Research Article

Laurentiu Vasiliu, Keith Cortis, Ross McDermott, Aphra Kerr, Arne Peters, Marc Hesse, Jens Hagemeyer, Tony Belpaeme, John McDonald, Rudi Villing, Alessandra Mileo, Annalina Capulto, Michael Scriney, Sascha Griffiths, Adamantios Koumpis, and Brian Davis*

CASIE – Computing affect and social intelligence for healthcare in an ethical and trustworthy manner

https://doi.org/10.1515/pjbr-2021-0026
received November 11, 2020; accepted June 24, 2021

Abstract: This article explores the rapidly advancing innovation to endow robots with social intelligence capabilities in the form of multilingual and multimodal emotion recognition, and emotion-aware decision-making capabilities, for contextually appropriate robot behaviours and cooperative social human–robot interaction for the healthcare domain. The objective is to enable robots to become trustworthy and versatile social robots capable of having human-friendly and human assistive interactions, utilised to better assist human users' needs by enabling the robot to sense, adapt, and respond appropriately to their requirements while taking into consideration their wider affective, motivational states, and behaviour. We propose an innovative approach to the difficult research challenge of endowing robots with social intelligence capabilities for human assistive interactions, going beyond the conventional robotic sense-think-act loop. We propose an architecture that addresses a wide range of social cooperation skills and features required for real human–robot social interaction, which includes language and vision analysis, dynamic emotional analysis (long-term affect and mood), semantic mapping to improve the robot's knowledge of the local context, situational knowledge representation, and emotion-aware decision-making. Fundamental to this architecture is a normative ethical and social framework adapted to the specific challenges of robots engaging with caregivers and care-receivers.

Keywords: social human–robot interaction, sHRI, computing affect, emotion analysis, healthcare robots, robot-assisted care, robot ethics

1 Introduction

One of the very distinct human intelligence abilities that distinguish us from machines is our ability to gauge, sense, and appropriately respond to emotions. However, ever increasing advances in AI and hardware technology are enabling machines to extract emotion from our verbal and non-verbal communication. Despite these advancements, there has been a narrow adoption of "emotion-aware technology" in social robotic applications due to the many scientific and technical hurdles involved. Many of these are related to the difficulty of dealing with the complexities of real-world human interactions, which has frequently resulted in poor results or even failure of non-robotic interactive AI applications. The complexities that robot applications focused on social human–robot interaction (sHRI) have to overcome is immense, resulting in most of the sHRI robots being more "robotic toys" than genuine social robots. The motivation for the research agenda we present in this article is to equip robots in the healthcare application area with multimodal affective

*** Corresponding author: Brian Davis,** School of Computing, Dublin City University, Dublin, Ireland, e-mail: brian.davis@dcu.ie
Laurentiu Vasiliu: Peracton Ltd., Dublin, Ireland
Keith Cortis, Ross McDermott, Alessandra Mileo, Annalina Capulto, Michael Scriney: School of Computing, Dublin City University, Dublin, Ireland
Aphra Kerr: Department of Sociology, Maynooth University, Kildare, Ireland
Arne Peters: Informatik 6 - Lehrstuhl für Robotik, Künstliche Intelligenz und Echtzeitsysteme Fakultät für Informatik, Technische Universität München, Munich, Germany
Marc Hesse, Jens Hagemeyer: Cognitronics & Sensor Systems Group, Center for Cognitive Interaction Technology (CITEC), Universität Bielefeld, Bielefeld, Germany
Tony Belpaeme: IDLab, Department of Electronics and Information Systems, Ghent University, Ghent, Belgium
John McDonald, Rudi Villing: Department of Computer Science, Maynooth University, Kildare, Ireland
Sascha Griffiths: NoosWare BV, Amsterdam, The Netherlands
Adamantios Koumpis: Berner Fachhochschule, Business School, Institute Digital Enabling, Bern, Switzerland

capabilities of enabling human-friendly and human assistive interactions that can be accomplished only by recognising the user's emotional state.

Smart interface technology is ubiquitous in all areas of our lives; we use conversational smart assistant interfaces like Amazon's Alexa, we use facial recognition for authentication, and we use our voice to control our connected devices. However, research shows that user interaction with smart interfaces that are not "emotionally intelligent" results in one-directional commands rather than genuine dialogue between humans and machines [1].

Unsurprisingly, previous sHRI applications have had limited adoption as they failed to live up to expectations and users found that their lack of empathy, social intelligence, and inability to understand context led to inappropriate responses or no response at all, eventually resulting in frustration and dissatisfaction [2,3]. The market is recognising the rising consumer demand for a more personalised experience, where a robot can recognise emotions (such as joy, trust, fear, surprise, sadness, anticipation, anger, and disgust) based on Robert Plutchik's eight basic emotions [4], considering not only what the user wants but also appreciating how they feel in that moment and modifying the interaction accordingly.

A user-centred design [5] is crucial to technology innovation and acceptance [6] and is a core part of our research, as new assistive interactive technology often fails, because factors which affect how humans perceive technology were not taken into account by developers at the design stage, or insufficient attention was paid to contextual barriers and ethical challenges. To enable meaningful and trustworthy social interaction with social agents, a person needs to perceive their dialogue partner as an autonomous entity, requiring both a physical presence and the possibility to directly interact and emote appropriately. This propensity to anthropomorphise increases meaningful social interaction between robots and people [7] and helps interactive assistive technologies succeed.

Our core premise is that future autonomous robots, from the simplest service robot to the most sophisticated individualised support robot, will all require some level of affective and social cognition to succeed in a dynamic and complex human-populated environment. We propose leveraging technologies and techniques from the fields of *Affective Computing* [8], *Natural Language Processing* (NLP) [9], *Computer Vision* (CV) [10], and *Complex Decision-Making in HRI* [11,12] to develop an "emotion-aware architecture" called **C***omputing* **A***ffect and* **S***ocial* **I***ntelligenc***E** (referred to in the text as "**CASIE**," which refers to a robot that makes use of the technology we propose). By allowing the robot to manage the complexities associated with real-world human interactions, "*CASIE robots*" can facilitate the adoption of assistive healthcare robotic applications.

1.1 Social robots in healthcare

The COVID-19 pandemic has clearly demonstrated that our healthcare systems and workforce are operating close to their limits, and in some EU regions, even before the current crisis, some countries' vulnerability to future shocks and stresses had already been identified in the "2019 State of Health in the EU" report [13]. There is now an urgent need to adopt innovative technologies that can help reduce workload and stress on the health systems and healthcare professionals, we need to be better prepared for the next crisis.

The number of people who could benefit from social robots used in healthcare is vast, the following applications shown in the following list have been identified as particularly promising for increased adoption of social robots. Not because we already have operational social robotics solutions, but because these healthcare settings have been extensively explored in recent years using mock-ups and remote control, i.e. non-autonomous or semi-autonomous robot prototypes [14].

1. **Hospitals**:
   (a) Offering support to patients, such as companionship, informing patients, encouraging them to adhere to a healthcare programme using social assistance [15].
   (b) Robots that are able to evaluate aspects of the current physical state of the patient in psychiatric clinics.
   (c) Interactive robots for reception and waiting rooms of hospitals and doctors' offices.
2. **Nursing homes**:
   (a) Helping residents to be more independent, supporting residents through offering entertainment and diversion, monitoring residents (specifically residents with dementia), providing companionship, and supporting health promoting activities [16–18].
   (b) Emotion-aware robots deployed in elderly care settings to enable more socially appropriate interactions with users based on their facial expression and emotions in speech.

3. **Care facilities and home use**:
   (a) Assisting people with cognitive impairments, such as autism spectrum disorders [19–21].
   (b) Socially and emotionally-aware robots that can help people in their daily life, such as dealing with loneliness and anxiety.

As we design sHRI applications for the healthcare domain, we must consider the effects that these robots can have not only on the care-receiver but also on the caregiver, how the robot fits into the overall network and dynamics of the user's social relationships, and, most importantly, when and where their use is appropriate and ethical [22]. Social robots have been shown to improve social engagement, reduce negative emotions and behavioural symptoms, and promote a positive mood and quality of care experience [23]. Patients who use socially assistive robots in a patient-centred manner are perceived to have higher emotional intelligence [24,25], which can influence caregivers to form a more favourable impression of the patient, directly leading to an improvement in the quality of care a patient may be given [26,27].

Basic companion/service robots have shown that they can improve users' quality of life, social and cognitive health, mitigate depression, increase social connectedness and resilience, and reduce loneliness [28]. In particular, the efficacy of companion/service robots used in care settings for people with dementia has been validated, even when the robot lacks emotion-aware capabilities [29]. These results demonstrate that sHRI applications can further improve care in healthcare settings where companion/service robots have already been implemented and enable new ones where a companion/service robot would have no impact. For example, social robots are being used in novel ways to improve human–human interactions.

Inspired by the context above, we propose enabling a robot to sense, analyse, and interpret an individual's behaviour and mood from what they say and how they say it, from their speech, subtleties of tone, facial expressions, micro-expressions, gestures, and body language. The recognition of the nuances of non-verbal communication is essential for meaningful sHRI; they influence how messages are perceived and understood. For example, reading body language is integral to how we navigate social situations. Empowering a robot to recognise all these verbal and non-verbal communications enables the robot to respond more appropriately with emotion-aware behaviours, communication, and social interaction. This social intelligence capability empowers robots to interact more naturally with people in everyday real-world scenarios, hence further increasing the quality of the sHRI, allowing them to be deployed in new domains and applications which require social intelligence while also delivering a contextually appropriate interactive experience and not a standard one-directional command interaction.

The growth and demand for advanced social robotic applications were highlighted in a Microsoft Research report on AI [30], where the combination of robotics and AI to perform advanced tasks was ranked second only to machine learning as the most useful technology for European companies deploying AI solutions. The report emphasises the importance of social intelligence capabilities for building the future AI applications. However, social intelligence competencies were listed last, emphasising the scarcity of available resources and knowledge, which in part is limiting the adoption of new sHRI applications.

