Computational Models of Consciousness-Emotion Interactions in Social Robotics: Conceptual Framework

Remigiusz Szczepanowski, Małgorzata Gakis,
Krzysztof Arent and Janusz Sobecki

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.72369

Abstract

There is a little information on how to design a social robot that effectively executes consciousness-emotion (C-E) interaction in a socially acceptable manner. In fact, development of such socially sophisticated interactions depends on models of human high-level cognition implemented in the robot's design. Therefore, a fundamental research problem of social robotics in terms of effective C-E interaction processing is to define a computational architecture of the robotic system in which the cognitive-emotional integration occurs and determine cognitive mechanisms underlying consciousness along with its subjective aspect in detecting emotions. Our conceptual framework rests upon assumptions of a computational approach to consciousness, which points out that consciousness and its subjective aspect are specific functions of the human brain that can be implemented into an artificial social robot's construction. Such research framework of developing C-E addresses a field of machine consciousness that indicates important computational correlates of consciousness in such an artificial system and the possibility to objectively describe such mechanisms with quantitative parameters based on signal-detection and threshold theories.

Keywords: social robot, consciousness-emotion interaction, machine consciousness, signal-detection theory

1. Introduction

It is widely acknowledged that a social robot should be an embodied agent which can communicate with people easily, using both verbal and nonverbal signals [1]. Such a robot needs to have a wide range of social and cognitive skills [2, 3] to understand human behavior and to be intuitively understood by people. However, it should be noted that nowadays there is a gap between the requirements concerning a social robot and their implementations. It is due to imperfection in technology and deficiencies in theory in various areas ranging from psychology through computer science up to classical robotics. Despite of intense technological efforts over the last two decades in terms of developing high-level cognition models for human-robot interaction (HRI), so far, robot's constructions have been hardly equipped with such competency. Here, we focus on issues concerning development on processing the interaction between consciousness and emotions in a social robot.

2. Designing human-robot interaction

Designing efficient HRI is a basic research problem of modern social robotics [1, 4]. This is mainly due to a technological struggle to make a construction of robots that is intended to share space with humans and support them in daily life in a socially acceptable manner. The joint efforts of modern research including cognitive psychology, developmental psychology, philosophy of mind, and modern technology such as artificial intelligence and machine learning show that creating effective HRI depends on the implementation of human high-level cognition into a robot's system. For example, emotions in the context of social robots have attracted a considerable attention for the last two decades [5]. It is expected that artificial emotions increase plausibility of interactions including predictability of a robot behavior. The well-known idea of a "theory of mind" describing our ability to mentalize others' internal states was captured by theoretical accounts by Baron-Cohen [6] and Leslie [7] and finally was used to construct a Cog humanoid robot with the usage of current technology [3]. In addition, endowing of a robot in a theory of mind [3] could allow the robot to detect, recognize, interpret, and react to a human behavior and hence to make interaction more human-like. There are a lot of works concerning emotions, computational models of emotions in psychology, and computer science, but there is no result to date that would considerably improve a social robot behavior. Attempts to implement and verify a computational model of emotions in a control system of a real robot have been undertaken systematically for many years. For example, emotional system of Kismet designed from scratch is strongly inspired by various theories of emotions in humans [2]. An affective, computational model of mind fearnot affective mind architecture (FAtiMA) [8] was implemented in the robot FLASH [9]. The works in [10–12] are examples of systems that were verified using agent-based modeling software and possess a potential in the context of implementation in robots. The experience gained to date points to three areas of challenges.

Nowadays, sensory systems of robots are insufficient to detect social events, such as human emotions, intentions, attention points, etc. Clear and natural expressing of emotions and other internal states by a robot requires advanced and expensive mechatronic solutions. Computational model of consciousness and emotions can be interpreted as compound components of a higher level part of the robot control architecture. Therefore, implementation of such models requires them to be formally complete and adequate that is not guaranteed by the current psychological research.

3. Consciousness-emotion interaction as functions of the human brain

Philosophy of designing robotic systems inspired by human high-level cognition, including attentional and perceptual processes, is commonly used and known as a biologically inspired approach (see [11–13]). Some studies have also indicated the possibility of implementing into the social robot a computational architecture, which is inspired by a neurobiological basis of the brain [14]. For instance, there are well-developed robotic control systems of high-level cognition that implement a feature-integration theory of attention [15] or a model of saliency-based attentional search mechanisms [16] that have been intensively verified both behaviorally and computationally [13].

Contemporary brain research suggests that the interaction of cognition and emotion may be crucial for a social robot's design [14, 17]. For instance, Pessoa [14] argues that the fundamental problem is to determine an organization of the robotic system in which cognition and emotion are intertwined in a general information-processing architecture. In general, such information-processing architecture should be viewed as a general theory that describes important components of the system and relations between such components [18]. In this way, the adopted architecture can determine an organization of the cognitive system and general principles of information processing in the robotic system. Therefore, goal-directed or conscious behavior of a social robot in terms of recognizing human affective states will require understanding how complex cognitive and affective processing should be mapped into a robotic information-processing system that performs computational algorithms to integrate C-E interactions effectively as the human brain does it. In fact, brain research indicates that there is no decisive evidence what kind of organization of information processing is ultimate to mediate the C-E interaction effectively. Many neuroscientists (see [19, 20]) indicate that there is a functional division in the brain between low-level processes of emotion regulation (for instance, linked with amygdala activation) and higher order processes that are associated with frontal and parietal cortical activity involved in conscious goal-directed behavior. In addition, according to modern neurobiological accounts (see [21]), the amygdala synchronizes and modulates access to affective stimuli in such a way that their representations are stronger (exert a stronger influence on behavior) than neutral stimuli. Thus, selection of specific architecture can determine how a specialized C-E interaction system should be organized; it should also enable to define specific components of such system that are attributed to specific brain structures as well as describe how computations underlie high-level cognitive processes underlying such interaction. Following this line of reasoning, it is possible that the architecture of the C-E interaction in the social robot may be either structured into blocks (a theoretical system that processes sequentially, in which the knowledge is hierarchical, etc.), or modules (there are independent, autonomous, distributed modules handled by a central processor, e.g. [22]) or represents some kind of non-modular organization in which information processing is inspired by neurocomputations for which simple interactions between processing units are going on [23]. It is therefore important in terms of social robots to set up theoretical criteria to analyze potential architecture of the C-E interaction in the brain regarding structural components and functionality of the social robot's system.

