Humanized Artificial Intelligence: What, Why and How

C. Miao[1], Z. Zeng[1], Q. Wu[1], H. Yu[1], and C. Leung[1,2]

[1]LILY Research Center, Nanyang Technological University, Singapore
[2]The University of British Columbia, Vancouver, Canada

{ ascymiao, i160001, wu.qiong, han.yu, cleung}@ntu.edu.sg

Abstract

Today's artificial intelligence (AI) systems have surpassed human capabilities for task-oriented applications such as image recognition and playing chess or Go. The study of AI has gradually shifted from task-oriented tool-like objects which can execute computational and analytical works, to humanoid characters which are characterized by cognitive and social intelligence. As the next generation of technology that will be seamlessly woven into everyday lives, in addition to accuracy and efficiency, AIs need to be more human-like in order to interact, facilitate and collaborate with humans on various tasks. However, within the AI research community, this topic remains less familiar to many researchers. In this paper, we define the characteristics of an ideal **Humanized Artificial Intelligence (HAI)** which possess human-like traits and can establish long-term close ties with humans. We also propose a framework for guiding the design and development of HAI. We review and highlight the intuitions and key techniques that can be used in each area listed in the framework, then discuss promising future research directions towards successful integration of HAI systems into human societies.

Keyword: humanized artificial intelligence (HAI), Home AI, Personal AI, artificial curiosity, artificial persuasion, explainability

I. Introduction

As artificial intelligence (AI) technologies enter many areas of our daily life, the study of artificial agent, robotic or software, has gradually shifted from task-oriented tool-like objects which can execute computational and analytical works, to humanoid characters which are characterized by cognitive and social intelligence. Early AIs show some humanoid features by responding with instant emotions, dialogues, and gestures. However, these characters soon lost their humanoid halo due to their short-term characteristic. Driven by the motivation to remedy this drawback of intelligent agents, a new research area, Humanized Artificial Intelligence (HAI), targeting at development of a closer and long-term human-computer relationship was born. HAI is not a subfield or a simple extension from existing AI research and applications. It requires insights from multiple disciplinaries including psychology, sociology, education, gerontology, and artificial intelligence of course. The study of HAI concerns about basic research which will greatly expand fundamental knowledge in the AI field. The study of HAI also sheds light on many possible future AI applications which is not limited to specific scenarios.

In this paper, we argue that HAIs are inherently cognitive and social intelligent. Following these two characteristics, we proposed a humanized AI framework which mainly consists of a learning module and an anticipative common-sense knowledge model. Based on this HAI framework, we further discussed the key techniques that can infuse human-like traits into Home AI and Personal AI, which are two major application areas of HAI. The emphasis of the Home AI is the ability to understand the living context of the inhabitant and responding accordingly. Hence, our review focused on sensing and multi-modal signal processing techniques which can be employed the Home AI. The emphasis of Personal AI is the ability to interpret and respond verbally and non-verbally to the human user. Hence, we reviewed the research efforts in infusing the following human-like traits into AI: 1) curiosity, 2) persuasion, 3) explainablity and 4) emotion. In the end, we point out several promising future directions and conclude.

II. The Humanized AI Framework

Today's AI systems have surpassed human capabilities for specific applications such as image recognition and playing chess or Go. However, such performance is based on the availability of vast amount of training data and real-time external feedbacks. In real-life application scenarios involving interactions with a single user, such assumptions may not always be valid. For example, an AI companion for an elderly user cannot expect to learn to interact with the user through trial and error following traditional reinforcement learning as such an approach may be upsetting. Similarly, a rehabilitation system may not have a large amount of data about how a user with dementia expresses his anger for training a deep learning model. In order for AI to play a greater role in addressing real-life challenges, we need a new AI paradigm that is adaptive, robust and learns proactively.

In this paper, we propose a fundamentally new humanized AI framework (Figure 1) which can proactively learn problem-solving technique from a small number of examples, so that humanized AI can power home and personal smart devices that support humans in aspects of their lives. In order for AI to engage humans in intuitive and meaningful ways, it is necessary to improve current machine learning and decision support technologies. The proposed framework will shed light on the fundamental research question of how to enable AI to process multi-modal data, adapt to complex environmental changes, autonomously form goals and learn from sparse external feedbacks, thereby achieving near human-level intelligence. More specifically, the framework also highlights the key techniques that can infuse human-like traits into Home AI and Personal AI. A humanized home AI will be able to help the inhabitant to lead a safe, active and efficient lifestyle in their own homes, offering all-round support while doing this in an unobtrusive way. A humanized personal AI will be able to learn with intrinsic motivation, while offering persuasive, interpretable and affective interactions with humans.