The pandemic is motivating hospitals and healthcare facilities to implement autonomous robotic systems more than ever. It is critical, particularly in close-contact situations, that these robots collaborate in a socially intuitive and trustworthy manner. They must be capable of perceiving human emotions, intentions, social boundaries, and expectations. These features will help humans to feel more secure, comfortable, and amiable when interacting with robots. In light of the pandemic, the robotics community issued a call to action for new robotic solutions for public health and infectious disease management, with a particular emphasis on increased adoption of social robots [31], as quarantine orders have resulted in prolonged isolation of individuals from social interaction, which may have a detrimental effect on their mental health. To tackle this problem, social robots could be deployed in healthcare and residential settings to maintain social interactions. The authors acknowledge the challenges inherent in achieving this goal, as social interactions require the development and maintenance of complex models of people, including their knowledge, beliefs, and emotions, as well as the context and environment in which they occur, which would be too challenging for the current generation of social robots architecture.

Europe's healthcare systems are becoming overburdened with numerous problems due to ageing populations, disparities in medical systems and social protection across countries, and crippling medical events that can put the global medical community under tremendous strain. Getting enough healthcare staff for the future will become an increasing challenge. In many cases, medical jobs are unappealing because of the low pay, night shift work, long hours, and the risk of being exposed to harmful viruses. The World Health Organization estimated in a 2016 study on a global strategy on human resources for health [32] that the expected healthcare staff shortage for the EU28 alone will reach 4.1 million in 2030, which includes 600k physicians, 2.3 million nurses, and 1.3 million other health care professionals [33]. In Europe, health workforce imbalances and shortages have long been a problem, and despite recent increases in workforce numbers, this improvement will not be sufficient to meet the needs of ageing populations. For example, this increased healthcare demand is projected to require up to 500k additional full-time healthcare and long-term care staff in Germany by 2030. In light of these circumstances, we argue that robotics can be an effective tool for resolving future staff issues. They can assist with specific social and care tasks while allowing the staff to focus on their core competencies and functions.

2 Studies and investigations

2.1 Bringing soft skills to robots

The next generation of social robots must be trustworthy, contextual, and culturally aware to provide meaningful assistance to the healthcare industry. The research agenda outlined in this article can significantly contribute to overcoming the challenges of enabling a robot to have human-friendly, human assistive, and emotion-aware interactions, accelerating the adoption of AI and social robots applications in healthcare and beyond.

Robotic deployment in healthcare applications is not a simple task, as each medical environment and medical condition presents unique challenges and varies in terms of legal and regulatory requirements. However, all of them require social and emotional recognition and concurrent adaptability. For example, in a nursing home setting, a care-receiver may be seeking treatment, guidance, or simply some small talk. The robot must be able to recognise their emotions and react appropriately and compassionately within a given context. This could be accomplished by the robot speaking slower and louder to someone who has difficulty hearing, slowing down when guiding people with walking disabilities, and using appropriate gestures, words, and tone during a conversation. As a result, emotion-aware social robots must be fundamentally aware of their current situation and capable of contextualising new information. Additionally, they must remember people and be capable of adapting to their current users over time.

2.2 User informed design

From a user perspective, CASIE will need to interact with two distinct types of end users, each with distinct needs and expectations of the system, as illustrated in Figure 1.

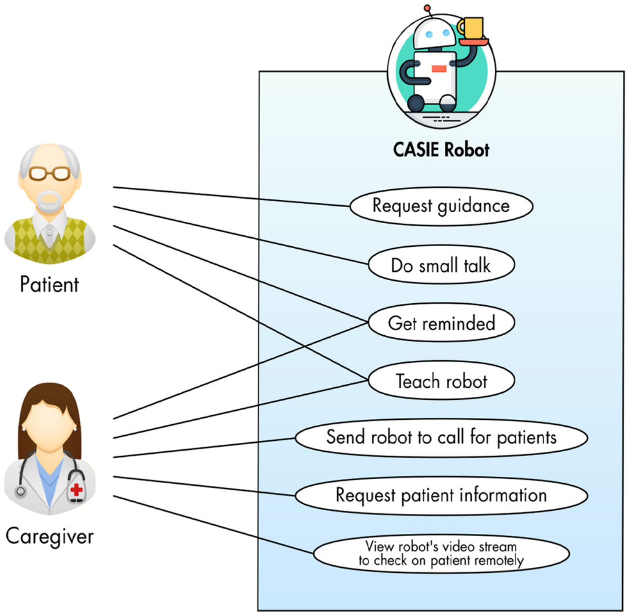
CASIE Robot
Patient
Caregiver
Request guidance
Do small talk
Get reminded
Teach robot
Send robot to call for patients
Request patient information
View robot's video stream to check on patient remotely

Figure 1: Exemplary list of role-specific use cases from a list of potential applications.

1. **Care-receiver**: The first group are patients or residents of hospitals or nursing homes, which we consider the care-receivers. The robot is responsible for their well-being and being their primary means of contact or assisting the doctors with their care. It can connect them to medical staff and be a social connection to the outside world (e.g. for isolated elderly in a nursing home).
2. **Caregiver**: Additionally, caregivers can benefit from robotic assistance. These may include hospital doctors and nurses, nursing home staff, and family members who deploy a robot assistant to assist their elderly relatives at home. For this group of users, the robot serves as a tool rather than a companion. Caregivers prefer direct interaction, assigning specific tasks to the robot and expecting direct access to the robot's knowledge base.

Analysis of care-receiver data, particularly from private conversations between care-receivers and robots, holds enormous potential for treatment improvement, as patients share information in a completely different way when communicating with robots. When speaking with a human doctor, a subconscious need to justify and explain oneself arises, because people are conscious that robots do not judge; they tend to be more honest [6]. Thus, the data gathered by a social robot may make a significant difference in the treatment of a variety of medical conditions such as mental health problems.

Importantly, Western societies' healthcare and professional care workforce's are generally highly feminised. The history of technology, particularly healthcare technology, has revealed implicit and explicit gender biases and stereotyping in technology design [34,35]. Additionally, men and women express emotions, feelings, and illness symptoms differently, and their expressions vary with specific illnesses. The World Health Organization asserts that gender plays a critical role in mental health and illness [36].

A user-centred approach that considers the role of gender in technology development ensures that the robotic platform is informed from the start by caregiver and care-receiver knowledge, and the iterative development of the architecture in tandem with the intended use cases enables the developers to adjust for unintended gender or other biases.

3 Comments on studies

3.1 State-of-the-art for the addressed disciplines and fields

The successful deployment of CASIE in healthcare settings depends on a number of critical key aspects. The most important one is the user acceptance of both care-receivers and caregivers. Besides ethical elements, this also involves the system's reliability and ease of use. CASIE robots should be simple enough to allow roll-out after a single day workshop with caregivers, such as medical staff, who need to be able to operate the system without a technical background, including updating the system's configuration (e.g. updating the map of the environment), installing software updates, and even doing minor repairs.

Because the CASIE is focused on robots in healthcare settings, a new level of robustness in robotics is required. A CASIE robot will constantly be required to recognise and manage situations it has never encountered before, such as new patients with unique behaviours, mixed languages or dialects, and changes in the situation while maintaining short response times to facilitate fluent conversations. To illustrate, studies show that human speakers have extremely fast response times, from 250 ms, depending on the spoken language, and frequently interrupt their dialogue partner before their sentences are completed [37]. Given that many of today's state-of-the-art NLP systems are cloud-based or edge-based, a CASIE robot should provide basic integrated language processing functionality as a failsafe. On the other hand, CASIE will be required to support external systems via interfaces (e.g. an appointment calendar of a hospital).

Ideally, CASIE robots would be affordable to consumers, allowing for widespread adoption of social robot assistants. We can circumvent this issue by making CASIE as hardware-independent as possible, allowing it to run on a wide variety of current robot platforms. Depending on technological advancements, the system could be added to lower-cost consumer robots. However, a CASIE robot must earn the trust of both end-user groups. While this is influenced by a variety of factors, including the system's reliability and ease of use, it is heavily influenced by emotional awareness and the ability to take appropriate actions. Developing an emotion-aware architecture for robots pushes the boundaries of several technical, ethical, and legal disciplines. As such, we view progress in this field of research through the lens of the following nine complementary and overlapping challenge areas.

3.2 Challenges

3.2.1 Challenge 1: Affective (emotive) speech and language processing

While processing emotion from speech is difficult, it is necessary for empathic communication. Detecting and comprehending emotion in speech are critical for computer-assisted technologies [38], as it is determining the speaker's intent [39]. Additionally, speech synthesis enhances the effectiveness of machine–human interactions [40]. When it comes to sHRI in a specific domain, such as health, noisy incomplete spoken language input presents a number of difficulties.

While these issues are typically resolved when processing edited texts (e.g. web news), they become significantly more problematic when analysing short, noisy text utterances (incomplete sentences, missing words, speech processing errors). For linguistic processing tasks such as tokenisation, sentence boundary detection, part-of-speech tagging, and syntactic parsing, such noisy input complicates attempts to recognise, classify, and connect concepts/entities within linguistic content to a knowledge base for concept/aspect-based emotion analysis of text (using opinion mining techniques) [41], which requires associating an emotion with a specific target entity.

Developing human-like dialogue-based NLP presents particular challenges in addition to those mentioned previously, including real-time processing in accordance with human language processing time frames, concurrent language processing to enable the generation of responses during an ongoing utterance, the ability to process multimodal linguistic cues, for example, deictic terms accompanied by body movements which constrain possible interpretations of linguistic expressions, and bi-directional exchange of information flow [42].