4. Consciousness-emotion interaction and machine consciousness approach: establishing formal assumptions

Besides specific architecture of cognition for social robots, the essential problem of designing effective HRI is to analyze conscious behavior of the robot by considering human conscious knowledge and therefore considering subjectivity experience that accompanies consciousness (phenomenal aspects of consciousness; see [24, 25]). In our opinion, such a research problem should be embedded within the area of machine consciousness that can identify critical computational correlates of consciousness [26] to establish HRI. According to this computational approach, consciousness and its subjective experience can be explained by higher level cognition that is grounded in neurocomputations in the brain [25]. This approach not only allows for development of machine consciousness but also attempts to explain a so-called hard problem of consciousness that is related to inability to objectively measure phenomenal aspects of consciousness (see [27]). In fact, the theories of machine consciousness have been successively implemented in artificial environments (e.g., system CLARION; see [28]); some attempts were made in terms of implementing them into robotic systems [29].

Given such philosophical physicalism [30], we assume that consciousness of the robot can be addressed within an information-processing framework in terms of behavior control, information integration, attention and access to the information, or ways of expressing internal states of the robot. According to this framework, social robots are embodied, socially intelligent agents, operating in the human environment [1, 11, 12]. Our conceptual framework attempts to solve the problem of modeling consciousness-emotion interactions using the machine consciousness approach. Below, we will demonstrate that feasibility of hypothesized computational correlates of consciousness for the C-E interaction in a social robot system is formally allowed within on a signal-detection theory (SDT) [31] and a threshold theory [32, 33].

5. Modeling consciousness-emotion interaction using a combination of signal-detection and threshold approaches

According to Reggia and colleagues [25], the machine consciousness approach indicates that a possible computational correlate of consciousness is representational property defined as a possible way of encoding incoming information in the cognitive system. It is postulated in this account that such representations may be a pattern of neuronal activity that is encoded in the current states of the neuronal network [34]. For example, in a study on visual awareness with backward masking [35], patterns of conscious behavior are described as human ability to detect emotion under a forced-choice condition within a series of signal (e.g., mimic fear expression) and noise trials (e.g., neutral face expression) (see [36]). The assumption that consciousness is the ability to differentiate signal from noise based on choice behavior has enabled researchers to use a signal-detection theory (SDT) to quantify consciousness of emotion with objective sensitivity parameters [37]. It is therefore clear that conscious behavior identified with the SDT parameters can be used as a computational correlate to determine objective representations of the C-E interactions in the social robot's construction.

The computational approach to consciousness [25] also points out that an additional potential computational correlate of consciousness is represented by relational properties between the components of human knowledge. According to Reggia and colleagues [25], assumptions of higher order theory (HOT) of consciousness [36] nicely fit with this computational aspect. In particular, HOT postulates a computational correlate of consciousness which is the relationship between stimulus representation and its corresponding subjective knowledge of being conscious of the first-order representation (metarepresentation) [25]. In fact, modeling studies of consciousness and emotion [33, 37] showed that an adequate computational approach which considers the relation between consciousness and emotion can be described within the SDT framework. Szczepanowski [33] with his original proposal has shown that the SDT computational model may consider the fact that consciousness and emotions interact with one another. In addition, such computational SDT model of consciousness allows for a hierarchy of the information processing associated with conscious detection of emotion, that is, higher order processing requires prior discriminations of emotion at the lower level. This suggests that the relational relationship between the components of knowledge underlying architecture of the C-E interaction could be crucial for a social robot's design.

The machine consciousness framework also indicates that consciousness is characterized by a specific information-processing mode [25, 38]. Some theoretical accounts emphasize effectiveness of such conscious processing, and it has been argued that the information content in conscious state is processed globally [38]. For instance, Dehaene [39] who is an advocate of such line of reasoning has shown that global processing in the brain may be linked with activation of extensive long-distance neuronal connections that link several separate brain areas, including prefrontal areas that are not activated in another processing mode [38]. Indeed, such conscious processing mode may stand for a computational correlate of consciousness that explains the nature of conscious access that involves subject's disposition to action and mobilizes and integrates mental functions that operate independently and differ in terms of tasks under the unconscious condition [38]. In the context of conscious affective processing, it seems likely that activation of the global processing mode may operate on an "all-or-none" or discrete fashion when emotional stimuli enter consciousness [37, 40]. In fact, Szczepanowski [33] based on a Krantz threshold theory [32] demonstrated that preferences for affective representation to access consciousness may be the threshold processing. Thus, preferential conscious processing of emotion in the brain may arise from the fact that activation strength of affective stimuli to enter consciousness is characterized in the discrete manner [33, 37, 40]. This implies that in the case of affective information, the robotic system could be implemented with the global processing mode based on thresholds to be able for handling effective and natural HRI.

