Journal of Human Centered Technology Vol. 1: No. 1 (2022) 23 – 33
==========
Journal of Human Centered Technology
==========
https://humentech.utm.my

Conceptual design of a socially intelligent agent with triadic empathy and theory of mind for mental health support

Azizi Ab Aziz[1*], Mohamad Farif Jemili[2]

[1,2] Relational Machines Group, Human-Centred Computing Research Lab, School of Computing, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia

*Corresponding Author azizіaziz@uum.edu.my

Received 9 November 2021; Accepted 16 January 2022; Available online 31 January 2022
https://doi.org/10.11113/humentech/v1n1.12

Abstract:
For socially intelligent agents to become a fully digital therapist to support individual with mental health problem in the future, they need to know how to socially interact with humans. One of the key ingredients to allow this skill to take place is an ability to exhibit empathic behaviours. Despite a number of socially intelligent agents were build with empathic behaviour, they only cover single empathy behaviour, contrary to more complex triadic empathic behaviours. In this article, the conceptual design and model to implement triadic empathy in socially intelligent agents is presented.

Keywords: Empathic robot; Social presence, Cognitive modelling, Therapeutic robot, Human-robot interaction.

1. Introduction

Empathy is essential in human-human interaction because it motivates prosocial behaviour and promotes kind acts such as helping, caring, and trust. This concept also relates to understanding or feeling what another person is experiencing within their frame of reference. Within mental health domains, empathy is vital in mental health care and has improved mindsets toward members of stigmatised groups. Furthermore, empathy enables the human therapist to form a therapeutic alliance by acknowledging the client's viewpoint and objectives, their distinctive personality needs and personality, and communicating effectively with them. As it shows positive impacts towards human-human mental health support, it is expected that similar results for technology-based support, especially with socially intelligent agent platforms. Socially intelligent agents (SIAs) are computational artefacts (agents) that express emotion and collaborate with humans. They can serve as learning companions, service robots, and therapists, among other things. Besides, SIAs provide some powerful tools for investigating cognitive mechanisms underlying social intelligence in a human context. To date, SIAs with basic empathy capabilities have been developed to create meaningful relationships with humans as the basis of social cooperation [1] and prosocial behaviours [2-3].

In this article, the triadic empathy model with the Theory of Mind (ToM) was proposed to improve SIA's empathic displays and behaviours for the application of mental health. However, these empathic SIAs are restricted to a single empathic behaviour, whereas humans exhibit at least three types of empathic behaviours (triadic empathy): cognitive [4], affective [5], and companionship [6]. This article is organised as follows. The following section will present essential concepts in socially intelligent agents and their applications in mental health support. After that, the explanation of theoretical concepts and interplays in triadic empathy and ToM will occur, followed by the description of the design overview conceptual models for SIAs with triadic empathy and ToM. Finally, the conclusions and future work in this area conclude the article.

2. Socially Intelligent Agent

SIAs have a recent history of artificial intelligence and robotics. They are viewed as an expanding and increasingly significant research area with an active area of research activities and highly collaborative approaches. Therefore, the underlying background, concepts, and theories in cognitive sciences and human development are crucial to designing algorithms, computational frameworks, and models capable of allowing SIAs technology to be used in several practical applications. Examples of SIAs are social robots, virtual humans in computer games, or conversational agents (interactive chatbots). In general, SIAs can be described as artificial beings that can engage in social interaction with humans, using both verbal (e.g., text, speech) and non-verbal (e.g., facial expressions, gestures) modalities. Such agents convey and recognise emotions; communicate with high-level discourse; learn models of or acknowledge other agents; develop and maintain social relationships; utilise natural cues (e.g., gaze and social gestures); and may learn or develop social capabilities [1,7]. Figure 1 shows some examples of socially intelligent agents.

(a) (b) (c)

Figure 1. Types of socially intelligent agents, (a) chatbot, (b) virtual agent and (c) social robot

Research in SIAs covers many aspects of disciplines, from human development theories to bio-sensing engineering. Because many people will routinely work with SIAs for many hours each day in the future, it is essential to understand how working with SIAs would become more organic. Also, as humans are social creatures, one strategy is comprehending what it means for SIAs and humans to have a social relationship. This idea of SIAs leads to essential requirements to be fulfilled in the future. Among others are; i) a suitable social environment (including humans) [1,7], ii) adequately rich communicative ability – (i.e. a conversational language that enables the fine-grained modelling of others' thoughts [8] and feelings leading to action in that language), 3) general anticipatory modelling abilities [7-8], 4) a capacity to separate various types of perception, including the observation of specific others' actions; one's own actions; and other sensory information [1,9], 5) an ability to comprehend other agents as distinguishable individuals [8], 6) the ability to predict the decisions of others [8,9], 7) the model to recognise one's own decisions [3,7], and 8) the skill to utilise model frameworks learned for one intention for something else [10]. Some of these are prerequisites on an agent's internal architecture, while others are imposed on the societal structure in which it develops.