Handling abusive language, such as offensive, obscene, culturally, and socially insensitive remarks, changing the subject, and detecting utterances spoken in multiple languages are also well-known challenges when processing human-to-robot dialogue. Additionally, extracting the necessary features for emotion recognition from speech can take several dozen seconds per utterance, which can be overcome using deep learning algorithms [43]. These approaches have significantly advanced dialogue generation, particularly in terms of social intelligence [44].

The challenges are exacerbated further when dealing with noisy and domain-specific non-English input. This raises the following research questions: how do you develop native emotion analysis applications and neural language models in the absence of sufficient language resources? And how, in this context, can Machine Translation be used to support domain-specific, concept-based multilingual emotion analysis of short text content?

3.2.2 Challenge 2: Spatial perception

While much information can already be obtained from linguistic interaction, CASIE's focus is also on the visual perception of humans and the robot's environment. In addition to the voice analysis described above, both facial and body pose recognition to understand a user's intentions and emotional state is required.

For humans to accept robots as socially intelligent entities, they must exhibit social intelligence in several forms. Recent advances in deep neural networks have led to a step change in the performance of emotion classification, person detection [45], body pose estimation [46] algorithms, and, therefore, a CASIE robot will have to incorporate such advances as a core part of its perception system allowing it to work effectively with and among people. AI facial coding technology for recognising basic human emotions and attention states through a combina tion of a Convolutional Neural Network and a Temporal Convolutional Network is well established but has had limited adoption in healthcare robotics applications.

It is important that a socially intelligent robot can move around in human environments. The approach we favour for CASIE robots is to feature a modern simultaneous localization and mapping (SLAM) system. While many working solutions for indoor SLAM have already been demonstrated, our CASIE conceptual model is about interconnecting mapping, object recognition, and the robot's knowledge base while considering the limitations of our target platforms, both with regard to available sensors and computing power. The last decade has seen research in SLAM move towards handling dynamic environments [47], numerous different approaches have been demonstrated, such as deforming the scene in a rigid as possible method [48], estimation of joints [49], or warp fields [50]. As CASIE robots are intended to face moving people, or even beds, and bigger objects being moved around, the framework we suggest requires taking temporal factors into account to extract the actual minimal map over time while tracking certain objects over extended periods.

Moreover, CASIE robots will need to navigate large environments, such as nursing homes or hospitals, making the implementation more complex due to memory and computing power limitations. Further integration with object recognition techniques is necessary to enable robots to access contextual knowledge and investigate methods for simplifying the teaching of new objects. A critical skill for a socially intelligent robot is navigating in a socially acceptable manner and may optionally include escorting or guiding someone to a destination. Given the importance of people in the context of the robot's operation, we will build on recent advances in person detection and body pose estimation to compute a social map of the robot's environment to augment semantic and geometric maps with suitable human location, pose, and dynamics data. This will entail combining the estimated 2D pose with the output of the depth channel of the robot's RGBD sensor to upgrade the pose to a full 3D model, allowing the resulting data to be grounded relative to the robot's model of the environment. We will extend the social map with a predictive model of human dynamics, initially based on filtering the body poses. The overall aim of our approach will be to extend traditional robot navigation solutions to observe social rules regarding proxemics [51], approaches, interactions, guiding, and following.

3.2.3 Challenge 3: High-level control

To respond in a contingent manner to interaction events, and specifically to the emotional and affective states of the user, CASIE robots will require a control mechanism that is sensitive to these aspects of the external world. While low-level aspects of the robot's control (such as dialogue management, social navigation, or non-verbal behaviour) or delegated to low-level control mechanisms, a high-level control mechanism is needed to drive the robots' long-term behaviour. One feasible approach is to rely on non-deterministic finite state machines to switch between different behaviours [52]. However, while this approach can handle small-scale interaction in which the programmer can foresee most actions, we expect that long-term interaction in complex domains will require a robotic planner. The novel aspect here is making decisions and chunked plans on affective data, handling incomplete information, and managing potential conflicting decision resolutions. Automated reasoning with incomplete information, sometimes referred to as default reasoning, focuses on computational methods to efficiently generate non-deterministic solutions, and then pruning such solutions based on preferences (or penalties) to rank possible final outcomes. Default reasoning has only recently been applied for handling streaming data with substantial limitations in scalability (e.g. the LARS framework [53]). On the other hand, reasoning under uncertainty requires the handling of knowledge and inference mechanisms that are probabilistic in nature.

These two approaches, traditionally used to solve different problems, will be combined to handle incompleteness and uncertainty in dynamic scenarios. In healthcare scenarios such as those described in Section 2.2, there is a need for a scalable hybrid approach of this sort that can consider qualitative and quantitative aspects of dynamic reasoning combined with multiple real-time criteria for complex decision-making.

Such approaches can be seen as an advantage only if we can deal with the potential reduction in the quality of information represented by incompleteness and uncertainty. However, decision-making may fall short as it is not able to generate plans and reason about potential future outcomes of actions. The main challenge is represented by the interplay between logical and probabilistic inference to help reduce the complexity of logical reasoning and support learning from observations.

3.2.4 Challenge 4: Knowledge base

The CASIE requirements pose a technical challenge of determining an appropriate system architecture that enables the storage and query of knowledge and provides a sufficiently detailed API for other components and processes. Numerous research questions arise as a result of this: What is an appropriate graph model? A knowledge base is a graph of vertices and edges. There are numerous ways to represent such a structure within a system [54]. There is no one "best fit" model, and a suitable model is derived from a combination of the system requirements and the underlying data which resides in the knowledge base. How to determine an efficient path through the knowledge graph? Querying a graph requires a graph traversal, which can be a time-consuming process. An efficient query processor requires the ability to prune the available state space of all edges within the graph to minimise the number of possible paths that can satisfy a query.

Furthermore, frequently accessed paths can be indexed [55] using polyglot persistence [56] to minimise query processing time. How to learn from historical queries to predict and cache frequently posed queries? A common method in database optimisation is the caching of frequently stored queries in memory for instant retrieval. Due to the high number of queries being posed to the graph and the requirement to respond effectively, a query-cache is required. Within a graph, this involves identifying sub-graphs [57] that are frequently visited during query processing and caching these in memory.

3.2.5 Challenge 5: Multimodal data fusion

To generate a meaningful and engaging affective dialogue, a robot must be able to interact with humans on a shared sensory level, where communication is often enriched by facial expressions, body language, voice pitch, and the context in which the communication occurs. A dialogue system must be capable of capturing and aggregating all of these stimuli in order to direct the system's response. Apart from the numerous challenges involved in processing and extracting relevant data from each of these sources, this task entails additional difficulties associated with synchronising these disparate streams and selecting relevant portions to include. To extract emotions from multiple modalities, it is necessary to model each source's temporal dynamics and align the extracted features [58].

All of this pre-processing must occur in real-time, burdening these systems with complexity when dealing with large amounts of user-generated data. There is also a personal dimension to detecting emotions from heterogeneous sources: there is no standard model for conveying emotions, which can be expressed in a variety of ways (e.g. some people express anger by emphasising their voice pitch, while others use body language) [59]. As a result, a one-size-fits-all algorithm may struggle to capture these nuances, as it may fail to recognise the context in which the dialogue occurs.

3.2.6 Challenge 6: Affective (emotive) dialogue and speech production

A coherent, consistent, and interesting dialogue requires several key components: emotion awareness and expressiveness, personalisation, and knowledge [60]. Emotion recognition components extract emotions from utterances using either annotated dialogue corpora or external emotion classifiers to create an end-to-end dialogue system. Each of these components poses a challenge to non-English languages, particularly those with limited resources.

Engaging content is generated when the system's responses are personalised based on the patient's history and personality. By fusing medical knowledge with the patient's cultural and social context, it is possible to generate engaging and pertinent dialogues. It is critical that we combine them all to optimise the performance of the robot and the overall care-receiver experience. To accomplish this, we will train the robots in the relevant medical data/discourses. This will necessitate a number of experiments. Through the incorporation of reinforcement learning, the dialogue system will be able to adjust its response and learn from previous interactions. To achieve the desired outcome, the robot's dialogue generation must be aligned with the appropriate emotion. The current state-of-the-art enables this expressive speech synthesis [61,62]. The main challenges will be selecting the appropriate emotion automatically based on the spoken text [63] and adapting the dialogue to the patient's language and context.

3.2.7 Challenge 7: Non-verbal interaction

Other than verbal interaction, non-verbal interaction is very important for robots to understand. This requires components for interpreting the social environment: reading emotion from facial expression and body posture, the interpretation of hand and arm gestures, the interpretation of intent, and the assessment of proxemics and social space. In addition, components to express non-verbal behaviour will be required, such as non-linguistic utterances, motion, and space to interact with the social environment.

Next to the purely technical challenges of creating a powerful SLAM and tracking system for keeping track of walking humans (see Challenge 2), spatial, social interactions [64], and, in particular, social aspects of navigation around humans need to be investigated [65–69]. Such interactions include proxemics [50], avoiding, giving way, approaching, guiding, and following, among others. Although considerable research has been published in this area [70], deploying such capabilities robustly in a real-world context, such as a crowded hospital waiting room environment, remains a significant research and engineering challenge.

Here the following questions need to be addressed: How close does a robot need to stay behind a person to follow someone without giving the feeling of tailgating or getting lost behind? How far can a robot move ahead when guiding someone? How can we signal an approaching or crossing person that we noticed him or her? Given the healthcare settings, particular challenges include ensuring that spatial, social interaction caters to the diverse range of abilities and needs of the target populations.

3.2.8 Challenge 8: Hardware requirements

In terms of hardware requirements, on the one hand, the sensors on the robots must meet the algorithms' specifications. For instance, microphones must be sensitive enough, and cameras must have a high enough resolution and repetition rate. Additionally, all sensors must be calibrated on-site. On the other hand, the robot's computing power must be sufficient to execute real-time algorithms which require low latencies locally. The deployment architecture is also determined by the robots' performance and the local network infrastructure. As a result, algorithms must be shared between the robot, local base stations (Edge), and the cloud. Subsequently, software deployment is contingent on the robot and local infrastructure, increasing development effort and, in many cases, precluding deployment for economic reasons. To enable widespread adoption of robotics, a hardware-independent implementation must be developed.