Thus, with the abovementioned assumptions, our conceptual framework shows that the computational organization underlying the C-E interaction in the robotic system should correspond to an architecture of affective computing in the brain [14, 41] and should be based on computational correlates of consciousness [25] by including (i) a low-level representation correlate which enables robot's objective conscious perception of emotion, (ii) a metacognitive correlate of robot's subjective knowledge of emotion, and (iii) a conscious processing mode based on global access to the emotion content. Here, we will explain in detail the idea of modeling computational correlates of C-E interactions with mathematical frameworks.

6. Signal-detection theory to encode objective consciousness of emotion in a social robot

The SDT theory assumes that the ability of human subject to perceive a stimulus is described by the probability of deciding whether a signal or noise stimulus was presented in a given trial [31]. The fluctuations of a stimulus presented within series of trials, for example, manipulated with a time exposure, or visibility of the stimulus, are determined by Gaussian probability density functions [31, 33, 42]. Because of presentations of two stimulus types under the forced-choice detection condition, participant within experimental condition produces correct (a hit (*H*) and correct rejection (CR)) and incorrect responses (a false alarm (*FA*) and miss (*M*)). The ability to detect a stimulus is then described by a sensitivity parameter $d'_{\text{Type I}}$, which conceptually corresponds to a difference in mean values from the probability distributions for the signal and noise. In addition to the sensitivity measure, the detection theory also provides a bias measure $c_{\text{Type 1}}$, which determines the participant's tendency to favor either "yes" or "no" responses during the detection process. Based on probability distributions, the receiver operating curve (ROC) is computed whose course determines the participant's ability to detect stimuli. According to the SDT, the task performance above the chance level will indicate conscious perception as measured by a significant nonzero sensitivity index ($d'_{\text{Type I}} > 0$). Similar conclusions are formulated when a size of the area under the ROC curve is above the level of 0.5 which is characterized by the so-called parameter $A'_{\text{Type 1}}$. In fact, according to Lau [42], the SDT sensitivity measure of consciousness in detection tasks is not sufficient and in terms of consciousness, it is important to determine decision criteria for detecting a stimulus based on the c parameter rather than discrimination ability per se. For instance, the SDT interpretation of behavior in blindsight patient with visual cortex damage who deny any visual sensation in the resultant visual field defect but can nonetheless detect the visual emotion stimuli presented in the area [43] would indicate a nonzero value $d'_{\text{Type I}}$ and paradoxically conscious perception. Therefore, in terms of the consciousness measure, establishing and maintaining appropriate decision-making are critical when detecting stimuli, rather than using sensitivity values $d'_{\text{Type I}}$ which rather would refer to the basic effectiveness of the information processing [42].

In terms of machine consciousness, it seems to be clear that the SDT approach by estimating sensitivity of first-order detection of emotion and bias can allow to determine computational correlates of social robot's objective knowledge about human affective states. A hypothetical robotic system (see **Figure 1**) with the functionality of objective consciousness of emotion may be equipped with emotion recognition algorithms that constantly analyze human expressions based on sequences of affective stimuli within time events and will then result in online SDT computations that simulate objective consciousness about recognized human affective state. In such a way, the use of the detection theory will enable to capture one of the key properties of conscious knowledge associated with choice behavior [44] in a possible robotic system.

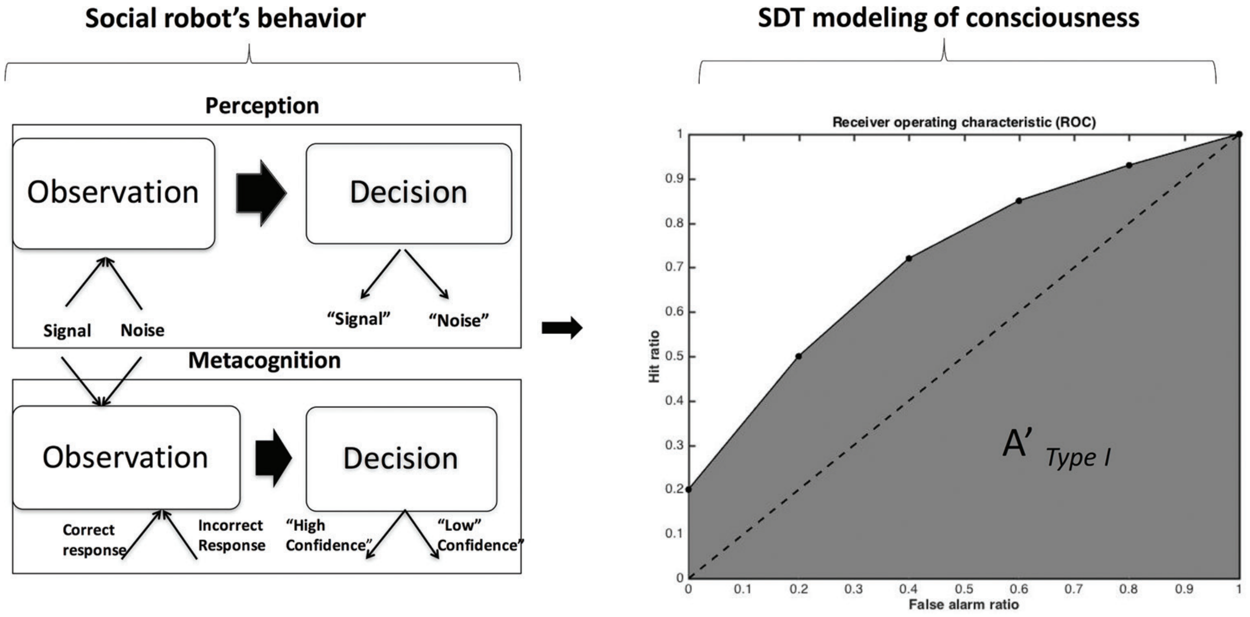

Figure 1. General idea of robotic system with measurement function of objective conscious perception of human emotions (inspired by [35]).