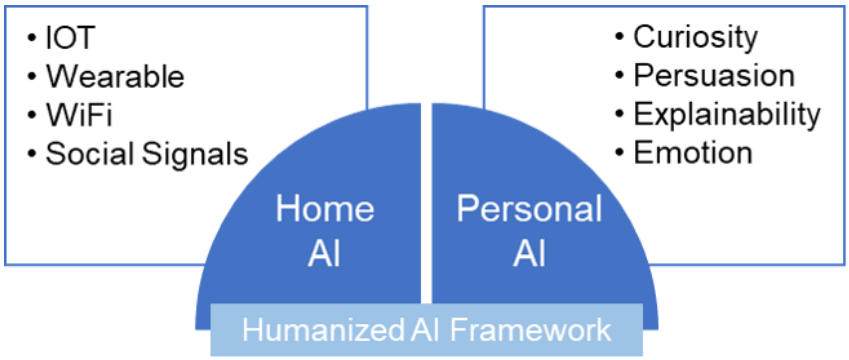

Figure 1 The Proposed Humanized AI Framework

Humans do not reason from scratch every time we encounter a new problem. We leverage on our existing knowledge and experience about the world to anticipate the likely outcomes of our decisions in new problems. Based on this observation, the proposed humanized AI framework learns problem-solving techniques from a small number of example human decisions. The architecture of this framework will broadly encompass two main parts: 1) a learning module which observes how human beings solve a given problem and imitates the techniques, and 2) an anticipative common sense knowledge model which represents the knowledge about how a given environment works and simulate multiple subsequent steps of actions for the AI agent to evaluate the long-term implications of various decision options and perform causal reasoning. The learning module will be designed following the methodology of ensemble learning to accommodate multiple learning techniques to be flexible and extensible. Using techniques such as knowledge graph and case-based reasoning, the common-sense model imbued with key learning points about previously encountered problems will allow AI to predict likely outcomes for new problems and continuously update itself.

This framework will be further enriched with curious reinforcement learning to support autonomous decision-making. Instead of assuming the availability of abundant real-time external feedbacks, we suggest designing a new architecture which can guide self-directed learning by AI through computational curiosity modelling. Such an architecture aims to replace external reward signals with intrinsic reward signals similar to human curiosity to drive the AI to explore and self-learn.

The proposed Humanized AI framework will represent a fundamental advance in the field of AI and serve as the technical foundation for the Home AI and Personal AI discussed below.

III. Home AI

A humanized home AI will be able to help the inhabitant to lead a safe, active and efficient lifestyle in their own homes, offering all-round support while doing this in an unobtrusive way. For this purpose, care and support needs to be delivered in a personalized manner. The first step towards realizing personalized care and support is to gain comprehensive insight into each individual's health condition, lifestyle, and habits.

Various sensing technologies can be employed by a Home AI to understand the living context of an inhabitant. For example, such techniques include personal area sensor networks incorporating Internet-of-Things (IoT), wearable devices and WiFi-based unobtrusive in-home privacy preserving technologies to collect longitudinal activity data. Results from decentralized partially observable Markov decision processes (Dec-POMDP) can be used to improve the robustness and accuracy of decisions under sensor malfunctions, and uncertainties in the environment and user behaviors.

Based on the multi-modal sensing data, social signal processing methods can be designed to gain actionable insights into personal activity patterns and physical, cognitive and emotional health. Availability of such personal data can lead to many potential research directions, such as detection and analysis of abnormal changes in physiological vital signs, diet, exercise, social interactions, and their significance on general health condition. In addition, models that can aggregate information from multiple sensors to predict the risk of falls, stroke and neurodegenerative diseases can be proposed and validated based on personal data.

