper should not be underestimated, there are likely a substantial number of situations where a robot, for all practical purposes, need not be perfect in order to be perceived as attentive, empathic, or emotional. Thus, to improve perceived realism of robotic behaviour, it may not always be necessary to understand precisely the emotional state of participants throughout the entirety of the experiment. While this would, of course, be advantageous, even humans are not necessarily the gold standard of affect detection that we might think intuitively take them for (Kappas, 2010). Rather, humans have been shown to be heavily influenced by contextual factors (Russell et al., 2003), to perform surprisingly poorly at emotional lie-detection tasks (Ekman and O'Sullivan, 1991), and to tend to fail at tasks involving interoception, such as tracking one's own heart-beats (Katkin et al., 1982).

What are the implications if social robots after all do not have to be perfect at distinguishing, e.g., polite social smiles from genuine smiles? It is possible that robots may not even have to excel at general inferences about ongoing changes in action tendencies (see Frijda et al., 1989) or predict perfectly the likely future actions of a human partner from physiology alone across a broad range of contexts. While improved measurement devices are undoubtedly an important piece of the puzzle, we argue that an actual understanding of the situation may turn out to be equally important. At present, more work appears to be needed on the design of critical experimental situations where the pattern of all available information allows clear predictions on the appropriateness of a set of different behavioural response options for the robot. For example, an increase in physiological arousal coupled with the participant's eye-gaze and a smile directed at the robot could be a fairly clear indicator that the interaction is going well, and that the robot might continue further along the current path. At other times, physiological data, including information about head orientation or gaze synchronicity, might be used successfully to adjust the precise timing of certain pre-arranged sets of statements. Finally, yet other data might be used in concert with physiological and expressive data, e.g., response latencies or button presses recorded from an ongoing task (e.g., Leite et al., 2013). If we can use physiology to improve social robots at certain key moments of an interaction, we may already be on a good way to improve our understanding of context-sensitive emotional responding in HRI at a more general level.

Through careful experimental design, the context-dependency of emotions in HRI may, at least in part, be transformed from a challenge into a characteristic that can be systematically employed to improve realism and fluency of social robotics. However, for this to occur, substantial additional basic research is still needed concerning the role of social context in physiological and expressive measures of emotion in HRI.

REFERENCES

Baron-Cohen, S., Tager-Flusberg, H., Cohen, D. J., 2000. *Understanding other minds: Perspectives from developmental cognitive neuroscience*, Oxford University Press. Oxford, UK.

Bargh, J. A., Chen, M., Burrows, L., 1996. Automaticity and social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*, 230-244.

Batson, C. D., 2009. These things called empathy: Eight related but distinct phenomena. In: J. Decety & W. Ickes (Eds.), The Social Neuroscience of Empathy (pp. 3–15), MIT Press. Cambridge, UK.

Boucsein, W., 2012. *Electrodermal activity*, Springer. New York.

Boucsein, W., Fowles, D. C., Grimnes, S., Ben-Shakhar, G., Roth, W. T., Dawson, M. E., Filion, D. L., 2012. Publication recommendations for electrodermal measurements. *Psychophysiology*, *49*, 1017-1034.

Bradley, M. M., Codispoti, M., Cuthbert, B. N., Lang, P. J., 2001. Emotion and motivation I: Defensive and appetitive reactions in picture processing. *Emotion*, *1*, 276-298.

Calvo, R. A., D'Mello, S., 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, *1*, 18-37.

Cacioppo, J. T., Tassinary, L. G., 1990. Principles of psychophysiology: Physical, social, and inferential elements. New York: Cambridge University Press.

Cacioppo, J. T., Tassinary, L. G., Berntson, G. G., 2000. Psychophysiological science. In: J.T. Cacioppo, L.G. Tassinary, G.G. Berntson (Eds.), Handbook of Psychophysiology (3-23), Cambridge University Press. Cambridge, UK, 2nd edition.

Cacioppo, J. T., Tassinary, L. G., Berntson, G. G., 2007. *Handbook of Psychophysiology*, Cambridge University Press. UK, 3rd edition.

Chu, W.-S., de la Torre, F., Cohn, J. F., 2013. Selective transfer machine for facial action unit detection. *IEEE International conference on Computer Vision and Pattern recognition*. Portland, OR.

Dautenhahn, K., 2007. Methodology & themes of human-robot interaction: A growing research field. *International Journal of Advanced Robotic Systems*, *4*, 103-108.

Dawson, M. E., Schell, A. M., Filion, D. L., 2000. The electrodermal system. In: J. T. Cacioppo, L. G. Tassinary, G. G. Berntson (Eds.), *Handbook of Psychophysiology* (200-223), Cambridge University Press. Cambridge, UK.

Ekman, P., Friesen, W. V., 1978. *The Facial Action Coding System*, Consulting Psychologists Press. Paolo Alto, CA., 2nd edition.

Ekman, P., O'Sullivan, 1991. "Who can catch a liar?" *American Psychologist*, *46*, 913-920.

Fridlund, A. J., 1991. Sociality of solitary smiling: Potentiation by an implicit audience. *Journal of Personality and Social Psychology*, *60*, 229-240.

Fridlund, A. J., 1994. *Human facial expression: An evolutionary view*. Academic Press. San Diego, CA.

Fridlund, A. J., Cacioppo, J. T. 1986. Guidelines for human electromyographic research. *Psychophysiology*, *23*, 567-588.

Frijda, N. H., Kuipers, P., ter Schure, E., 1989. Relations among emotion, appraisal, and emotional action readiness. *Journal of Personality and Social Psychology*, *57*, 212-228.

Hess, U., Banse, R., Kappas, A., 1995. The intensity of facial expression is determined by underlying affective state and social situation. *Journal of Personality and Social Psychology*, *69*, 280-288.