7. Signal-detection approach to encode metacognitive consciousness of emotion in a social robot

Objective theories of consciousness link consciousness to the ability of detecting incoming external stimuli by choice [45]. According to this view, consciousness is described as sensory processing that ignores first-person experience underlying subjective knowledge (metacognition) of its own representation of processing the incoming information. The problem of consciousness and its relations to metacognition has been viewed a central topic in consciousness research [46] which fits well the HOT approach [47]. This theory of consciousness is now considered to be a main framework that explains how people are aware of their conscious states [47]. On the one hand, HOT implies an assumption that consciousness depends on the presence of metacognition [48, 49]; on the other hand, there are opposite claims that metacognition is a prerequisite for the emergence of consciousness [50, 51]. According to this second assumption, consciousness is a first-person metarepresentation which refers to the ability to acquire knowledge about first-order mental states [52]. This second HOT view is well documented by studies on conscious learning with a neuronal network approach in which the brain via learning processes the information about external world and creates its own re-representation on how it is to be in a conscious state of the processed information [53, 54]. In fact, both the HOT theory and connectionist model are consistent with the signal-detection framework (see [53]).

Following the HOT view on consciousness, our conceptual framework assumes that a computational correlate of consciousness is relational property between first-order representation of emotion and metacognition [25]. Adopting such architecture of the machine consciousness indicates that metarepresentation is distinct from first-order representation and may require separate neuronal structures in the brain. In fact, brain research provides empirical evidence for feasibility of such architecture of neurocomputations showing that metarepresentation may be associated with activation of prefrontal and parietal regions [36], while low-level representation may be responsible for fast emotion recognition which depends on the amygdala [17, 55]. There is also convincing evidence of independence between these two types of knowledge representations from behavioral measurements of dissociation between correctness of performance in perceptual tasks and metacognitive awareness of such performance [33, 55] as well as neuronal instances of such dissociations in the brain [17, 56]. Common-sense intuition of brain activity also supports such view claiming that conscious knowledge about the stimulus does not relate to physical qualities of the perceived stimulus, but considers internal representations of the stimulus, which in turn refer to specific brain activation associated with stimulus perception [53]. It is worth mentioning that metacognition as higher level cognition, including monitoring, control processes, and evaluation, is sequential by nature [18]. Several computer modeling studies, for example, post-decision wagering procedures [57], demonstrate that metacognitive sequential strategies lead to consciousness of a stimulus. In the same vein, our brain study on metacognition with event-related potentials (ERPs) showed that metacognitive knowledge is crucial for conscious processing of emotion [58]. Similarly, a masking study with neural network simulations [54] shows that metacognitive knowledge can be underlined by a specific computational base for making conscious and unconscious decisions in terms of emotion detection. Unquestionably, empirical studies on consciousness and metacognition linked to the problem of accuracy of metacognitive knowledge, and its neurobiological and computational basis suggests that HOT is a theory that can be empirically verified.

Here, it is important to indicate that Szczepanowski [33] has shown that the relation between consciousness and emotion predicted by HOT can be modeled numerically with a signal-detection theory. In particular, SDT modeling has shown that under the emotion detection condition, subjective experience that expresses subjective feelings that accompany the first-order representation of affective stimuli can be embraced in the model by including participant's confidence responses [33, 55]. With regard to such SDT and HOT views, metacognition about task performance can be measured with a secondary sensitivity parameter $d'_{\text{Type II}}$ (see **Figure 2**). Evaluation of metacognition can include also $c_{\text{Type II}}$ parameter which is a second-order bias that identifies metacognitive strategies leading either to under- or overconfidence in task performance evaluations [46]. In this way, the second-order SDT measurements of consciousness provide objective information on subjective feelings of perceived affective stimuli.

In fact, Szczepanowski [33] has demonstrated that the SDT model of consciousness can embrace a hierarchical organization of affective processing, that is, objective information of performance in the emotion detection task must be reflected in a hierarchically higher level of processing. In this computational model of consciousness, there is an objective sensitivity measure of the perceived affective information (e.g., parametric first-order sensitivity $d'_{\text{Type I}} > 0$ or nonparametric $A'_{\text{Type I}} > 0.5$) as well as an objective measure of metacognition (e.g., parametric second-order sensitivity $d'_{\text{Type II}} > 0$ or nonparametric $A'_{\text{Type II}} > .5$ indices). The validity of this hierarchical SDT model was empirically proved with visual masking experiments with emotional faces (e.g., [35, 57]). In fact, the modeling outcomes based on SDT show that human consciousness with accompanying