2.1 Socially Intelligent Agent in Mental Health Support

In general, SIAs are getting popular to reinforce other digital health support by offering healthcare information, evaluation, and therapeutic interventions. Throughout many cases, when compared to telepresence or virtual agents, SIAs elicit more desirable reactions from people, including better ratings on aspects such as general perception, personal choice, participation, supportiveness, attractiveness, and satisfaction. It is crucial to highlight that SIAs can understand and socially interact with individuals while displaying intervention strategies to users similar to web-based and mobile apps (e.g., skills training, health tracking). Due to their diverse capabilities, SIAs may incorporate conventional app- and telehealth-related supports with a socially interactive companion, providing users with a more intriguing and attentive platform. Within specific mental health support perspectives, SIAs can be leveraged in various aspects of applications. These aspects include provide decision and cognitive assessment [1,7], educate patients through multimodal actuation [2], provide companionship and support activities [1,9], provide intervention [8], personalised intentions that can lead to better goal attainment [10], and tailored progress feedback for discrepancy awareness [2,8].

To date, several studies explored the implementation of SIAs in various aspects of self-support (e.g., intervention, training, screening) [11], cognitive behavioural therapy [1], relationship therapy [12], and behavioural change [13]. Not only can SIAs applications lessen the burden of health care providers (e.g., psychotherapists), but they can also decrease healthcare expenditures and assist patients who have trouble connecting with humans in social settings due to factors such as global pandemics, social distancing, location, and time. In this study, a social robot platform will be used as a platform to deliver therapeutic recommendations and actions.

3. Triadic Empathy and Theory of Mind

Human beings are naturally social beings, and our interactions with others foster an empathetic connection that enables social understanding. Both triadic empathy and Theory of Mind (ToM) are firmly ingrained in human evolutionary history and are essential for social functioning.

3.1 Theoretical Concepts on Triadic Empathy with Theory of Mind

First, the dynamics of empathy can be viewed as a triadic interplay, including cognitive empathy, emotional/ affective empathy, and compassionate empathy. Cognitive empathy describes the perception and (accurate) identification of others' feeling states. This type of empathy has been shown to predict positive social outcomes, such as helping others' behaviour and compassion. Second, emotional (affective) empathy, or emotion contagion, describes the emotional mirroring of others' feeling states. This type of empathy helps a person to build emotional connections with others. Lastly, compassionate empathy, or emotional responses of sympathy, caring, and generosity for another is thought to be a prevalent but not guaranteed result of the other two types of empathy [4]. Compassion is frequently conceived of as a distinct prosocial affective response in and of itself, associated with desirable impacts such as benevolent behaviours. These empathy concepts are connected through ToM. ToM is the attribution of mental states to others, enabling individuals to comprehend or anticipate another person's actions and respond accordingly [14]. The advancement in neurosciences (as a result of the development of advanced neuroimaging machines) allows scientists to provide neurological evidence for the interaction between ToM and triadic empathy. Numerous brain areas have been recognised as contributing to ToM and empathy. For example, emotional empathy is related to the insula, amygdala and anterior cingulate cortex (ACC) [5]. These processes are supported by distinct, albeit interconnected brain networks.

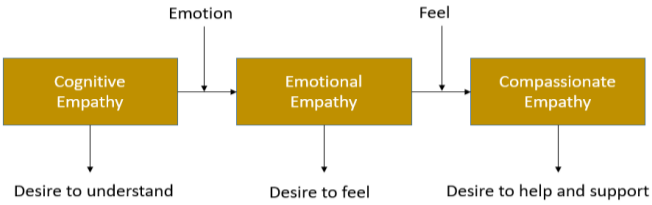

Figure 2. Types of triadic empathy

Whenever a cognitive empathic response is elicited, the ToM brain network (e.g., medial prefrontal cortex, superior temporal sulcus, temporal poles) and the affective ToM network (primarily the ventromedial prefrontal cortex (vPfc)) are generally involved [5,14]. On the other hand, emotional empathic feedback is driven mainly by simulation and involves regions that facilitate emotional responses (i.e., amygdala, insula) [15]. As empathy represents a notion as the capacity to relate to another's emotional state, it is a complex socio-emotional behaviour that requires the interaction of high and low-level cognitive behaviour, such as belief-desire intention concepts. Thus, fundamental empathetic behaviours in open-world situations can probably not be achieved by making a predefined set. Therefore, rather than exist within the predetermined rules/actions, the SIAs will be benefited by allowing more robust action-selection mechanisms based on the "mind-reading" ToM concept.

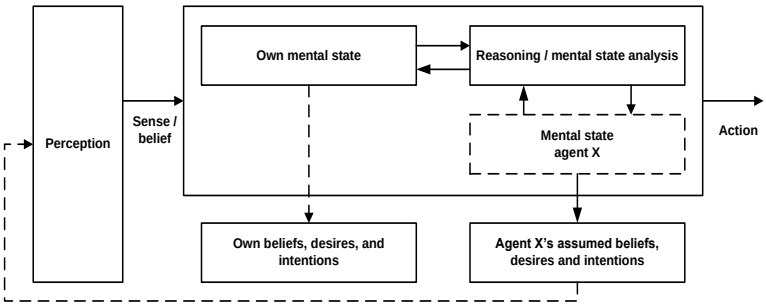

Figure 3. Concepts and components in theory of mind

Essentially, ToM implies that "mindreading," which relates to specific mental states and cause-effect networks of observed individuals, employs the same mental states within the observing socially intelligent agent. As a direct consequence, it is associated with the idea of own model is used to approximate the process of other individuals; a similar simulation standpoint can also be used to describe the emergence of imagination or social "assumption" [16]. This study

will incorporate the Theory of Mind concept as a mechanism to create a belief system model (based on the Belief-Desire-Intention model) for empathic actions execution.