A humanized Home AI should be able to coordinate all smart devices in home in order to provide ubiquitous computing service in a holistic manner. To illustrate this, consider a medication reminding scenario. Many seniors require medications in order to manage chronic multi-morbid conditions. With declining memory and cognitive capabilities, it is often difficult for a senior to take

the correct amount of the right medications at the right time. Although smart medicine dispensing devices are commercially available, the deterioration in prospective memory capability (McDaniel & Einstein, 2007) among the elderly has limited their usefulness. It is currently difficult for medical professionals to know if the failure of a treatment plan is due to limitations of the plan or patient non-adherence. A Home AI will be able to address the problem of treatment adherence in a holistic manner. Depending on the reason for non-adherence as perceived by the sensing network, algorithmic approaches can be designed to dynamically persuade (in the case of deliberate non-adherence) or remind (in cases where non-adherence is due to forgetfulness) post-surgery or chronic disease patients to adhere to their prescribed recovery regimes. Instead of relying on a single smart medicine dispensing machine, a Home can employ a collection of memory technologies that can be incorporated into display areas commonly found in a home environment (e.g., TVs, digital photo stands, smartphones, wall projections, etc.).

IV. Personal AI

Data collected by Home AI constitute a rich source of longitudinal information about individual's physical, cognitive conditions as well as their lifestyle and habits. Such personal data can be used to create personalized profiles to guide the design and implementation of a personalized AI which can support its human user in aspects of his/her daily lives, no matter is it for productivity, entertainment or lifestyle management.

In order to provide rich natural intuitive interactions, it is necessary to investigate how to incorporate human-like characteristics into the Personal AIs. In this paper, we have identified four important human-like characteristics:

- **Curiosity.** This characteristic is especially useful from a machine learning perspective. As reward signals in real world interactions with humans may be scarce, traditional reinforcement learning approaches may be ineffective. With curiosity modelling, Personal AI will be able to generate innate reward signals which encourage them to

explore and learn about the environment and the target user in the absence of external rewards.

- **Persuasion.** Persuasion refers to the process of altering others' thoughts, attitudes and behaviors and is crucial in human-human interaction. It is also an important characteristic for a Personal AI which will need to make suggestions/ recommendations, support decision making and help manage lifestyle of its user. Persuasion modelling can improve the effectiveness of Personal AI in human-agent interaction and build an independent agent "personality" which can stand for its own thoughts and opinions.
- **Explainability.** The advent of big data era has brought great success to machine learning and led to an explosion of new AI models and applications. However, much of machine learning remains opaque as a "black box". In order for humans to develop trust in AI systems, it is necessary to design explainable AI (XAI) systems which are able to explain their decisions and behaviors and behave in a consistent way.
- **Emotion.** Emotion is a very important form of non-verbal communication. In order to establish close emotional tie with humans, Personal AI should be able to sense and respond to the emotional state of its user appropriately and address the user's social needs. Infusing emotion into Personal AI help develop more affective, personalized and believable social interactions. Emotions that expressed as consistent and long-term behavioral tendencies can also help to form the "personality" of the Personal AI.

In the following part of this section, we will discuss each of the four characteristics in details, reviewing existing attempts in modelling these characteristics and highlighting key techniques that can be used.

A. Curiosity

As a variance of intrinsically motivated learning, curiosity-based learning has come into the scope of AI for just a few years. However, curiosity has a long psychological background ever since 1890, when William James published *The Principles of Psychology* (James, 1890) in which he

described two kinds of curiosity. The first is an instinctual or emotional response, in which attention is aroused by seeing something new. And the second is a "scientific curiosity" and "metaphysical wonder" in which the "brain responds to an inconsistency or a gap in its knowledge, just as the musical brain responds to a discord in what it hears" (Borowske, 2005). James has laid the foundation of psychological study of curiosity for the early twentieth century.

Then, in 1960s, Daniel Berlyne (Berlyne, 1960) took a new angle in investigation of curiosity and made a new division of curiosity: diversive and specific. Diversive curiosity is a general tendency for a person to seek novelty, take risks, and search for adventure. Specific curiosity is a tendency to investigate a specific object or problem aiming to understand it. He further suggested the factors underlying to stimulate curiosity which involve complexity, novelty, uncertainty, and conflict.

Nowadays, curiosity has been studied as a mechanism for achieving and maintaining high levels of well-being and meaning in life (Kashdan & Steger, 2007). All these psychological evidences give rise to the consideration of implementing curiosity into machine learning as an intrinsic motivation.