3.2.9 Challenge 9: Ethical and social considerations

CASIE robots will be designed with trustworthiness and ethics in mind. They must operate within a high-level normative framework that is tailored to the unique challenges of care, communication, and robotic ethics. This framework must be informed by empirical data pertaining to the unique ethical and social challenges associated with robots operating in a healthcare setting in various countries. The framework will also need to evolve in tandem with critical public (e.g. AI HLEG – the EU's high-level expert group on artificial intelligence) and professional governance considerations for designing and deploying social robots (e.g. IEEE).

3.3 Proposed robotic focused software architecture

The proposed CASIE platform's architecture is depicted in Figure 2, which adapts the established robotics control loop concept – sense, think, act – to a design focused on

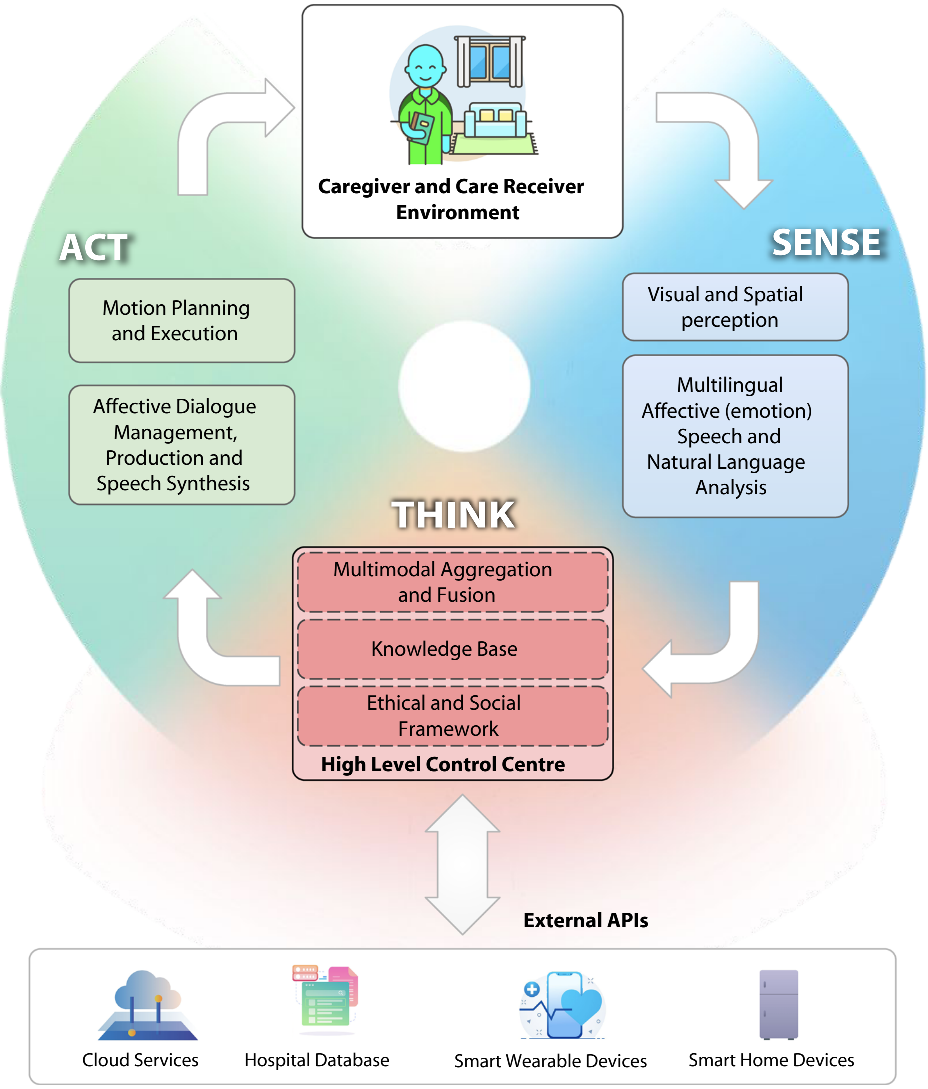
Caregiver and Care Receiver Environment
ACT
SENSE
Motion Planning and Execution
Affective Dialogue Management, Production and Speech Synthesis
Visual and Spatial perception
Multilingual Affective (emotion) Speech and Natural Language Analysis
THINK
Multimodal Aggregation and Fusion
Knowledge Base
Ethical and Social Framework
High Level Control Centre
External APIs
Cloud Services
Hospital Database
Smart Wearable Devices
Smart Home Devices

Figure 2: High-level overview of the planned CASIE architecture.

sHRI with emotion processing capabilities. CASIE robots are designed to process audio input to analyse speech and tone, as well as video streams to detect faces and emotions in order to understand their environment. Unlike the conventional approach of a simple control loop, the core idea is to use this input data not only as a basis for CASIE's decision-making components but also to build up a knowledge base, enabling the robot to remember faces, conversation topics, and even context from its environment while utilising remotely stored knowledge. A CASIE robot would be capable of detecting human emotions as well as locating a missing object (in its knowledge base) in the environment, such as a lost set of keys. Finally, CASIE must carry out the originally planned actions, which may include a combination of screen and speech output in order to carry out physical motions. Each component is described in the following section.

3.3.1 CASIE components

Multilingual affective (emotion) speech and natural language analysis – This functionality is required to process spoken input (to determine the emotion and source language) and to generate text from speech (see Challenge 6).

First, the source language must be identified, followed by the emotion elicited by the dialogue in accordance with the dialogue's intention. This technique can be based on Deep Learning with a subsample of the audio analysed quickly. Deep Learning techniques can also be used to analyse the speaker's prosody (tone of voice) and emotion.

The first step towards decoding a user's speech and interpreting their intent, emotion, sentiment polarity, and expectations is speech-to-text conversion. This functionality can be implemented using a hybrid knowledge-based/deep learning (Long Short-Term Memory, Artificial Recurrent Neural Network [71] NLP pipeline, such as the open-source platform developed [72] in EU H2020 project "*SSIX*": https://cordis.europa.eu/project/id/645425). To classify emotions, we could modify the SSIX aspect-based sentiment pipeline for short texts [72]. This involves pre-processing linguistic data, including tokenisation, sentence splitting, lemmatisation, part-of-speech tagging, and syntactic parsing, followed by Named Entity Recognition and Classification. A similar approach resulted in developing a multi-label maximum entropy social emotion classification model, which uses social emotion lexicons to identify entities and behaviours that elicit various social emotions [73]. Additionally, a pSenti lexicon and learning-based hybrid approach developed for concept-level sentiment analysis could be applied to emotion analysis [74]. The National Research Council's (NRC) Word-Emotion Association Lexicon (EmoLex) [75] resource could be used to add support to over 100 languages [76].

Visual and spatial perception – This functionality, which is implemented as a module in the CASIE architecture, functions as a complement to the language processing functionality. While the latter focuses on speech processing, this module focuses on Computer Vision and other sensor readings that can be interpreted spatially, such as camera images, but may also include data from ultrasonic sensors or joint positions (see Challenge 2).

This module comprises a number of parallel pipelines in the proposed CASIE architecture, including those for face recognition, pose and body language recognition, object recognition, localisation, and mapping. Depending on the output, the data may be processed by the decision-making components or may be directly stored in the local knowledge base (e.g. changes to the map or updated locations of objects).

Multimodal aggregation and fusion – Human communication typically makes use of a variety of verbal and non-verbal cues beyond simple utterances and textual statements, including voice inflection, facial expression, and body language (see Challenge 5). CASIE's dialogue system must aggregate and fuse these disparate data in order to obtain an accurate estimate of the emotions and sentiment polarity conveyed during the user interaction. This functionality is implemented as a component in the architecture by aggregating classifiers trained independently on each modality. Aggregation techniques vary considerably, ranging from simple majority voting to exert rules and ensemble learning.

Additionally, this module will examine more advanced techniques for feature-level fusion that make use of recent advances in deep neural networks for the purpose of learning robust feature representations. While representing multimodal inputs in a common feature space may have the advantage of capturing correlations between different features, an open challenge remains the incorporation of temporal interactions between the various modalities. The component will make use of the SSIX platform's analysis pipelines for classifier aggregation/fusion. The statistical analysis component of SSIX, the "X-Score Engine," provides fine-grained sentiment metrics on analysed textual content via an API. It generates continuous metrics called "X-Scores" that provide insight into a target entity's sentiment behaviour via a custom Named Entity Recognition pipeline. This component will be modified to aggregate emotion scores derived from various classification outputs.

Control, decision making, and planning – It is easy to see how a robot will be required to make complex real-time decisions as part of the various use cases. There is a High-Level Control Centre for this purpose, which comprises three interconnected components that cater to the requirements of diverse use cases (see Challenge 3). The first component is a non-deterministic Finite-State Machine that controls the robots' behaviour in the "here and now" and for a decision that is unlikely to have any long-term impact. The second component is Emotion-Based Decision-Making, which addresses the problems of using emotion and affect states (gleaned from voice, events, and video data) to choose between possibly conflicting actions. The following questions need to be solved within the component implementation process, such as what parameters from emotion states should be used? What emotion patterns should the robot look for? How do we reuse decision mechanisms across scenarios and robot implementations? The third component is the Robotic Planner, which is a probabilistic planner used to plan a series of actions that have the highest probability of reaching a goal set by the robot users. The planner will need to deal with incomplete, partially observable, stochastic information, and uncertain post conditions; all elements inherent to the use of interactive robots in dynamic scenarios.

The three components each handle a different temporal aspect of the robots' control, with the non-deterministic Finite State Machines handling immediate events and the planner handling actions with a long-term horizon.