Jonhston, L., Miles, L., Macrae, C. N., 2010. Why are you smiling at me? Social functions of enjoyment of non-enjoyment smiles. *British Journal of Social Psychology*, *49*, 107-127.

Kappas, A., 2010. Smile when you read this, whether you like it or not: Conceptual challenges to affect detection. *IEEE Transactions on Affective Computing*, *1*, 38-41.

Kappas, A., Küster, B., Basedow, C., Dente, P., 2013. A validation study of the Affectiva Q-Sensor in different social laboratory situations. *Poster presented at the 53rd Annual Meeting of the Society for Psychophysiological Research*. Florence, Italy.

Katkin, E. S., Morell, M. A., Goldband, S., Bernstein, G. L., Wise, J. A., 1982. Individual differences in heartbeat discrimination. *Psychophysiology*, *19*, 160-166.

Kuhlen, A. K., Brennan, S. E., 2013. Language in dialogue: When confederates might be hazardous to your data. *Psychonomic Bulletin & Review*, *20*, 54-72.

Küster, D., 2008. *The relationship between emotional experience, social context, and the face: An investigation of processes underlying facial activity* (Doctoral dissertation). Retrieved from http://www.jacobs-university.de/phd/files/1199465732.pdf.

Leite, I., Henriques, R., Martinho, C., Paiva, A., 2013. Sensors in the wild: Exploring electrodermal activity in child-robot interaction. *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*. ACM/IEEE Press.

Krumhuber, E., Kappas, A., 2005. Moving smiles: The role of dynamic components for the perception of the genuineness of smiles. *Journal of Nonverbal Behavior*, *29*, 3-24.

Krumhuber, E., Manstead, A.S.R., Kappas, A., 2007. Temporal aspects of facial displays in person and expression perception: The effects of smile dynamics, head-tilt, and gender. *Journal of Nonverbal Behavior*, *31*, 39-56.

Lang, P. J., Greenwald, M. K., Bradley, M. M., Hamm, A. O., 1993. Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, *30*, 261-273.

Larsen, J. T., Norris, C. J., Cacioppo, J. T., 2003. Effects of positive and negative affect on electromyographic activity over zygomaticus major and corrugator supercilii. *Psychophysiology*, *40*, 776-785.

Manstead, A. S. R., Fischer, A. H., Jakobs, E., 1999. The social and emotional functions of facial displays. In P. Philippot, R.S. Feldman, E.J. Coats (Eds.), *The social context of nonverbal behavior* (pp. 287-313), Cambridge University Press. Cambridge, UK.

Mauss, I. B., Robinson, M. D., 2009. Measures of emotion: A review. *Cognition & Emotion*, *23*, 209-237.

Paulhus, D. L., 1991. Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & Wrightsman Eds., *Measures of personality and social psychological attitudes* (pp. 17-59), Academic Press, Inc. San Diego, CA.

Paulhus, D. L., 2002. Socially desirable responding: The evolution of a construct. In H.I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49-46), Erlbaum. Mahwah, N.J.

Payne, F. H., Dawson, M. E., Schell, A. M., Kulwinder, S., Courtney, C. G., 2013. Con you give me a hand? A comparison of hands and feet as optimal anatomical sites for skin conductance recording. *Psychophysiology*, *50*, 1065-1069.

Picard, R. W., 1999, Affective Computing for HCI. *Proceedings HCI, 1999*. Munich, Germany.

Prkachin, G. C., Casey, C., Prkachin, K. M., 2009. Alexithymia and perception of facial expressions of emotion. *Personality and Individual Differences*, *46*, 412-417.

Rantanen, V., Venesvirta, H., Spakov, O., Verho, J., Vetek, A., Surakka, V., Lekkala, J., 2013. Capacitive measurement of facial activity intensity. *IEEE Sensors Journal*, *13*, 4329-4338.

Riek, L. D., 2012. Wizard of Oz studies in HRI: A systematic review and new reporting guidelines. *Journal of Human-Robot Interaction*, *1*, 119-136.

Rosengren, K. E., 2000. *Communication: An introduction*, Sage. London.

Russell, J. A., Bachorowski, J.-A., Fernández-Dols, J.-M., 2003. Facial and vocal expressions of emotion. *Annual Review of Psychology*, *54*, 329-349.

Scherer, K. R., 2003. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*, 227-256.

Schmidt, K. L., Ambadar, Z., Cohn, J. F., Reed, L. I., 2006. Movement differences between deliberate and spontaneous facial expressions: Zygomaticus major action in smiling. *Journal of Nonverbal Behavior*, *30*, 37-52.

Searle, J. T., 1976. A classification of illocutionary acts. *Language in Society*, 5, 1-23.

Theunis, M., Tsankova, E., Küster, D., Kappas, A., 2012. Emotion in computer-mediated-communication: A fine-grained perspective. *Poster presented at the 7th Emotion Pre-Conference to the Annual Meeting of the Society for Personality and Social Psychology*. San Diego, California.

Valstar, M. F., Mehu, M., Jiang, B., Pantic, M., Scherer, K., 2012. Meta-analysis of the first facial expression recognition challenge. *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics*, *42*, 966-979.

va Boxtel, A., 2010. Facial EMG as a tool for inferring affective states. In A. J. Spink, F. Grieco, O. E. Krips, L. W. W. Loijens, L. P. J. J. Noldus, & P. H. Zimmerman (Eds.), *Proceedings of Measuring Behavior 2010* (pp. 104-108). Eindhoven, The Netherlands.

van Dooren, M., de Vries, J. J. G., Janssen, J. H., 2012. Emotional sweating across the body: Comparing 16 different skin conductance measurement locations. *Physiology & Behavior*, *106*, 298-304.