3.2 Limitations on Current Computational Empathy Models for Socially Intelligent Agents

Currently, computational models of empathy for SIAs suffer from a clear description and theoretical positioning as human true empathy perspectives are governed by a triadic (cognitive, emotional, compassionate) concept. To date, there is no clear evidence that an attempt has been made to combine these triadic empathies into a single computational framework. Previous studies of computational empathy models have predominantly concentrated on either one of those empathy constructs. For example, early works conducted by several researchers [1-3,8] only focused on either modelling cognitive or emotional and were somewhat limited only to the virtual agent. Later, more advanced models were developed to address the needs of general SIAs implementation, such as in trust promotion, emotion expression, cognitive testing and evaluation, human-robot interaction and conversational agent (e.g. chatbot). However, these models only addressed a single empathy construct.

Nevertheless, exceptional can be made based on several previous reports [17-19], where those researchers explored the possibilities to connect emotional and cognitive empathy as a dual-route model. However, these models still suffer limited to providing human-like empathy behaviour. Also, those models are not ready for real-world SIAs implementation as the primary purpose of the developer models is only to describe the theoretical aspects of a dual-empathy model from the neuron-biological stand. When it comes to the computational model of triadic empathy, little is known about the existence of its computational model as the triadic empathy requires another concept called Theory of Mind (ToM) to be introduced as a connector to link those three empathy concepts together. As a result of this combined idea, a complete computational model and triadic empathy implementation are missing in the current research on artificial empathy. Without this model, the socially intelligent agent will only work within a limited scope to display emphatic behaviours, thus hampering the idea of fluid social interaction between humans and agents [8,13].

4. Design Overview of Modules for a Socially Intelligent Agent with Empathy and Theory of Mind

A social robot (one of the SIA platforms) with an interactive touch screen display will be chosen for this work's deployment platform. Figure 4 shows the physical design of the robot-based deployment platform. The built robot stands about half a metre tall and is intended to sit on a table or work surface. In addition, there is a small touch-enabled input screen on the front for data entry. Off-the-shelf PC components are used for computation, and a low-cost servo controller controls motor speed. The robot's neck has one degree of freedom, and the paired eyes have two, permitting a full spectrum of horizontal motion (controlled by Tenvis P2P wireless network camera setting). A set of web cameras is positioned above the eyes, which provide the OpenCV face tracking system and emotion recognition modules with a robot view. In addition, this robot features custom LED mouth displays lip-synched to pre-recorded dialogue spoken using available text-to-speech voice.

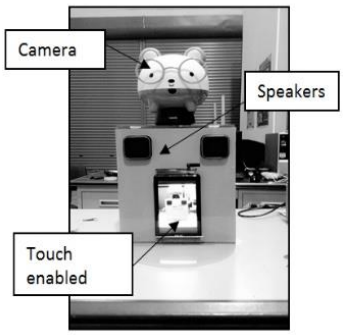

Figure 4. Physical design of a social robot platform

To allow more fluid robot functionality, several modules are designed. Generally, these modules will serve as a basis for implementing a computational framework to enable a socially intelligent agent to support stressed individuals. As depicted in Fig. 5, the conceptual design outcome is divided into five main modules – sensing, emotion analysis, personality and event evaluation, empathy analytics, behaviour selection, stress analytic and support, and feedback.

4.1 Sensing and Feedback Output Modules

In order to interpret and process the meaning of the stimuli that occur in an environment, three types of sensing and perception platforms will be used. These platforms will capture the desired information from stimuli provided by an environment and agent-human interactions. An internal table will provide touch-sensitive screen technology as a direct interface with the intervention-based application. This touch-screen input design makes it easier for the user to select the desired objects. In addition to this, the speech input will be equipped to allow more natural interaction between users and SIAs. This study will utilise the Python Speech Recognition Package (*PyPI*), which offers built-in features (e.g. accessing microphones, processing audio files and natural language processing). Both platforms will be utilised sparingly to allow more fluent human-agent interaction to take place. Moreover, visual sensors are a very popular, practical, and accessible channel for perception (vision-based input) [9]. The vision-based input utilises the high-quality webcam (Logitech HD 720p with 30fps webcam and 1.2 camera megapixel). This webcam provides a 60^{o} diagonal field of view (dFoV) with auto-light correction. The primary function of the visual input is to capture the user's facial expression for the emotion analysis module. A robot must use multimodal feedback after perceiving its user and environment to engage (e.g., display its behaviours and feedback) with its user in a pleasant, comprehensible, appropriate, and intelligent manner. The feedback module encompasses voice output (auditory feedback) and screen output (touch-screen).