Knowledge base – The robot's knowledge base represents its long-term memory. In general, its role is to store data that have been classified as relevant by the decision-making component. Moreover, it also acts as an abstraction layer for external data sources and services to provide a structured approach for the planning and decision-making components. It is to be expected that the knowledge base system may receive a high frequency of queries. As such, any queries posed to the system must be executed and responded to quickly and efficiently. A suitable system architecture for storing, querying, and updating the knowledge base would be composed of three components: a Query Processor, a Query Optimiser, and a Query-Cache. Central to a knowledge base is the Knowledge Graph. A query to the knowledge base ultimately requires a traversal of the knowledge graph. The query processor component aims to analyse the input query to the knowledge base and determine a path through the Knowledge Graph, which best satisfies the said query. Graph traversal can be a time-consuming process, and the knowledge base may have to respond to a high frequency of queries. As such, the purpose of the query optimiser is to analyse historical queries and determine what the following query to the system would be to improve response times. Predicted queries with a high probability of being posed to the knowledge base will be stored in the query-cache for immediate retrieval when a matching query is poised. The knowledge base will be exposed to other processes and components using a query API, allowing continued optimisation and upgrading of the knowledge base without interfering with other components and processes. Finally, there is a need for external APIs, and interfaces for external services, such as a hospital's database, are provided.

Affective dialogue management, production, and speech synthesis – This functionality is concerned with the dialogue management and Natural Language Generation (NLG) [77] components of the dialogue system. It is responsible for defining and implementing a Dialogue Manager for: (i) tracking the state of the current dialogue, (ii) updating the knowledge base where appropriate, and (iii) deciding the next dialogue action of the system based on the current state. The dialogue manager may also interface with the planning/behaviour selection component to initiate physical actions when needed. The NLG component will be responsible for translating the given action into natural language. Semantic-based representation learning techniques will be adopted to mitigate problems generated by changing user intents. This task will build on current state-of-the-art technology to modify the text before the Text-to-Speech and increase control over emotional signals (breath, vocal tract length, speed, and tone).

Ethical and social framework – While developing an appropriate ethical and social framework is a contribution in itself, this will also frame and impact the work of the technical challenges. For example, a key ethical and social consideration is the need to minimise the potential for gender bias in the choice of training data, language models, and facial recognition models for the architecture. We see the need for novel computational models and methods that can mitigate bias and make transparent how research deals with gender or other potential forms of bias in language, emotion, and material embodiment (which involve challenges 1, 6, 7, and 9). Combating computational bias is an ongoing challenge for AI systems, as they are only as good as the data we put into them. Inadequate data can contain implicit racial, gender, or ideological biases. Consideration will also be given to the importance, or not, of assigning gender to the robots and the possible impact of that on the overall research goals and outcomes. Robots using CASIE must be transparent and accountable in relation to how they deal with patient needs (in relation to a medical condition, gender, age, and language), in different caring contexts (nursing home, hospital, and private home) and social densities (individuals, small groups, larger groups). Local attitudes to robots in care contexts and the acceptability of robot autonomy will need to be accounted for; our approach considers the local barriers to robot acceptance and the potential positive impacts of social robot communication in different care contexts and situations [78,79].

Motion planning and execution – Challenges 7 and 8 are a collection of software modules responsible for executing non-verbal tasks formulated by the decision-making engine. This could be simple gestures or screen output during a conversation and more complex navigation goals requiring additional data from the robot's knowledge base, like a map of the environment.

3.3.2 CASIE compute architecture

Figure 3 provides an overview of CASIE's compute architecture. Starting from the computational capabilities embedded in the robot, which, depending on the particular type of robot, might be enhanced by CASIE as well, edge and cloud computing layers are used to map the different functions required.

4 Discussion

The CASIE platform sets out an ambitious research challenge to develop an innovative multimodal emotion-aware robotics platform that enables novel applications in healthcare and beyond. In this section, we presented the current state-of-the-art solutions relevant to CASIE areas. For completeness, below are additional robotic

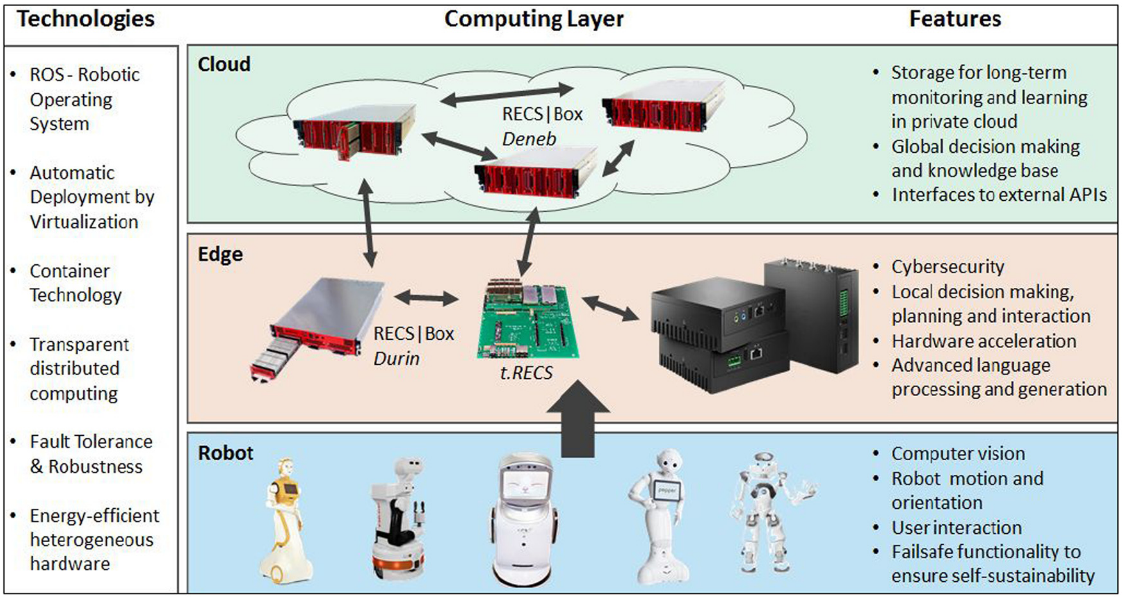
Technologies
Computing Layer
Features
• ROS - Robotic Operating System
• Automatic Deployment by Virtualization
• Container Technology
• Transparent distributed computing
• Fault Tolerance & Robustness
• Energy-efficient heterogeneous hardware
Cloud
RECS | Box Deneb
• Storage for long-term monitoring and learning in private cloud
• Global decision making and knowledge base
• Interfaces to external APIs
Edge
RECS | Box Durin
t.RECS
• Cybersecurity
• Local decision making, planning and interaction
• Hardware acceleration
• Advanced language processing and generation
Robot
• Computer vision
• Robot motion and orientation
• User interaction
• Failsafe functionality to ensure self-sustainability

Figure 3: Overview of CASIE Robot to Cloud computing architecture including major technologies and features.

solutions currently used in relevant healthcare applications which are equipped with varying degrees of social intelligence.

4.1 Other existing robotic solutions

Furhat – It has incredibly alive faces and gestures. It can engage and react to users, while a camera enables it to maintain eye contact. It can interact with humans the way we interact with each other. Merck has trialled it as a pre-screening medical robot to educate people on how to take better care of their health while simultaneously alleviating the embarrassment that people often feel when discussing stigmatised health issues. The trial showed how social robots provide a very intuitive and engaging way to interact with people to raise awareness, pre-screen, and potentially onboard people with high risks of certain medical conditions.

Care-O-Bot – It is a mobile robot assistant which can make simple gestures and express emotions. It was designed to actively support humans in domestic environments.

ElliQ – It is a social robot designed to be a friendly, intelligent, curious presence in older adults' daily lives, helping them, offering tips and advice, responding to questions, surprising them with suggestions. Using real-time sensory data, ElliQ understands situational context to proactively engage with users over the course of the day at the ideal moment, offering personalised suggestions that anticipate their needs and preferences.

Moxi – It is a socially intelligent hospital robot assistant that helps clinical staff with non-patient-facing tasks. Created with a face to visually communicate social cues and able to show its intention before moving to the next task, Moxi is built to foster trust between patients and caregivers.

Buddy Pro – It is an emotional companion robot that can hear, speak, see, and make head movements. It is built on an integrated end-to-end robotics framework and platform for robotics manufacturers and integrators to enable the delivery of highly relevant and customised service robots across several domains.

Sophia – It is a human-like robot endowed with a vibrant personality and holistic cognitive AI. Sophia can engage emotionally and deeply with people. It can maintain eye contact, recognise faces, understand speech, hold natural conversations, and learn and develop through experience. Sophia was designed to show deep engagement and report a warm, to create a real emotional connection.

4.2 Patents for emotion-aware technologies

Next, we examine relevant emotion-aware patents utilising text, audio, and video analysis techniques that are intended to be used in a social robot architecture:

"*Adapting robot behavior based upon human–robot interaction*" (D. A. Florencio, D. Guimarães, D. Bohus, U.S. Patent No. 9956687, 2018) – Microsoft wants to make social robots that adapt to human behaviour. Technologies pertaining to HRI, a task that is desirably performed by the robot, are to cause the human to engage with the robot. The model is updated while the robot is online, such that the behaviour of the robot adapts over time to increase the likelihood that the robot will successfully complete the task. The technology marks a move towards more dynamic human–computer interactions, signifying the increasing sophistication of intelligent devices.