Evidence also comes from the field of neuroscience. Studies by Kakade and Dayan (Kakade & Dayan, 2002) suggest that dopamine not only plays a critical role in extrinsic motivational control of behavior, but also in the intrinsic motivational control of behavior associated with novelty and exploration. Novel sensory stimuli induce an unpredicted rewards kind of activity of dopamine cells (Schultz, 1998). Reed (Reed, Mitchell, & Nokes, 1996) argues that novelty itself has rewarding characteristics because novel sensory stimuli would have less effect on the dopamine cells when they become familiar.

Schmidhuber (Schmidhuber, 1991) implemented curiosity as the motivation in his system for exploratory learning of the environment and basic sensory-motor coordination. Curiosity plays a fundamental role in minimizing the prediction error. Curiosity has been constantly utilized in developing self-evolving robots (Ugur, Dogar, Cakmak, & Sahin, 2007; Roa, Kruijff, & Jacobsson, 2009).

Wu et al. proposed the world's first generic computational model of curiosity for intelligent agents, based on a well-established curiosity theory in psychology and an agent modelling theory in AI. This computational model simulates the human curiosity assessment process in intelligent agents through a formalized abstract view which allows implementation in and adaptation to different application contexts. Wu has also developed a curiosity-driven AI learning algorithm which significantly improves the traditional extreme learning machine by allowing online adaptation of the neural network structure (Wu & Miao, 2015). This algorithm first selects interesting data from the input stream through the assessment of curiosity-stimulating factors, such as novelty, uncertainty, conflict and surprise. It then employs smart learning strategies which mimic a curious human learner.

B. Persuasion

Personal AIs are meant to offer not only a long-term companionship to the user, they should also have the tendency to persuade the user to do something good that the user would not otherwise do by themselves (Benyon & Mival, 2007). Persuasion is the process that changes people's belief, attitude, and behavior in a predetermined fashion (Fogg, 1998). Social persuasion abounds in human interactions. Our beliefs, attitudes and behaviors are constantly experiencing subtle changes influenced by our friends, parents, teachers and anyone around us. Katagiri et.al argued that the social factors of agent design such as affiliation, authority and conformity can also have an effective social persuasive power in human-computer interaction (Katagiri, 2001).

Persuasive technology has a great impact on education, health and Safety (Fogg, 1998). Pedagogical companion equipped with persuasive technology have been tested to be more motivational, and learner engaging (Okonkwo & Vassileva, 2001). The elderly is vulnerable to serious health problems, and normally are physical inactive. If their Personal AIs are equipped with persuasive tactics that gradually persuade the elderly to take a more active role in physical exercise and cultivate them with a better eating habit, their quality of life will be greatly improved, and life span extended. A recent hot research area is to combine persuasive technology with

ubiquitous computing to help people change to an ideal lifestyle (Consolvo, McDonald, & Landay, 2009). If well designed, Personal AI would be a suitable interface to persuade people due to its long-term, close characteristics in the relation with the user.

In order to model computational persuasion for AI, Kang et al. (Kang, Tan, & Miao, 2015) proposed a Model for Adaptive Persuasion (MAP). By adopting MAP, Personal AI can execute personalized strategies and provide personalized responses. MAP is also able to autonomously learn the individual users' traits or preferences and adjust its strategy according to the users' behavioral changes over time. MAP is a semi-connected network model that enables an agent to adapt its persuasion strategies through feedback. It is based on the Elaboration Likelihood Model (ELM), a proven theory of the thinking processes that might occur when we attempt to change a person's attitude through communication. Specifically, ELM assumes that there are two major routes to choose from when one is attempting to persuade others: the central route and the peripheral route. Which route to choose depends on a person's motivation and ability to process the message presented to him/her. Following ELM, the MAP enables Personal AI to adapt to the user's perceived personal state, better execute personalized strategies and provide personalized responses to the user.

Besides the ones reviewed above, existing literatures are largely focused on more broadly defined persuasive technology, rather than the modelling of computational persuasion. The persuasive technology has taken various forms in the literatures:

- *Affective cues*: The affect valence has an influence on cognition. In general, positive affect makes people rely more on simplified knowledge structures, stereotypes and judgment heuristics when taking in new information, while negative affect makes people more likely to scrutinize (Griskevicius, Shiota, & Neufeld, 2010). Burleson et al investigated into the persuasive ability of an affective learning companion in influencing the perseverance of learner when facing failure (Burleson & Picard, 2007). Cavazza et.al implements affective

strategies into an ECA companion to influence user attitudes in offices (Cavazza, et al., 2010).