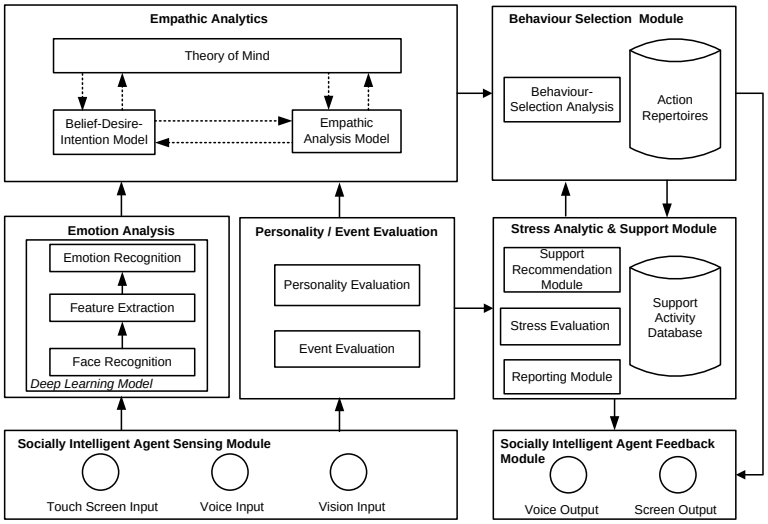

Figure 5. Conceptual design and interplays between modules

4.2 Personality and Event Evaluation

This module will analyse the users' personalities based on the Big Five Personality Traits Theory. This theory describes five broad personality traits: extraversion (also known as extroversion), agreeableness, openness, conscientiousness, and neuroticism (OCEAN). The Big Five Personality Traits have been extensively researched to determine their impact on a person's behaviour and persona, from heredity and environment to age and maturation [20]. The Big Five are uniformly distributed in a normal curve and statistically independent of one another. Also, these traits are genetically heritable and are stable over time. Unlike some other trait theories that categorise people as either introverts or extroverts, the Big Five Model posits that every personality trait exists on a spectrum or spectrum (as shown in Fig. 6). Individuals are thus ranked on a scale between the two extreme ends of the spectrum.

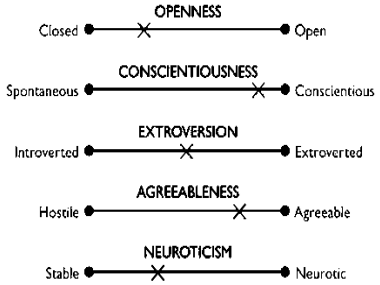

Figure 6. Examples of big five personality traits spectrum

For example, when measuring conscientiousness, one would not be categorised as purely spontaneous or conscientious but rather positioned on a scale indicating their level of conscientiousness. Thus, the individual personality differences can be effectively measured by ranking individuals on each of these traits. However, the maladaptive extremes of the Big Five traits will leave their adverse effects on mental health conditions [21]. Over-conscientiousness, for instance, foretells obsessive-compulsive disorder, whereas low conscientiousness indicates substance abuse as well as other "impulse control disorders". Another central concept is the personality trait of neuroticism, a widely known concept to be associated with depression, distress, anxiety, and bipolar disorders [20]. Therefore, this study will utilise a set of inputs from Big Five Personality Traits as a personality evaluation score in determining the tendency of individual's risk in poor mental health.

4.3 Stress Analytics and Support Module

The Depression-Anxiety-Stress 21 Scale (DASS 21) will be used to assess the user's mental health. The DASS-21 is a combination of three self-report scales developed to measure the negative emotional states of depression, anxiety, and stress. It is a shorter representation of the thorough DASS 42-items. Each of the three DASS-21 scales has seven items, which are categorised into subscales with similar content. First, individuals read each statement and then choose the number ranging from zero (never) to three (almost always), indicating how often the statement adhered to them in the prior week. Scores for depression, anxiety, and stress are computed by adding the scores for each item (Item #1-21) (as depicted in Fig. 7). The details of these items can be found in [22].

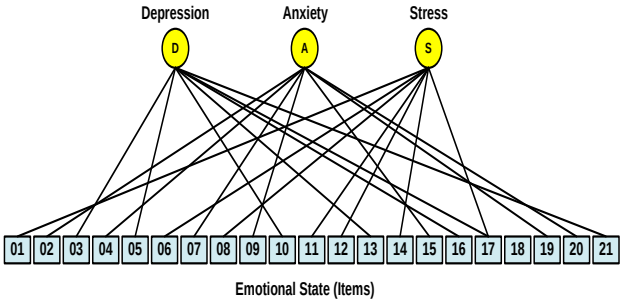

Figure 7. Items and its relations to depression, anxiety and stress

Furthermore, each subscale's scores are classified into five severity levels: normal, mild, moderate, severe, and extremely severe. The severity labels explain the complete spectrum of scores in the population, so 'mild' implies that an individual is higher than the population mean but likely lower than the typical severity of someone seeking support. It is essential to comprehend that this does not imply a minor level of disorder. Numerous studies on its reliability and validity have been published worldwide, demonstrating that the DASS-21 is a well-established instrument that measures depression, anxiety, and stress symptoms in both clinical and non-clinical samples of adults [22].

The event evaluation component relates to the current condition or event that users are experiencing. By combining DASS score evaluation and Big Five Personality Traits results, individuals' risk and condition in poor mental health can be assessed and analysed. Based on this information, a proper intervention or support programme could be formalised and administered to the users. In addition, users will be able to monitor their progress based on personalised reports generated from this module. As there is no "one-size-fits-all" in providing support, the support recommendation model will analyse possible preferences based on specific user personality, mental health status and previously administered aid. The basic incremental learning algorithm (e.g., K-Nearest Neighbours, Case-Based Reasoning) will be used as a classification model to provide related personalised support to the user. The mechanism for selecting the supportive behaviours is adaptive and considers previous interactions with that same user.