"*Object control system and object control method*" (S. Honda, A. Ohba, H. Segawa, Japan Patent No. WO2018-203501A1, 2018) – Sony has designed a "feeling deduction unit," a robot that can understand a user's mood and respond appropriately. By analysing a feed of data from a camera and sensors, the robot would notice the user's verbal, paralinguistic (e.g. speed, volume, and tone of voice), non-verbal cues, and the user's sweat and heart rates. The system would categorise these inputs based on an emotion index, such as joy, anger, love, and surprise. The robot would then respond in real-time through speech and gestures, for example, by throwing its arms up in celebration. If the robot observes that the user is living an irregular life, such as if the user is staying up late at night to play video games, it may prompt users by saying, "let's go to bed soon." This sets up the robot to have a more deeply integrated position in users' lives, beyond turning on the TV.

"*Human emotion assessment reporting technology system and method*" (R. Thirumalainambi, S. Ranjan, U.S. Patent No. 9141604, 2015) – A novel method of analysing and presenting results of human emotion during a conversational session, such as chat, video, audio, and combination thereof in real-time. The analysis is done using semiotic analysis and hierarchical slope clustering to give feedback for the session or historical sessions to the user or any professional. The method is useful for identifying reactions for a specific session or detecting abnormal behaviour and emotion dynamics. The unique algorithm is useful in getting instant feedback to help maintain or in the session or indicate a need for a change in strategy for a desired result during the session.

"*Emotion state prediction method and robot*" (M. Dong, U.S. Patent No. 2019038506, 2015) – It provides a method for a robot to continually predict the emotional status of a user. The method determines a user's initial emotion status, then predicts a second emotion status based on the first emotion status and a first emotion prediction model, where the second emotion status is the emotion status of the first user at the second moment, and the second moment is later than the first moment; and finally, based on the second emotion status, the system outputs a response to the user. According to the method, the emotion status prediction model can provide a timely warning or a communication skill suggestion for a robot or application, thereby further improving the conversation effect and enhancing user experience.

4.3 Current products and solutions in emotion-aware technologies

Emotion-aware technologies utilising text, audio, and video analysis techniques for specific tasks in the healthcare domain are already on the market, with the following being some emerging market solutions that relate to CASIE.

Winterlight labs – It quantifies speech and language patterns to help detect and monitor cognitive and mental diseases.

Ellipsis health – It provides natural speech analysis as a behavioural health vital sign used to measure anxiety and depression. Their system only requires a few minutes of natural speech to create a real-time assessment.

Eyeris – It offers a suite of face analytics, body tracking, action recognition, and activity prediction APIs. Eyeris technology is currently being used in automotive and social robotics commercial applications.

Clarigent health – It detects mental health issues early, with the goal of preventing suicide in at-risk children and adolescents. The technology is based on linguistics, including word selection and sentence construction. Their system can identify vocal biomarkers in at-risk youth and discovered a correlation with the use of absolutist words and certain pronouns, as well as the pace, breathiness, and inflection of speech.

OliverAPI – It is a speech emotion API that offers a variety of emotional and behavioural metrics. It allows both real-time and batch audio processing and can readily support heavy-duty applications.

DeepAffex – It is a cloud-based affective intelligence platform that utilises innovative facial blood-flow imaging technology to provide analysis of human physiology and psychological affect.

MATRIX Coding System [80] – It is an NLP content analysis system for psychotherapy sessions that transforms session transcripts into code. It offers therapists a direct observation of ongoing psychotherapy processes where analytics are used to tailor psychotherapy treatments.

Moxie – It is a social robot platform that enables children to engage through natural interaction, evoking trust, empathy, motivation, and deeper engagement to promote developmental skills. Moxie can perceive, process, and respond to natural conversation, eye contact, facial expressions, and other behaviour, and recognise and recall people, places, and things to create a unique and personalised learning experience for a child.

Ellie [81] – It is a virtual interviewer who can detect non-verbal cues and respond accordingly. The system analyses the patient's face and speech pattern before answering questions. Ellie's actions, movements, and speech mimic those of a real therapist, but not entirely, which is advantageous for patients who are fearful of therapy.

4.4 Future research potential and conclusions

From a robotics perspective, allowing robots to operate in social contexts and react to their human users' social and emotional circumstances is a significant and crucial step towards expanding their use and applicability in our society. While providing them with basic social intelligence may seem trivial from a human perspective, this represents a significant advancement that must be built upon from a machine perspective. The proposed social intelligence functionality from the initial design is hard-coded and mimetic, devised to have autonomy and learning capabilities. From a technological standpoint, we conclude that robotic social intelligence components' autonomy, decision-making, and learning capabilities are the most critical aspects that could be significantly enhanced and extended in future research. This is because decision-making is still in its infancy, as robotic functionality and complete robotic autonomy necessitate continuous heavy decision-making processes. Then learning capabilities will continuously enhance the robotic adaptability and will have to go hand in hand with improved computer hardware performance due to ever-increasing computing demands from more complex scenarios. In this regard, the scope for future advancement and possibilities resulting from the successful implementation of our proposed approach for emotion recognition and analysis technology into robots is endless, both in the AI and IT hardware domain. Applications range from social robotics, healthcare, entertainment, industry in general and manufacturing in particular, ecology, space exploration, tourism, and education, which can impact several markets and facets of our daily and future life.

At the same time, it can boost and open new possibilities and opportunities for further research and innovation. Some of these include integrating "CASIE" technologies with AI stand-alone technologies for entirely new types of applications or product enhancements. The AI market could use this technology to have much better data to improve their analysis and decisions and provide more valuable and human-like responses and use CASIE for very different AI-type applications, especially in the healthcare application area. CASIE can also boost service and social robots and, as a result, their future applications and innovations or other existing technologies like smart spaces or other autonomous vehicles.

This article presents a detailed approach and an in-depth attempt to enhance robotic interactions with social intelligence, relying heavily on new AI and IT technologies. The above material will be further developed and structured as an EU proposal submission for the Horizon Europe program in 2021. The authors also hold various elements of the required CASIE technology to continue their independent development and enhancements with novel functionalities.

Funding information: This work was supported, in part, by Science Foundation Ireland Grant 13/RC/2094 and co-funded under the European Regional Development Fund through the Southern and Eastern Regional Operational Programme to Lero – the Science Foundation Ireland Research Centre for Software (http://www.lero.ie), the Insight Centre for Data Analytics which is funded by Science Foundation Ireland (SFI/12/RC/2289 P2), and the ADAPT Centre for Digital Content Technology (https://www.adaptcentre.ie) which is funded under the Science Foundation Ireland Research Centres Programme (Grant 13/RC/2106).

Conflict of interest: Authors state no conflict of interest.

Data availability statement: Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

References

[1] C. Murad and C. Munteanu, "I don't know what you're talking about, HALexa," in *Proceedings of the 1st International Conference on Conversational User Interfaces – CUI '19*, ACM Press, 2019, DOI: https://doi.org/10.1145/3342775.3342795.

[2] G. Hoffman, "Anki, Jibo, and Kuri: What we can learn from social robots that didn't make it," *IEEE Spectrum*, 2019, http://spectrum.ieee.org/automaton/robotics/home-robots/anki-jibo-and-kuri-what-we-can-learn-from-social-robotics-failures.

[3] L. Tian and S. Oviatt, “A taxonomy of social errors in human–robot interaction,” *ACM Trans. Hum.-Robot Interact.* vol. 10, no. 2, art. 13, 2021, DOI: https://doi.org/10.1145/3439720.

[4] R. Plutchik, “A General Psychoevolutionary Theory of Emotion,” in *Theories of Emotion*, R. Plutchik and H. Kellerman, Eds., Academic Press, New York, 1980, pp. 3–33, DOI: https://doi.org/10.1016/B978-0-12-558701-3.50007-7.

[5] M. A. Goodrich and A. C. Schultz, “Human-robot interaction: A survey,” *Foundation Trends® Hum.-Comp Interact.*, vol. 1, no. 3, pp. 203–275, 2008, DOI: https://doi.org/10.1561/1100000005.

[6] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, “Would you trust a (faulty) robot?” in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human–Robot Interaction*, ACM, 2015, pp. 141–148, DOI: https://doi.org/10.1145/2696454.2696497.

[7] K. Darling, “‘Who’s Johnny?’ Anthropomorphic framing in human-robot interaction, integration, and policy,” in *ROBOT ETHICS* 2.0, P. Lin, G. Bekey, K. Abney, R. Jenkins, Eds., Oxford University Press, 2017. Available at SSRN: http://dx.doi.org/10.2139/ssrn.2588669.

[8] O. Celiktutan, E. Sariyanidi, and H. Gunes, “Computational Analysis of Affect, Personality, and Engagement in Human–Robot Interactions,” in *Computer Vision for Assistive Healthcare*, M. Leo and G. M. Farinella, Eds., Academic Press, United Kingdom, 2018, pp. 283–318, DOI: https://doi.org/10.1016/B978-0-12-813445-0.00010-1.

[9] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, Prentice Hall: Pearson Education International, 2014.

[10] P. Wiriyathammabhum, D. Summers-Stay, C. Fermüller, and Y. Aloimonos, “Computer vision and natural language processing,” *ACM Computing Surveys*, vol. 49, no. 4, art. 71, 2017, DOI: https://doi.org/10.1145/3009906.

[11] D. Nyga and M. Beetz, “Cloud-based probabilistic knowledge services for instruction interpretation,” in *Robotics Research, Springer Proceedings in Advanced Robotics, vol 3*, A. Bicchi and W. Burgard, Eds., Springer, Cham, 2018, pp. 649–664, DOI: https://doi.org/10.1007/978-3-319-60916-4_37.

[12] E. Triantaphyllou, B. Shu, S. N. Sanchez, and T. Ray, “Multi-criteria decision making: an operations research approach,” *Encyclopedia Electrical Electron. Eng.*, vol. 15, no. 1998, pp. 175–186, 1998.

[13] European Commission Directorate General for Health and Food Safety, “State of health in the EU: Companion Report,” 2019, https://data.europa.eu/doi/10.2875/71887.