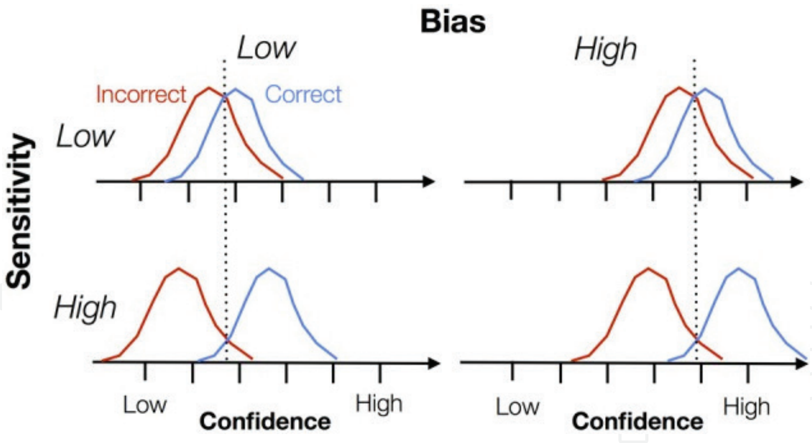

Figure 2. Measurements of consciousness and metacognition with second-order sensitivity and bias parameters based on SDT (source: [46]).

cognitive processes during detection of affective states may be a subject of empirical research. In addition, the interaction between consciousness and emotion can be related to decision-making processes which may be a result of computational-cognitive processes in the brain [33] and therefore potentially implemented into artificial environments of the social robot.

The above premises suggest that the hierarchical SDT model of consciousness can be used to determine computational correlates of robot's consciousness and its subjective experience of emotion. According to such SDT view, subjective conscious feelings of the robot may be related to execution of second-order operations on internally generated information from previous processing linked with detection of incoming stimuli from environment registered by a sensory system of the robot. We therefore assume that such conceptualization of machine consciousness within the robot is necessary to effectively regulate robot's behavior in terms of participation of metacognition in executing conscious control of cognition, in response to emerging affective information [18, 59].

8. Threshold approach to establish access consciousness for encoding C-E interaction by social robot

The third research domain for encoding C-E interactions by a social robot is to determine a computational correlate of global processing mode for consciousness of emotion. We assume that an adequate implementation of global information processing of emotion in the robotic system can be enabled by a threshold theory [33, 37]. As many experimental studies have demonstrated, representation of affective information is preferred to be accessed to conscious processing [60, 61]. For instance, in the area of consciousness research, backward masking studies provide substantial evidence that visual awareness occurs in the "all-or-none" fashion [62]. In the context of the masking task, this indicates that during stimulus detection, there is some sudden stepping-like burst of activation due to an incoming stimulus to enable transition between nonconscious and conscious states [63]. Some researchers suggest that such specific activation occurs in the brain as a threshold needed to activate access consciousness (see, for instance,

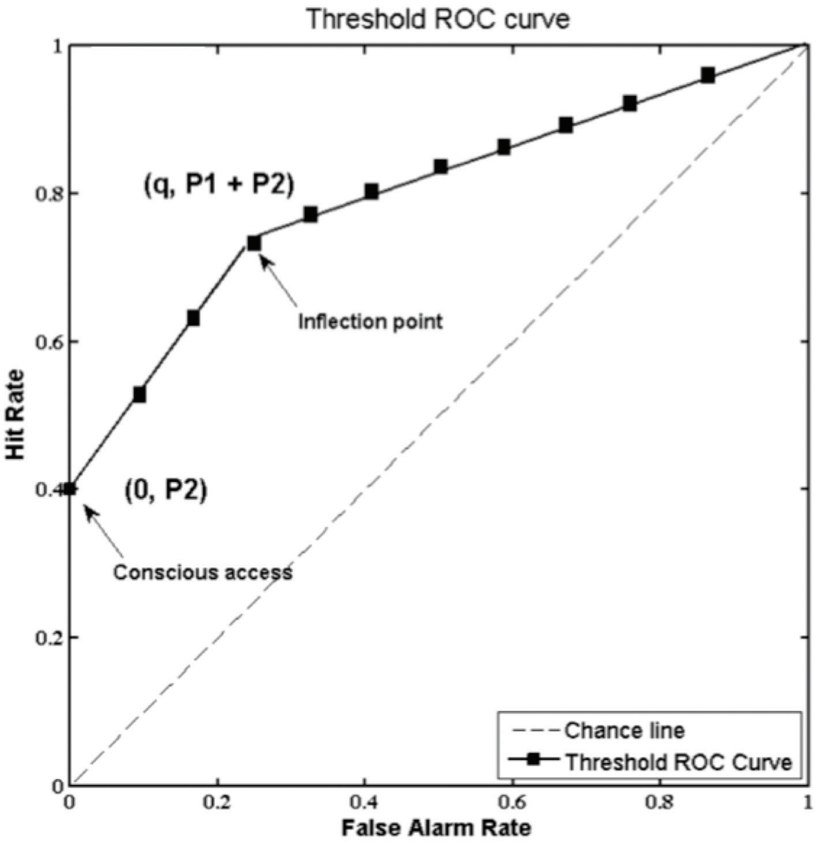

Figure 3. Linear ROC curve predicted by the Krantz's threshold model (source: [37]).

[63]). Indeed, Szczepanowski [33] has demonstrated that under a backward masking task, perception of fearful face happens in the "all-or-none" fashion and may be a factor explaining why this emotion information is preferable to conscious access. In particular, it appeared that in the visual masking experiments, several participants presented characteristic patterns of metacognition in terms of confidence in such a way that for the highest confidence, there are almost always hits without false alarms [37]. Because such observed response patterns followed ideal observer's behavior (hit responses without highest false alarms), the masking data have been successfully modeled with a Krantz's threshold detection theory [32, 33, 37]. This computational evidence that conscious perception of emotion is threshold-like processing implicates that under conditions in which stimulus strength is sufficiently large, the information content of the stimulus may be broadcasted in the system globally. This threshold-like information-processing approach to consciousness suggests that decision-making underlying emotion perception follows a discrete mental states' arrangement and its corresponding probabilities in terms of establishing conscious behavioral responses to affective information. Therefore, according to the outcomes from the threshold model, conscious processing in detecting emotion can activate global access to knowledge about emotion that manifests itself in ideal behavior of the observer.