4.4 Emotion Analysis

Similar to how practitioners (e.g., a counsellor) engage with patients during therapy sessions, the socially intelligent agent must be capable of interacting with users naturally throughout the session. This step involves identifying a thorough understanding of the user's emotions (e.g., happiness, sadness, fear). In this study, a computer vision model using Convolutional Neural Networks (CNN) Deep Learning recognises users' emotions was developed. The CNN algorithm processes data with a grid pattern inspired by the organisation of the animal visual cortex. Also, CNN has been designed to learn hierarchies of features from low to high-level patterns. In CNN, the convolution (Convolution 3x3 kernel size with Rectified Linear Unit (RLU) activation function) is performed on the input data using a filter to produce a feature map.

The pooling layer (Max Pooling with 2 × 2 pooling window and strides) is introduced to reduce continuous dimensionality, decreasing the feature map size while maintaining important information. Later, the Dropout technique is introduced to randomly selected neurons that are ignored during the training. This technique is crucial to reduce overfitting in this model for well-generalised results. In addition, this model is executed based on Adam (Adaptive Moment Estimation with learning rate α=0.001, regularisation β_1=0.9, β_2=0.998, $\varepsilon=10^{-8}$) optimiser and categorical cross-entropy as the loss function. Figure 8 shows the implemented result of this module.

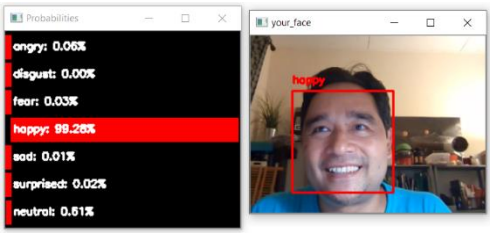

Figure 8. Examples of emotion analysis results

The accuracy is used as an evaluation metric for this model. The OpenCV and Python-Keras libraries are chosen as a computational platform to implement this module for this task. The Haar Cascade file is used to detect the real-time input from the cam. There are seven emotion conditions analysed: angry, disgust, fear, happy, sad, surprise, and neutral. Noted that all face recognition, feature extraction and emotion recognition processes are combined within the CNN Deep Learning model. Our initial experiment results have obtained up to 75% recognition accuracy during the testing process. The output from this emotion analysis model will be used as a perceived input (user's emotional state) as one of the observed (exogenous) factors to trigger empathic behaviours.

4.5 Empathic Analytics

The empathic analysis module is composed of three levels structure. First, the Empathic Analysis Model contains a set of formalised concepts that encapsulates important temporal dynamics and interplays for each process in affective, cognitive and compassionate empathy. These empathy components will be developed based on a hierarchical layer. Later, the belief-desire-intention model (BDI) will be used as a computational framework to display empathic behaviours based on each activated layer. The ToM model will integrate empathic analysis and BDI models as a reference mechanism. The details of this module will be covered in Section 5.

5. Integrating Empathy with Theory of Mind

This section presents the three main models that integrate triadic empathy with ToM. These models are combined under the empathic analytics module.

5.1 Empathy Analysis Model

First, the Empathic Analysis Model contains a set of formalised concepts that encapsulates important temporal dynamics and interplays for each process in affective, cognitive and compassionate empathy. These empathy components will be developed based on a hierarchical layer. The hierarchical layering approach will allow each layer to see the layers below as a "virtual controller" from which it gets percept's and to which it sends commands (Fig. 9(a)).

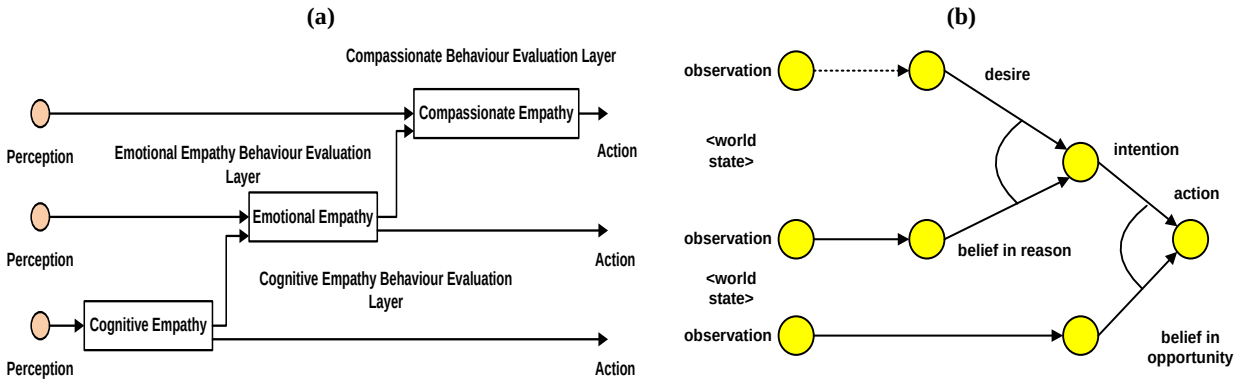

Figure 9. (a) Empathy behavioural layering and (b) belief-desire-intention

Using this approach, the primitive (low-level layer) empathic behaviour (cognitive empathy) will be given priority to be triggered compared to other behaviours (emotional/compassionate empathy). Also, the low-level layer executes much faster and react to those aspects of the world that need to be responded quickly. Likewise, people respond in fractions of a second for simple stimuli but plan at the highest level for complex behaviours in the real world. Therefore, the lower level can be viewed as "fast, and parallel" and the higher level as "slow, deliberate and based on reasoning" [23].