[14] M. Kyrarini, F. Lygerakis, A. Rajavenkatanarayanan, C. Sevastopoulos, H. R. Nambiappan, K. K. Chaitanya, et al., “A survey of robots in healthcare,” *Technologies*, vol. 9, no. 1, art. 8, 2021, DOI: https://doi.org/10.3390/technologies9010008.

[15] T. Belpaeme, P. Baxter, R. Read, R. Wood, H. Cuayahuitl, B. Kiefer, et al., “Multimodal child-robot interaction: Building social bonds,” *J. Hum.-Robot Interact.*, vol. 1, no. 2, pp. 33–53, 2013, DOI: https://doi.org/10.5898/jhri.1.2.belpaeme.

[16] A. D. Nuovo, F. Broz, N. Wang, T. Belpaeme, A. Cangelosi, R. Jones, et al., “The multi-modal interface of robot-era multi-robot services tailored for the elderly,” *Intell. Serv. Robot.*, vol. 11, no. 1, pp. 109–126, 2017, DOI: https://doi.org/10.1007/s11370-017-0237-6.

[17] R. Bemelmans, G. J. Gelderblom, P. Jonker, and L. de Witte, “Socially assistive robots in elderly care: A systematic review into effects and effectiveness,” *J. Am. Med. Directors Assoc.*, vol. 13, no. 2, pp. 114–120.e1, 2012, DOI: https://doi.org/10.1016/j.jamda.2010.10.002.

[18] R. Q. Stafford, E. Broadbent, C. Jayawardena, U. Unger, I. H. Kuo, A. Igic, et al., “Improved robot attitudes and emotions at a retirement home after meeting a robot,” in *19th International Symposium in Robot and Human Interactive Communication, IEEE*, 2010, pp. 82–87, DOI: https://doi.org/10.1109/roman.2010.5598679.

[19] J. J. Diehl, L. M. Schmitt, M. Villano, and C. R. Crowell, “The clinical use of robots for individuals with autism spectrum disorders: A critical review,” *Res. Autism Spect. Dis.*, vol. 6, no. 1, pp. 249–262, 2012, DOI: https://doi.org/10.1016/j.rasd.2011.05.006.

[20] B. Scassellati, “How social robots will help us to diagnose, treat, and understand autism,” in *Robotics Research, Springer Tracts in Advanced Robotics, vol. 28*, S. Thrun, R. Brooks, H. Durrant-Whyte, Eds., Springer, Berlin, Heidelberg, 2007, pp. 552–563, DOI: https://doi.org/10.1007/978-3-540-48113-3_47.

[21] S. Thill, C. A. Pop, T. Belpaeme, T. Ziemke, and B. Vanderborght, “Robot-assisted therapy for autism spectrum disorders with (partially) autonomous control: Challenges and outlook,” *Paladyn, J. Behav. Robot.*, vol. 3, no. 4, pp. 209–217, 2012, DOI: https://doi.org/10.2478/s13230-013-0107-7.

[22] T. Turja and A. Oksanen, “Robot acceptance at work: A multilevel analysis based on 27 EU countries,” *Int. J. Soc. Robot.*, vol. 11, no. 4, pp. 679–689, 2019, DOI: https://doi.org/10.1007/s12369-019-00526-x.

[23] L. Hung, C. Liu, E. Woldum, A. Au-Yeung, A. Berndt, C. Wallsworth, et al., “The benefits of and barriers to using a social robot PARO in care settings: a scoping review,” *BMC Geriatrics*, vol. 19, art. 232, 2019, DOI: https://doi.org/10.1186/s12877-019-1244-6.

[24] P. Salovey and J. D. Mayer, “Emotional intelligence,” *Imaginat. Cognit. Personal.*, vol. 9, no. 3, pp. 185–211, 1990, DOI: https://doi.org/10.2190/dugg-p24e-52wk-6cdg.

[25] M. Asada, “Towards artificial empathy,” *Int. J. Soc. Robot.*, vol. 7, pp. 19–33, 2015, DOI: https://doi.org/10.1007/s12369-014-0253-z.

[26] M. Chita-Tegmark, J. M. Ackerman, and M. Scheutz, “Effects of assistive robot behavior on impressions of patient psychological attributes: Vignette-based human–robot interaction study,” *J. Med. Internet Res.*, vol. 21, no. 6, art. e13729, 2019, DOI: https://doi.org/10.2196/13729.

[27] M. Escher, T. V. Perneger, and J.-C. Chevrolet, “National questionnaire survey on what influences doctors’ decisions about admission to intensive care,” *BMJ*, vol. 329, no. 7463, art. 425, 2004, DOI: https://doi.org/10.1136/bmj.329.7463.425.

[28] G. Odekerken-Schröder, C. Mele, T. Russo-Spena, D. Mahr, and A. Ruggiero, “Mitigating loneliness with companion robots in the COVID-19 pandemic and beyond: An integrative framework and research agenda,” *J. Serv. Manag.*, vol. 31, no. 6, pp. 1149–1162, 2020, DOI: https://doi.org/10.1108/josm-05-2020-0148.

[29] G. D’Onofrio, D. Sancarlo, M. Raciti, M. Burke, A. Teare, T. Kovacic, et al., “MARIO Project: validation and evidence of service robots for older people with dementia,” *J. Alzheimer*

Dis., vol. 68, no. 4, pp. 1587–1601, 2019, DOI: https://doi.org/10.3233/JAD-181165.

[30] Microsoft, "Artificial Intelligence in Western Europe: How 277 major European companies benefit from AI," 2019, https://info.microsoft.com/WE-DIGTRNS-CNTNT-FY19-10Oct-09-ArtificialIntelligenceinWesternEurope-MGC0003181_01Registration-ForminBody.html.

[31] G.-Z. Yang, B. J. Nelson, R. R. Murphy, H. Choset, H. Christensen, S. H. Collins, et al., "Combating COVID-19 – The role of robotics in managing public health and infectious diseases," *Sci. Robot.*, vol. 5, no. 40, art. eabb5589, 2020, DOI: https://doi.org/10.1126/scirobotics.abb5589.

[32] World Health Organization, "Global strategy on human resources for health: Workforce 2030," 2020, https://www.who.int/publications/i/item/9789241511131.

[33] J.-P. Michel and F. Ecarnot, "The shortage of skilled workers in Europe: Its impact on geriatric medicine," *Europ. Geriatr. Med.*, vol. 11, no. 3, pp. 345–347, 2020, DOI: https://doi.org/10.1007/s41999-020-00323-0.

[34] J. Wajcman, "Feminist theories of technology," *Cambridge J. Econom.*, vol. 34, no. 1, pp. 143–152, 2010, DOI: https://doi.org/10.1093/cje/ben057.

[35] C. Tannenbaum, R. P. Ellis, F. Eyssel, J. Zou, and L. Schiebinger, "Sex and gender analysis improves science and engineering," *Nature*, vol. 575, no. 7781, pp. 137–146, 2019, DOI: https://doi.org/10.1038/s41586-019-1657-6.

[36] M. Bergin, J. S. Wells, and S. Owen, "Gender awareness, symptom expressions and Irish mental health-care provision," *J. Gender Stud.*, vol. 25, no. 2, pp. 141–154, 2016, DOI: https://doi.org/10.1080/09589236.2014.917950.

[37] T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, et al., "Universals and cultural variation in turn-taking in conversation," *Proc. Nat. Acad. Sci.*, vol. 106, no. 26, pp. 10587–10592, 2009, DOI: https://doi.org/10.1073/pnas.0903616106.

[38] R. W. Picard, "Affective computing: challenges," *Int. J. Hum.-Comp. Stud.*, vol. 59, no. 1–2, pp. 55–64, 2003, DOI: https://doi.org/10.1016/s1071-5819(03)00052-1.

[39] V. Mitra, S. Booker, E. Marchi, D. S. Farrar, U. D. Peitz, B. Cheng, et al., "Leveraging acoustic cues and paralinguistic embeddings to detect expression from voice," in *Proc. Interspeech 2019*, 2019, pp. 1651–1655, DOI: https://doi.org/10.21437/interspeech.2019-2998.

[40] B. Weiss, *Talker Quality in Human and Machine Interaction*, Springer International Publishing, Switzerland, 2020, DOI: https://doi.org/10.1007/978-3-030-22769-2.

[41] D. Maynard, K. Bontcheva, and I. Augenstein, "Natural language processing for the semantic web," *Synthesis Lect. Semant. Web Theor. Technol.*, vol. 6, no. 2, pp. 1–194, 2016, DOI: https://doi.org/10.2200/s00741ed1v01y201611wbe015.

[42] M. Scheutz, R. Cantrell, and P. Schermerhorn, "Toward humanlike task-based dialogue processing for human robot interaction," *AI Magazine*, vol. 32, no. 4, pp. 77–84, 2011, DOI: https://doi.org/10.1609/aimag.v32i4.2381.

[43] P. Fung, D. Bertero, Y. Wan, A. Dey, R. H. Y. Chan, F. B. Siddique, et al., "Towards empathetic human–robot interactions," in *Computational Linguistics and Intelligent Text Processing, CICLing 2016, Lecture Notes in Computer Science, vol. 9624*, A. Gelbukh, Ed., Springer, Cham, 2018, pp. 173–193, DOI: https://doi.org/10.1007/978-3-319-75487-1_14.

[44] Y. Zhang and M. Huang, "Overview of the NTCIR-14 short text generation subtask: emotion generation challenge," in *Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies*, June 10–13, 2019 Tokyo Japan, 2019, pp. 316–327.

[45] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 42, no. 2, pp. 318–327, 2020, DOI: https://doi.org/10.1109/tpami.2018.2858826.

[46] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 43, no. 1, pp. 172–186, 2021, DOI: https://doi.org/10.1109/tpami.2019.2929257.

[47] M. Berger, A. Tagliasacchi, L. M. Seversky, P. Alliez, G. Guennebaud, J. A. Levine, et al., "A survey of surface reconstruction from point clouds," *Comp. Graph. Forum*, vol. 36, no. 1, pp. 301–329, 2016, DOI: https://doi.org/10.1111/cgf.12802.