The abovementioned outcomes suggest that global access to affective content in terms of metacognition (meta-knowledge) involves thresholds [33]. In other words, access consciousness may be activated for the highest confidence ratings on the "all-or-none" basis. In this way, conscious access to representation of emotions and metacognition can be quantified with signal parameters predicted in the Krantz model [32]. In the three-state threshold model (see

Figure 3), there are three mental states associated with perception, that is, the absence of ~D detection, D detection, and D* superdetection, and two thresholds, that is, upper and lower ones [32]. Detection of a target stimulus (probabilities P1 and P2) leads to mental states of D and D * (detection and superdetection), while detection of stimulus noise, described by the probability q, leads to a lack of detection ~D. The decision space described in the threshold detection theory is rectangular, and the ROC curve is linear as shown in [33]. It was demonstrated that participant who can consciously access to the stimulus content produces ideal observer behavior that can be estimated the P2 parameter [33]. Hence, the threshold model can predict situations in which the highest confidence is generated when there is conscious access to emotion content. Indeed, computational evidence for the threshold-like processing is an important discovery, since, so far, another view on perception has dominated in experimental psychology claiming that perception is continuous and should be described primarily by the Gaussian distribution [31]. Thus, in our conceptual framework of machine consciousness, we assume that conscious detection of emotion by the social robot engages global processing mode in the "all-or-none" fashion, and we propose to model these C-E interactions with the use of an innovative computational approach based on the Krantz's threshold theory [32].

9. Conclusions

As opposed to a typical application of industrial robots, a social robot needs to be considered as a social being with whom humans should be cooperating given a specific task structure. Therefore, the basic research aims of social robotics should be to determine computational models of the consciousness-emotion interaction designed to be implemented into a robotic platform. The request of preciseness in the context of computational models of emotions requires more research including related areas such as models of C-E interactions. This is a new research area in social robotics, and therefore it is potentially attractive from the perspective of development of computational models of emotion that are suitable for implementation in robots and contribute a new quality to the behavior of robots. It is believed that extending social robot competences and functionalities of HRI with C-E interactions will result in increasing acceptability of the social robot by the end user.

It seems that the abovementioned modeling outcomes of the C-E interaction based on the signal and threshold approaches are original contributions not only in the field of cognitive psychology but are crucial in the area of social robotics in terms of the possibility to implement high-level cognition into a social robot that effectively processes HRI in social domain [3, 4, 41, 64, 65]. In our conceptual framework, consciousness of emotion is the ability to detect affective information in the forced-choice condition, regardless choice decisions are low-level representation (features of the stimulus) or metarepresentation (subjective knowledge). In this way, consciousness may be attributed to an extremely simple function that can be associated with detection of different types of signals in the mind [33] and simply implemented into a social robot's design. In fact, adoption of the computational approach to consciousness that are based on quantitative detection parameters indicates that consciousness along with its subjective aspect is a specific function of the human brain and can be implemented into an artificial social robot's construction.

We believe that simplicity of such signal and threshold detection approaches that allow studying consciousness and its accompanying perceptual and metacognitive processes with the quantitative analysis will be optimal for implementing the C-E interactions into a social robot's system. Our successful attempts to operationalize desirable C-E interactions in the social robot within the signal-detection and threshold frameworks may provide valuable guidelines for implementation formal characteristics of conscious behavior into a social robot's construction and subsequently will be generalized for a much broader area of HRI. Finally, our understanding of cognitive mechanisms underlying consciousness and its subjective aspects will be the input to advance cognitive sciences, including philosophy of mind. In this way, our project will build a cross-disciplinary approach in designing effective HRI and machine consciousness that combine cognitive sciences and social robotics.

Acknowledgements

We would like to thank our respective universities and departments for the financial support. This research was partially funded by Institute of Psychology at University of Zielona Gora (Poland) and Wrocław University of Science and Technology (Poland) under Grant No. 0401/0142/17 and Grant No. 0401/0098/17.

Conflict of interest

Remigiusz Szczepanowski declares that he has no conflict of interest. Małgorzata Gakis declares that she has no conflict of interest. Krzysztof Arent declares that he has no conflict of interest. Janusz Sobecki declares that he has no conflict of interest.

Author details

Remigiusz Szczepanowski[1]*, Małgorzata Gakis[2], Krzysztof Arent[3] and Janusz Sobecki[4]

*Address all correspondence to: rszczepanowski@uz.zgora.pl

1 Institute of Psychology, Faculty of Education, Psychology and Sociology, University of Zielona Góra, Zielona Góra, Poland

2 Faculty of Psychology in Wroclaw, SWPS University of Social Sciences and Humanities, Wroclaw, Poland

3 Department of Cybernetics and Robotics, Electronics Faculty, Wroclaw University of Science and Technology, Wroclaw, Poland

4 Department of Computer Science, Faculty of Computer Science and Management, Wroclaw University of Science and Technology, Wrocław, Poland

References

[1] Breazeal C, Dautenhahn K, Kanda T. Social robotics. In: Siciliano B, Khatib O, editors. Springer Handbook of Robotics. 2nd ed. Cham: Springer International Publishing; 2016. pp. 1935-1972. DOI: 10.1007/978-3-319-32552-1_72