5.2 Belief-Desire-Intention Model

One of the most well-known models of reasoning agents in the sense of agent-based modelling employs beliefs, desires, and intention (BDI) as abstractions to explain a system's behaviours. According to Fig. 9(b), the aspects that distinguish the BDI model are as follows:

- beliefs depict an agent's expectation about the present world state or the likelihood that a given plan of action will contribute to a given world state;
- desires consist of a set of possibly inconsistent choices an agent has about a set of world states; and
- intentions reflect an agent's commitment to a particular action, limiting the consideration of intended outcomes.

A standard BDI interpreter's procedure can be described as a process that begins with an agent reflecting its sensor input and updating its belief base. With this updated belief base, the agent chooses several goals based on a set of desires and commits to achieving them. Later, plans are determined to accomplish that goal by utilising commitment intentions. Eventually, these intentions are carried out in the instantiated plans through meaningful actions (or intentions). This notion can be streamlined as follows (using an agent *X* as an example) [16];

- For any desire β, world state property Ψ, and action α such that an agent *X* beliefs it has a reason for (X, β, Ψ, α) holds: $desire(X, \beta) \wedge belief(X, \Psi) \rightarrow intention(X, \alpha)$
- For any world state property Ψ, and action α such that an agent *X* beliefs it has an opportunity for (X, Ψ, α) holds: $intention(X, \alpha) \wedge belief(X, \Psi) \rightarrow performs(X, \alpha)$

In this model, action is carried out after the subject has the intention to carry it out and believes that such conditions in the world have been met, allowing the subject to carry out the predetermined action. Beliefs are formed as a result of observations. The intention to execute a particular type of action is formed whenever there is a desire and the belief that only certain conditions in the world state exist, making it feasible that implementing this action will meet the above desire [24]. Whether or not a given action is sufficient to satisfy a given desire is determined by the current world state; thus, this belief could be influenced by other world-state beliefs. As a result, the BDI-model integrates such a pattern of reasoning to elucidate behaviour in a more refined form. Unlike a one-step process from a desire to action, an intention is derived as an intermediate phase, and the action is formed from the intention.

5.3 Empathy and Theory of Mind Model

To understand the world around them (e.g. social interaction, activities, and cues), a socially intelligent agent needs to interpret users' actions and "their" own actions in terms of mental states (Theory of Mind / ToM). For example, this interpretation of actions can be viewed as "ϕ beliefs that ρ intends ϕ to persuade α that ψ". There are two prominent views within the computer science and artificial intelligence community to model this concept; *Theory-Theory* (TT) and *Simulation-Theory* (ST) [16,25]. According to the first point of view, the Theory-Theory concept implies that an agent has an innate theory of the structure and functioning of the human mind. It suggests that humans use notions like their own beliefs and intentions to describe and interpret their own and other people's behaviour. Contrary to the Theory-Theory approach, Simulation-Theory was introduced to allow a mental projection of ourselves based on another person's perspective. This concept allows humans to mimic the mental state of another individual. Therefore, instead of relying on a complete theory, the Simulation-Theory approach is more efficient in its simplicity. In this article, the proposed model of triadic empathy will incorporate the Simulation-Theory approach in its design. The design process begins with integrating the BDI model for each empathy model (cognitive, emotional and compassionate). Later, the ToM model will be integrated into perspective-taking and assumed beliefs and intentions of another agent (e.g. agent *X*). It is worth mentioning that the underlying principle of simulation theory is that the agent uses its own reasoning power to explain the user's, so not all of one's own reasoning phases must be included in the theory [14,19]. According to Figure 10, the agent, in addition to representing its own mental state, has representations of mental states about other agents (dotted lines). The agent can start contemplating the mental state of another agent to conclude its behaviour based on its own judgement and perception (from the world/environment) [26]. In other words, it holds true its own reasoner to the mental states referred to it. As a result, the agent rationalises the user's mental constructs as if they had been their own. The agent can use its assertions about the users as input for its own reasoning process (activation level), and its actions can be based on them [16,27]. The activation level of each empathy process can be used to approximate these "assumptions." The numerical activation level of each E_x empathy process can range between [0, E_x^{max}], where E_x^{max} is an integer value

determined empirically (e.g. 1). Even though these processes are always active (based on the circumstances or amount of stimulus), their intensity must surpass a certain threshold before they can be expressed extrinsically (or trigger another layer of empathy level). The activation of each empathy generation process (e.g., *cognitive empathy*) can be computed by the equation:

$$E_x(t) = \frac{1}{1 + e^{-(\Sigma(P_x(t) + B_x(t) + S_x(t)) - \delta t))}} \quad (1)$$

where $P_x(t)$ is the activation level of its affiliated agent's perception from its environment, users, or stimulus; $B_x(t)$ is a collection of beliefs in the agent's mental state. Another concept, $S_x(t)$, adds a persistent level to the active empathy and δ_t is a constant decay term that restores an empathy level to its bias value once the empathy level becomes active. The computed empathy will be accumulated (known as *Accumulated Empathy Drive*, D_x) and updated (either towards upward change or otherwise) throughout time. This contribution pattern can be considered a "delay condition" (regardless of accumulating or decaying contributions). This representation can be described as follows:

$$D_x(t+\Delta t) = D_x(t) + \tau.[(E_x(t)-D_x(t))-\varepsilon](1- D_x(t)).D_x(t).\Delta t \quad (2)$$

assuming both change rate factor, τ ($0 < \tau < 1$) and temporal decay term, $\varepsilon > 0$ and $0 < \Delta t < 1$