- *Embodiment*: Brinol reviewed the social psychological literature on how embodied agent influence attitudes with their body postures and movement (Brinol & Petty, 2008). Bodily responses such as nodding head, arm flexion serve as cues of the extent of thinking.
- *Negotiation*: Persuasive tactics are key factors in successful negotiations. The commitments to actions are determined by the answers of the recipient (agree or disagree) which affect the persuasive tactics (threats or rewards) that can be used in the next round of negotiation (Schelling, 1980). A research branch called persuasive negotiation is widely studied; Jennings et.al gave the general requirements for persuasive negotiation mechanism (Jennings, et al.).
- *Games*: Persuasive games have been successful in marketing, advertising, research and design for the past 30 years, and have become an effective platform for leading advocacy groups and lifestyle brands to communicate. An example game is Fish'n'Steps, which encourages users to lead a healthier life by taking more steps every day. The daily count of steps of the user is related to the emotional state, activity and growth of a virtual fish in a virtual fish tank.
- *Text message, Blogs and podcasts*: Blogs and podcasts are tools that can promote a sense of community in a group. The study of Firpo aimed at changing the attitude and behavior of a group of students to promote their sense of community and their research showed that social presence and credibility play a big role in the persuasion (Firpo, Kasemvilas, Ractham, & Zhang, 2009).

C. Explainability

A major source of distrust in AI stems from the fact that many AI systems are black-boxes, which evolve to behave in unexpected ways. In order to enhance the acceptance of AI recommendations

by humans, transparency is an important first step. Explainable AI based on argumentation allows an AI system to automatically explain why certain recommendations are being made, under what conditions recommendations are likely to succeed/fail, and how a user can react to correct an error (Langley, Meadows, Sridharan, & Choi, 2017). Thus, it is important to build more explainable AI, so that humans can understand, trust and effectively manage the emerging AI systems.

Integrating the taxonomies proposed in (Gunning, 2016) and (Biran & Cotton, 2017), and summarized by (Zeng, Miao, Leung, & J., 2018), existing research in Explainable AI (XAI) can be categorized into three broad approaches: (1) explanations based on features, (2) model approximation and (3) interpretable models.

For explanations based on features, usually a noninterpretable complex model and its predictions are given. This approach focuses on generating justifications for the predictions by extracting and identifying the features that have significant effects on the prediction results. Martens et al. (Martens, Huysmans, Setiono, Vanthienen, & and Baesens) explain the results of an SVM classifier by extracting rules that can produce similar results to the SVM based on a small subset of features. Landecker (Landecker, et al., 2013) interpret the classification results of hierarchical networks by studying the degree of importance of different components to the classification results. Hendricks et al. (Hendricks, et al., 2016) generate explanations for image classification results of a CNN using a LSTM, based on both prominent image features and class discriminative features.

The second approach, model approximation, involves model-agnostic methods that infer an explainable model from any black-box models. Robnik-Sikonja and Kononenko (Robnik-Sikonja & Kononenko, 2008) decompose a model's prediction to the level of individual features by comparing the model results when a feature value is present and absent. More recently, Ribeiro, Singh, and Guestrin (Ribeiro, Singh, & Guestrin, 2016) explain a prediction instance by constructing interpretable model locally around it, which is only an accurate approximation of the global model in the vicinity of that instance.

The third approach, interpretable model, aims to construct models that are inherently structured and interpretable, such as rule lists and decision trees. Si and Zhu (Si & Zhu, 2013) use And-Or-Trees to represent the possible component structures of objects in images, which can be used as compositional models for explaining results of object detection. Lake, Salakhutdinov, and Tenenbaum (Lake, Salakhutdinov, & Tenenbaum, 2015) learn a generative model of character images, and explain a character recognition result using the generation process of that character.

Despite much work having been done in this field, there are still several gaps that need to be bridged. Firstly, existing XAI models can answer the question “why this decision or conclusion”, while they cannot answer “Why not”. Secondly, most of them offer explanations as a form of evidences rather than reasons.