[2] Breazeal C. Emotion and sociable humanoid robots. International Journal of Human-Computer Studies. 2003;**59**:119-155. DOI: 10.1016/S1071-5819(03)00018-1

[3] Scassellati B. Theory of mind for a humanoid robot. Autonomous Robots. 2002;**12**(1):13-24. DOI: 10.1023/A:1013298507114

[4] Lemaignan S, Warnier M, Sisbot EA, Clodic A, Alami R. Artificial cognition for social human–robot interaction: An implementation. Artificial Intelligence. 2017;**247**:45-69

[5] Marsella S, Gratch J, Petta P. Computational models of emotion. In: Scherer KR, Banziger T, Roesch E, editors. Blueprint for Affective Computing (Series in Affective Science). 1st ed. Oxford University Press; 2010. pp. 21-46. ISBN: 9780199566709

[6] Baron-Cohen S. Mindblindness: An Essay on Autism and Theory of Mind. Cambridge: MIT Press; 1997

[7] Leslie AM. ToMM, ToBY, and Agency: Core architecture and domain specificity. In: Hirschfeld LA, Gelman SA, editors. Mapping the Mind: Domain Specificity in Cognition and Culture. Cambridge: Cambridge University Press; 1994. pp. 119-148

[8] Dias J, Mascarenhas S, Paiva A. FAtiMA modular: Towards an agent architecture with a generic appraisal framework. Emotion Modeling. Springer International Publishing; 2014. pp. 44-56

[9] Arent K, Tchoń K. Roboty społeczne—Postępy robotyki Prace Naukowe Politechniki Warszawskiej. Elektronika. 2012;**182**(2):629-648 (in Polish)

[10] Bach J. Principles of Synthetic Intelligence PSI: An Architecture of Motivated Cognition. New York: Oxford University Press; 2009. Oxford Scholarship Online, 2009. DOI: 10.1093/acprof:oso/9780195370676.001.0001

[11] Marsella SC, Gratch J. EMA: A process model of appraisal dynamics. Cognitive Systems Research. 2009;**10**:70-90

[12] Fong T, Nourbakhsh I, Dautenhahn K. A survey of socially interactive robots. Robotics and Autonomous Systems. 2003;**42**(3):143-166

[13] Itti L, Koch C. Computational modelling of visual attention. Nature Reviews Neuroscience. 2001;**2**:194-203

[14] Pessoa, Luiz. Do Intelligent Robots Need Emotion?. Trends in Cognitive Sciences, 2017;**21**(11):817-819

[15] Treisman AM, Gelade G. A feature-integration theory of attention. Cognitive Psychology. 1980;**12**:97-136

[16] Itti L, Koch C. A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research. 2000;**40**(10):1489-1506

[17] LeDoux JE, Brown RA. Higher-order theory of emotional consciousness. Proceedings of the National Academy of Sciences. 2017;**14**(10):2016-2025

[18] Nosal CS. Psychologiczne Modelu umysłu [Psychological Model of the Human Mind]. Warszawa: PWN; 1990

[19] Tsuchiya N, Adolphs R. Emotion and consciousness. Trends in Cognitive Sciences. 2007;**11**(4):158-167

[20] Pessoa L. The Cognitive-emotional Brain: From Interactions to Integration. Cambridge: MIT Press; 2013

[21] Mitchell DG, Greening SG. Conscious perception of emotional stimuli brain mechanisms. The Neuroscientist. 2012;**18**(4):386-398

[22] Fodor J. The Modularity of Mind. Cambridge, MA: MIT Press; 1983

[23] Rumelhart DE, McClelland JL. Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Foundations. Vol. 1. Cambridge, MA, USA: MIT Press; 1986

[24] Block N. On a confusion about a function of consciousness. Behavioral and Brain Sciences. 1995;**18**:227-247

[25] Reggia JA, Huang DW, Katz G. Exploring the computational explanatory gap. Philosophies. 2017;**2**(1):5

[26] Reggia JA. Conscious machines: The AI perspective. In AAAI Fall Symposium Series; September, 2014; North America

[27] Chalmers DJ. The Conscious Mind: In Search of a Fundamental Theory. New York, Oxford: Oxford University Press; 1996

[28] Sun R, Franklin S. Computational models of consciousness. In: Zelazo P, Moscovitch M, editors. Cambridge Handbook of Consciousness. New York: Cambridge University Press; 2007. pp. 151-174

[29] Reggia JA. The rise of machine consciousness: Studying consciousness with computational models. Neural Networks. 2013;**44**:112-131

[30] Searle J. Mind: A Brief Introduction. New York: Oxford University Press; 2004

[31] Macmillan NA, Creelman CD. Detection Theory: A User's Guide. (Mahway, New Jersey: Lawrence Erlbaum Associates, Inc.); 2005

[32] Krantz DH. Threshold theories of signal detection. Psychological Review. 1969;**76**(3): 308-324

[33] Szczepanowski R. Świadome i nieświadome przetwarzanie emocji w mózgu. Modelowanie w ramach teorii detekcji sygnałów [Conscious and unconscious processing of emotion in the brain. Modeling with signal detection approach]. Warsaw: PWN; 2014

[34] Cleeremans A. Computational correlates of consciousness. Progress in Brain Research. 2005;**150**:81-98

[35] Szczepanowski R, Pessoa L. Fear perception: Can objective and subjective awareness measures be dissociated. Journal of Vision. 2007;**7**(4):10