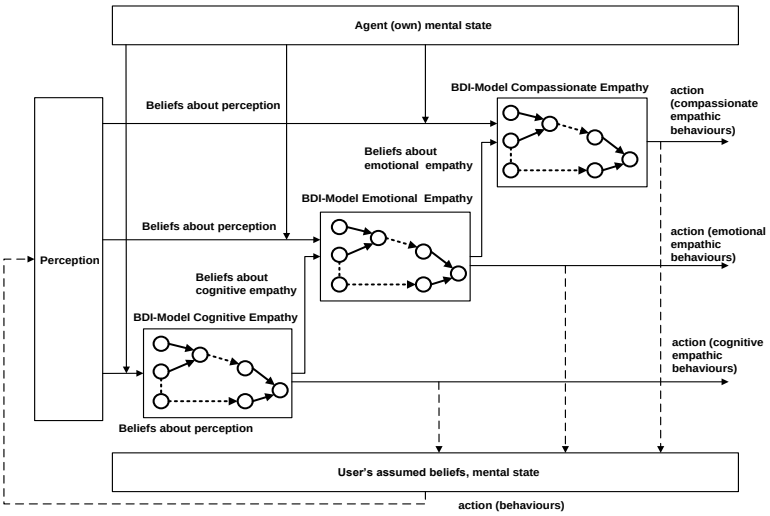

Figure 10. Integrated model of BDI, empathic behavior layer, and ToM

In general, based on accumulated D_x, each E_x is partitioned into three conditions; an under-stimulated, an activation of empathic behaviour, and the activation of the next empathy layer. Figure 11 visualised these conditions.

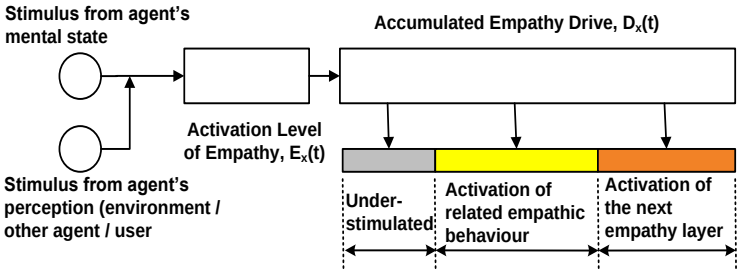

Figure 11. Empathy drive

By default, the accumulated drive remains in its under-stimulated condition until it encounters its high stimulus from the agent's mental state and perception. The activation of related empathic behaviour will occur if the accumulated empathy drive is sufficient enough to allow an agent to exhibit ranges of emphatic gestures towards users. However, if the current drive level is too intense, the next layer of empathy (related to the other *belief-desire-intention* empathy models) will be triggered while showing related empathic gestures. Throughout time, without appropriate stimulus levels (or if the intensity of the stimulus is too low), the temporal decay term will re-establish the current empathy drive level to the baseline condition (under-stimulated).

6. Conclusion

This article explores the ongoing work as a first step towards developing a socially intelligent agent with triadic empathy and ToM. In the context of mental health support, the conceptual design stage has been completed with seven main modules. The interaction between all modules was explained based on generated output from each module. In particular, the detailed explanation of the empathic analytics module was covered with the description of empathy analysis, belief-desire-intention, and the integrated empathy-ToM models. As for the next step, these modules will be used a computational deployment platform on the developed social robot platform. The evaluation process will be based on human experiments (undergraduate students) on stress-induced scenarios for SIAs with empathy and without one. One of the main hypothetical assumptions for this future experiment is that respondents interacting with a triadic empathic SIA will have better therapy outcomes than those who are not.