[48] O. Sorkine and M. Alexa, *"As-rigid-as-possible surface modeling,"* in *Geometry Processing*, A. Belyaev and M. Garland, Eds., 2007, DOI: http://dx.doi.org/10.2312/SGP/SGP07/109-116.

[49] W. Chang and M. Zwicker, "Global registration of dynamic range scans for articulated model reconstruction," *ACM Trans. Graph.*, vol. 30, no. 3, pp. 1–15, 2011.

[50] R. A. Newcombe, D. Fox, and S. M. Seitz, "DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE*, 2015, pp. 343–352, DOI: https://doi.org/10.1109/cvpr.2015.7298631.

[51] M. L. Walters, K. Dautenhahn, R. Te Boekhorst, K. L. Koay, D. S. Syrdal, and C. L. Nehaniv, "An empirical framework for human–robot proxemics," *Procs. of New Frontiers in Human-Robot Interaction: symposium at the AISB09 convention*, 2009, pp. 144–149.

[52] A. Coninx, P. Baxter, E. Oleari, S. Bellini, B. Bierman, O. B. Henkemans, et al., "Towards long-term social child-robot interaction: Using multi-activity switching to engage young users," *J. Hum.-Robot Interact.*, vol. 5, no. 1, pp. 32–67, 2016, DOI: https://doi.org/10.5898/jhri.5.1.coninx.

[53] H. Beck, M. Dao-Tran, and T. Eiter, "LARS: A logic-based framework for analytic reasoning over streams," in *SOFSEM 2018: Theory and Practice of Computer Science, Lecture Notes in Computer Science, vol. 10706*, Springer, Cham, A. Tjoa, L. Bellatreche, S. Biffl, J. van Leeuwen, J. Wiedermann, Eds., 2018, pp. 87–93, DOI: https://doi.org/10.1007/978-3-319-73117-9_6.

[54] D. Paulius and Y. Sun, "A survey of knowledge representation in service robotics," *Robot. Autonom. Sys.*, vol. 118, pp. 13–30, 2019, DOI: https://doi.org/10.1016/j.robot.2019.03.005.

[55] J. Wang, N. Ntarmos, and P. Triantafillou, "Indexing query graphs to speedup graph query processing," *OpenProceedings.org*, 2016, pp. 41–52, https://openproceedings.org/2016/conf/edbt/paper-30.pdf.

[56] F. Gessert, W. Wingerath, and N. Ritter, "Polyglot persistence in data management," in *Fast and Scalable Cloud Data Management*, Springer, Cham, 2020, pp. 149–174, DOI: https://doi.org/10.1007/978-3-030-43506-6_7.

[57] H. Rong, T. Ma, M. Tang, and J. Cao, “A novel subgraph K^+-isomorphism method in social network based on graph similarity detection,” *Soft Comput.*, vol. 22, no. 8, pp. 2583–2601, 2017, DOI: https://doi.org/10.1007/s00500-017-2513-y.

[58] H. Gunes and M. Piccardi, “Automatic temporal segment detection and affect recognition from face and body display,” in *IEEE Transactions on Systems, Man, and Cybernetics, Part B* (*Cybernetics*), vol. 39, no. 1, pp. 64–84, 2009, DOI: https://doi.org/10.1109/tsmcb.2008.927269.

[59] L.-P. Morency, R. Mihalcea, and P. Doshi, “Towards multimodal sentiment analysis,” in *ICMI ’11: Proceedings of the 13th International Conference on Multimodal Interfaces*, ACM Press, 2011, pp. 169–176, DOI: https://doi.org/10.1145/2070481.2070509.

[60] Y. Ma, K. L. Nguyen, F. Z. Xing, and E. Cambria, “A survey on empathetic dialogue systems,” *Inform. Fusion*, vol. 64, pp. 50–70, 2020, DOI: https://doi.org/10.1016/j.inffus.2020.06.011.

[61] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, et al., “Towards end-to-end prosody transfer for expressive speech synthesis with tacotron,” in *International Conference on Machine Learning*, PMLR, 2018, pp. 4693–4702.

[62] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, Y. Battenberg, Y. Shor, et al., “Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis,” in *International Conference on Machine Learning*, PMLR, 2018, pp. 5180–5189.

[63] G. Hillaire, F. Iniesto, and B. Rienties, “Humanising text-to-speech through emotional expression in online courses,” *J. Interact. Media Edu.*, vol. 2019, no. 1, p. 12, 2019, DOI: https://doi.org/10.5334/jime.519.

[64] R. Kirby, “Social robot navigation,” *Ph.D. dissertation*, The Robotics Institute, Carnegie Mellon University, 2010.

[65] A. Peters, T. P. Spexard, H. Marc, and W. Petra, “Hey robot, get out of my way – A survey on a spatial and situational movement concept in HRI,” *Ambient Intell. Smart Environ.*, vol. 9, pp. 147–165, 2011, DOI: https://doi.org/10.3233/978-1-60750-731-4-147.

[66] A. Peters, *“Spatial coordination-human and robotic communicative whole-body motions in narrow passages,”* Ph.D. dissertation, Universitat Bielefeld, 2012, https://pub.uni-bielefeld.de/record/2594360.

[67] C. Lichtenthäler, A. Peters, S. Griffiths, and A. Kirsch, “Social navigation – identifying robot navigation patterns in a path crossing scenario,” in *Social Robotics, ICSR 2013, Lecture Notes in Computer Science, vol. 8239*, G. Herrmann, M. J. Pearson, A. Lenz, P. Bremner, A. Spiers, U. Leonards, Eds., Springer, Cham, 2013, pp. 84–93, DOI: https://doi.org/10.1007/978-3-319-02675-6_9.

[68] C. Lichtenthäler, T. Lorenzy, and A. Kirsch, “Influence of legibility on perceived safety in a virtual human–robot path crossing task,” in *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, IEEE, 2012, pp. 676–681, DOI: https://doi.org/10.1109/roman.2012.6343829.

[69] C. Dondrup, C. Lichtenthäler, and M. Hanheide, “Hesitation signals in human-robot head-on encounters: a pilot study,” in *HRI ’14: Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction*, ACM, 2014, pp. 154–155, DOI: https://doi.org/10.1145/2559636.2559817.

[70] K. Charalampous, I. Kostavelis, and A. Gasteratos, “Recent trends in social aware robot navigation: A survey,” *Robot. Autonom. Syst.*, vol. 93, pp. 85–104, 2017, DOI: https://doi.org/10.1016/j.robot.2017.03.002.

[71] Y. Goldberg, “Neural network methods for natural language processing,” *Synthes. Lectur. Hum. Lang. Technol.*, vol. 10, no. 1, pp. 1–309, 2017, DOI: https://doi.org/10.2200/s00762ed1v01y201703hlt037.

[72] B. Davis, K. Cortis, L. Vasiliu, A. Koumpis, R. McDermott, and S. Handschuh, “Social sentiment indices powered by, X-Scores,” in *the Second International Conference on Big Data, Small Data, Linked Data and Open Data – ALLDATA 2016*, 2016, http://www.proceedings.com/29767.html.

[73] T. Gaillat, B. Stearns, G. Sridhar, R. McDermott, M. Zarrouk, and B. Davis, “Implicit and explicit aspect extraction in financial microblogs,” in *Proceedings of the First Workshop on Economics and Natural Language Processing*, ACL, 2018, pp. 55–61, DOI: https://doi.org/10.18653/v1/w18-3108.

[74] J. Li, Y. Rao, F. Jin, H. Chen, and X. Xiang, “Multi-label maximum entropy model for social emotion classification over short text,” *Neurocomputing*, vol. 210, pp. 247–256, 2016, DOI: https://doi.org/10.1016/j.neucom.2016.03.088.

[75] A. Mudinas, D. Zhang, and M. Levene, “Combining lexicon and learning based approaches for concept-level sentiment analysis,” in *WISDOM ’12: Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, ACM Press, 2012, art. 5, pp. 1–8, DOI: https://doi.org/10.1145/2346676.2346681.

[76] S. Poria, A. Gelbukh, E. Cambria, A. Hussain, and G.-B. Huang, “EmoSenticSpace: A novel framework for affective commonsense reasoning,” *Knowledge-Based Syst.*, vol. 69, pp. 108–123, 2014, DOI: https://doi.org/10.1016/j.knosys.2014.06.011.

[77] A. Gatt and E. Krahmer, “Survey of the state of the art in natural language generation: Core tasks, applications and evaluation,” *J. Artif. Intell. Res.*, vol. 61, pp. 65–170, 2018.

[78] M. Coeckelbergh, “Personal robots, appearance, and human good: A methodological reflection on roboethics,” *Int. J. Soc. Robot.*, vol. 1, no. 3, pp. 217–221, 2009, DOI: https://doi.org/10.1007/s12369-009-0026-2.

[79] S. Vallor, “Carebots and caregivers: Sustaining the ethical ideal of care in the twenty-first century,” *Philos. Technol.*, vol. 24, art. 251, 2011, DOI: https://doi.org/10.1007/s13347-011-0015-x.

[80] M. Bar, A. Saad, D. Slonim-Atzil, R. Tuval-Mashiach, T. Gour, N. Baron, et al., “Patient – therapist congruent exchanges engaged with the potential-to-experience is associated with better outcome of psychotherapy,” *Psychol. Psychother. Theory Res. Pract.*, vol. 94, no. S2, pp. 304–320, 2020, DOI: https://doi.org/10.1111/papt.12274.

[81] K. Chlasta, K. Wołk, and I. Krejtz, “Automated speech-based screening of depression using deep convolutional neural networks,” *Proc. Comp. Sci.*, vol. 164, pp. 618–628, 2019, DOI: https://doi.org/10.1016/j.procs.2019.12.228.