[36] Lau HC, Rosenthal D. Empirical support for higher-order theories of conscious awareness. Trends in Cognitive Sciences. 2011;**15**(8):365-373

[37] Szczepanowski R. Conscious access to fear-relevant information is mediated by threshold. Polish Psychological Bulletin. 2011;**42**(2):56-64

[38] Baars BJ. The conscious access hypothesis: Origins and recent evidence. Trends in Cognitive Sciences. 2002;**6**(1):47-52

[39] Dehaene S. Consciousness and the brain: Deciphering how the brain codes our thoughts. New York: Penguin. 2014. ISBN: 978-0-670-02543-5

[40] Szczepanowski R, Traczyk J, Fan Z, Harvey L Jr. Preferential access to emotion under attentional blink: Evidence for threshold phenomenon. Polish Psychological Bulletin. 2015;**46**(1):127-132

[41] Picard RW, Vyzas E, Healey J. Toward machine emotional intelligence: Analysis of affective physiological state. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2001;**23**(10):1175-1191

[42] Lau HC. A higher order Bayesian decision theory of consciousness. Progress in Brain Research. 2007;**168**:35-48

[43] de Gelder B, Pourtois G, van Raamsdonk M, Vroomen J, Weiskrantz L. Unseen stimuli modulate conscious visual experience: Evidence from inter-hemi-spheric summation. Neuroreport. 2001;**12**(2):385-391

[44] Seth AK, Dienes Z, Cleeremans A, Overgaard M, Pessoa L. Measuring consciousness: Relating behavioural and neurophysiological approaches. Trends in Cognitive Sciences. 2008;**12**(8):314-321

[45] Eriksen CW. Discrimination and learning without awareness: A methodological survey and evaluation. Psychological Review. 1960;**67**:279-300

[46] Fleming SM, Lau HC. How to measure metacognition. Frontiers in Human Neuroscience. 2014;**8**:443

[47] Rosenthal DM. Consciousness and Mind. Oxford: Clarendon Press; 2005

[48] Dennett DC. Are we explaining consciousness yet? Cognition. 2001;**79**:221-237

[49] Dennett DC. Sweet Dreams: Philosophical Obstacles to a Science of Consciousness. Cambridge: The MIT Press; 2005

[50] Koriat A. Metacognition and Consciousness. In: Zelazo P, Moscovitch M, editors. Cambridge Handbook of Consciousness. New York: Cambridge University Press; 2007. pp. 289-325

[51] Karmiloff-Smith A. Beyond Modularity: A Developmental Perspective on Cognitive Science. Cambridge, MA: MIT Press; 1992

[52] Timmermans B, Schilbach L, Pasquali A, Cleeremans A. Higher-order thoughts in action: Consciousness as an unconscious re-description process. Philosophical Transactions of the Royal Society B: Biological Sciences. 2012;**367**:1412-1423

[53] Cleeremans A. Frontiers: The radical plasticity thesis: How the brain learns to be conscious. Frontiers in Consciousness Research. 2011;**2**(86):1-12

[54] Szczepanowski R, Wierzchoń M, Szulżycki M. Neuronal network and awareness measures of post-decision wagering behavior in detecting masked emotional faces. Cognitive Computation. 2017;**9**(1):457-467

[55] Szczepanowski R. Signal detection approach in modeling consciousness-emotion interactions. Acta Neuropsychologica. 2017;**15**(1):89-96

[56] Lau HC. Are we studying consciousness yet? In: Weiskrantz L, David M, editors. Frontiers of Consciousness: Cichele Lectures. Oxford: Oxford University Press; 2008. pp. 245-258

[57] Szczepanowski R. Absence of advantageous wagering does not mean that awareness is fully abolished. Consciousness and Cognition. 2010;**19**(1):426-431

[58] Wierzchoń M, Wronka E, Paulewicz B, Szczepanowski R. Post-decision wagering affects metacognitive awareness of emotional stimuli: An event related potential study. PLoS One. 2016;**11**(8):e0159516

[59] Cichoń E, Szczepanowski R. Mechanizmy tłumienia niepożądanych odczuć i myśli w ujęciu metapoznawczym [Metacognitive approaches toward supression mechanisms of unwanted thoughts and emotions]. Rocznik Kognitywistyczny. 2015;**8**:79-89

[60] Milders M, Sahraie A, Logan S, Donnellon N. Awareness of faces is modulated by their emotional meaning. Emotion. 2006;**6**(1):10-17

[61] Yang E, Zald DH, Blake R. Fearful expressions gain preferential access to awareness during continuous flash suppression. Emotion. 2007;**7**(4):882

[62] Sergent C, Dehaene S. Is consciousness a gradual phenomenon? Evidence for an all-or-none bifurcation during the attentional blink. Psychological Science. 2004;**15**(11):720-728

[63] Dehaene, S. Conscious and Nonconscious Processes: Distinct Forms of Evidence Accumulation? In: Biological Physics. Rivasseau V, editor. Springer Basel; 2011. pp. 141-168. ISBN: 978-3-0346-0427-7

[64] Paiva A, Leite I, Ribeiro T. Emotion modeling for social robots. In Calvo R, D'Mello S, Gratch J, Kappas A, editors. The Oxford handbook of affective computing, New York: Oxford University Press; 2014. pp. 296-308

[65] Pereira A, Leite I, Mascarenhas S, Martinho C, Paiva A, Lamers MH, Verbeek FJ. Using Empathy to Improve Human-Robot Relationships, Human-Robot Personal Relationships: Third International Conference, HRPR. Berlin Heidelberg: Springer; 2010. pp. 130-138