References

[1] T. Turja, T. Rantanen and A. Oksanen, Robot use self-efficacy in healthcare work (RUSH): development and validation of a new measure, AI & Society, 2019, 34:137–143. https://doi.org/10.1007/s00146-017-0751-2
[2] B. De Carolis, S. Ferilli and G. Palestra, Simulating empathic behaviour in a socially assistive robot, Multimedia Tools and Applications, 2017, 76(4):5073–5094. https://doi.org/10.1007/s11042-016-3797-0
[3] E. Bagheri, P. G. Esteban, H. L Cao, A. De Beir and D. Lefeber, and B. Vanderborght, An autonomous cognitive empathy model responsive to users' facial emotion expressions, ACM Transactions on Interactive Intelligent Systems, 2020, 10(3):1–23. https://doi.org/10.1145/3341198
[4] T. Singer, and C. Lamm, The social neuroscience of empathy, Annals of the New York Academy of Sciences, 2009, 1156:81–96. https://doi.org/10.1111/j.1749-6632.2009.04418.x
[5] A. A. Marsh, The neuroscience of empathy, Current Opinion in Behavioral Sciences, 2018, 19:110-115. https://psycnet.apa.org/doi/10.1016/j.cobeha.2017.12.016.
[6] A. Wykowska, E. Wiese, A. Prosser and H. J. Müller, Beliefs about the minds of others influence how we process sensory information, PloS One, 2014, 9(4): e94339. https://doi.org/10.1371/journal.pone.0094339
[7] S. Varrasi, S. D. Nuovo, D. Conti and A. D. Nuovo, A social robot for cognitive assessment, HRI '18: Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, 2018, 269–270. https://doi.org/10.1145/3173386.3176995
[8] S. Schneider and F. Kummert, Comparing robot and human guided personalisation: Adaptive exercise robots are perceived as more competent and trustworthy, International Journal of Social Robotics, 2020, 13:169–185. https://doi.org/10.1007/s12369-020-00629-w
[9] A. A. Scoglio, E. D. Reilly, J. A. Gorman and C. E. Drebing, Use of social robots in mental health and well-being research: systematic review, Journal of Medical Internet Research, 2019, 21(7):e13322. https://doi.org/10.2196/13322
[10] M. C. Martini, C. A. Gonzalez and E. Wiese, Seeing minds in others – Can agents with robotic appearance have human-like preferences?, PLOS ONE, 2019, 11(2): e0149766. https://doi.org/10.1371/journal.pone.0146310
[11] A. A. Aziz, A. Saad and F. Ahmad, CAKNA: A personalised robot-based platform for anxiety states therapy, citizen-centric smart cities services workshop, The 13th International Conference of Intelligent Environments, 2017, 22:141–150. https://doi.org/10.3233/978-1-61499-796-2-141
[12] F. Dino, R. Zandie, H. Abdollahi, S. Schoeder and M. H. Mahoor, Delivering cognitive behavioral therapy using a conversational social robot, 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019, 2089-2095. https://doi.org/10.1109/IROS40897.2019.8968576
[13] A. Abdulrahman, D. Richards, H. Ranjbartabar and S. Mascarenhas, Belief-based agent explanations to encourage behaviour change, Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents (IVA '19), 2019, 176–178. https://doi.org/10.1145/3308532.3329444
[14] M. Brüne and U. Brüne-Cohrs, Theory of mind—evolution, ontogeny, brain mechanisms and psychopathology, Neuroscience and Biobehavioral Reviews, 2006, 30(4):437–455. https://doi.org/10.1016/j.neubiorev.2005.08.001
[15] A. Abu-Akel and S. Shamay-Tsoory, Neuroanatomical and neurochemical bases of theory of mind, Neuropsychologia, 2011, 49(11):2971–2984. https://doi.org/10.1016/j.neuropsychologia.2011.07.012
[16] D. C. Ong, J. Zaki and N. D. Goodman, Computational models of emotion inference in theory of mind: A review and roadmap. Topics in Cognitive Science, 2019, 11(2):338–357. https://doi.org/10.1111/tops.12371
[17] Z. A. Memon and J. Treur, An agent model for cognitive and affective empathic understanding of other agents, in: N. T. Nguyen, Transactions on Computational Collective Intelligence VI. Lecture Notes in Computer Science, Springer: Berlin, Heidelberg, 2012, 7190:56–83.
[18] M. Asada, Development of artificial empathy, Neuroscience Research, 2015, 90:41–50. https://doi.org/10.1016/j.neures.2014.12.002

[19] A. Kerasidou, Artificial intelligence and the ongoing need for empathy, compassion and trust in healthcare, Bulletin of the World Health Organization, 2020, 98(4):245–250. https://doi.org/10.2471/BLT.19.237198
[20] M. Jirásek and F. Sudzina, Big five personality traits and creativity, Quality Innovation Prosperity, 2020. 24(3):90. https://doi.org/10.12776/qip.v24i3.1509
[21] A. Furnham and H. Cheng, The Big-Five personality factors, mental health, and social-demographic indicators as independent predictors of gratification delay, Personality and Individual Differences, 2019, 150:109533. https://doi.org/10.1016/j.paid.2019.109533
[22] A. A. Zolotareva, Systematic review of the psychometric properties of the depression anxiety and stress scale-21 (DASS-21), V. M. Bekhterev Review of Psychiatry and Medical Psychology, 2020, 2:26–37. https://doi.org/10.31363/2313-7053-2020-2-26-37
[23] F. Toyoshima, A. Barton and O. Grenier, Foundations for an ontology of belief, desire and intention, Formal Ontology in Information Systems, 2020, 140–154. https://doi.org/10.3233/FAIA200667
[24] K. Su, X. Luo, A. Sattar and M. A. Orgun, The interpreted system model of knowledge, belief, desire and intention, AAMAS '06: Proceedings of The Fifth International Joint Conference on Autonomous Agents and Multiagent Systems, 2006, 220–222. https://doi.org/10.1145/1160633.1160668
[25] M. Harbers, K. V. Bosch, and J. C. Meyer, Modeling agents with a theory of mind: Theory-theory versus simulation theory, Web Intelligence and Agent Systems: An International Journal, 2012,10(3):331–343. https://doi.org/10.3233/WIA-2012-0250
[26] T. Kampik, J. C. Nieves and H. Lindgren, Explaining Sympathetic Actions Of Rational Agents. New York City: Springer International Publishing, 2019.
[27] A. Rosenfeld and A. Richardson, Explainability in human-agent systems, Autonomous Agents and Multi-Agent Systems, 2019, 33(3). https://doi.org/10.1007/s10458-019-09408-y