Good explanations need to reveal the underlying reasoning process and are best presented in human-interpretable terms. Zeng et al. (Zeng, et al., 2018) have taken an argumentation-based approach to XAI. Argumentation is the study of how reasonable decisions can be reached by constructing for and against arguments and evaluating these arguments accordingly. Argumentation-based approach to decision making appears to be closer to the way humans deliberate, evaluate alternatives and make decisions. This leads to a transparent decision-making process and the ability to offer understandable reasons underlying the decisions made. In this approach, arguments are generated for different decision alternatives, which are then evaluated against some decision criteria. The argument structures used during the evaluation process offers a formalism from which explanations can be extracted.

D. Emotion

Romano attributed emotion as a very important form of non-verbal communication so as to truly be one’s best friend. Emotion has recently been implemented into artificial intelligence in developing effective, personalized and believable social interactions (Romano, 2007).

There are three levels of affect states distinguished by the temporal discourse: emotions refer to "behavioral dispositions" that last for only seconds or minutes, moods are "states" that have similar

effects but last over a longer temporal discourse of hours or days, and personality are "traits" that reflect comparatively stable and long-term behavioral tendencies.

Emotion

In the last decades, there have been a great many computational models developed and Marsella et.al (Armony, 2005) has made a comprehensive review of existing computational models and projects that utilize these models. They categorized the existing computational models into five classes:

- Appraisal, which emphasize the relation between emotion and cognition and concern with the relationship between individual's beliefs, desires, intentions and events;
- Dimensional: instead of treating emotions as discrete entities, dimensional theory conceptualizes emotions and represent them as points in a continuous space;
- Anatomical: anatomically inspired models try to simulate the neural links and processes that undergoes the emotion transitions of organisms;
- Rational: views emotion as adaptive functions and tries to abstract these adaptive functions from human emotion processes;
- Communicative: views emotion as a communicative system, by sharing one's mental states with others to facilitate social coordination.

Mood

Though mood has been highly recognized as an intermediate state between emotion and personality, little literature has been found to form a comprehensive computational model for mood as OCC for emotion and Big-five Factor for personality.

Personality

Personality is a long-term and comparatively stable trait of individuals that underlies and influences the current emotional disposition. The most widely accepted markers for personality trait is the "Big-Five Factor structure" by Goldberg (Goldberg, 1992), which describe personality

in terms of 1) Surgency or Extraversion, characterized by sociable, outgoing, confidence, etc; 2) Agreeableness, characterized by friendliness, pleasant, etc; 3) Consciousness or Dependability, characterized by helpful, hard-working, etc; 4) Emotional Stability or, Neuroticism, characterized by adjusted etc; 5) Culture, Intellect, or Openness, characterized by imaginative, flexibility etc.

Emotion has been widely studied and implemented in early sociable robot which has been recognized as the first generation of artificial companion. The Seal robot Paro (Kidd, Taggart, & Turkle, 2006), was equipped with simple emotional reactions, and the field study in a nursing home showed its ability in improving the sociability of the user. Huggable (Stiehl, et al., 2006), another social robot developed by MIT media lab, is a teddy-bear with touch sensitive skin and voice coil actuator, also serves the similar pet-like functionalities as Paro. HOMIE (Kriglstein & Wallner, 2005), an artificial dog with emotional behavior developed to provide the elderly with entertainment to disperse their feeling of solitude. Field studies by Heerink demonstrated that these sociable robots do have a positive impact on the elderly and the social presence of the artificial companions enhances the acceptance by the elderly (Heerink, Kröse, Evers, & Wielinga, 2008).

In the current artificial companion design, emotions are studied as facilitation for conducting real life conversations. Cavazza et.al developed an affective ECA which can sooth emotion at work for adults (Cavazza, Camara, & Turunen, 2010). Salvi et.al studied the emotional speech synthesis using a statistical classification method (Salvi, Tesser, Zovato, & Cosi, 2010). Also, emotion is highly recognized as an effective persuasion tool in affecting human attitudes and behavior. Cavazza et.al also implemented persuasive abilities in their ECA to influence the attitude of the workers at work using affective strategies (Cavazza, et al., 2010).

Studies in emotional artificial intelligence are not limited in self-oriented emotions, but also in emotions related to others empathic emotions. These companions can analyze the emotional feedback and respond to the original with proper emotional states in real-time. They will provide the originals with a sense of being cared about and easy to build a rapport relationship between the

ACs and the originals. The early start of building empathic is the job interview companion, which can accompany the user in a virtual job interview environment, recognizing and measuring the affective states of the user in a form of empathetic feedback. McQuiggan et.al developed a unified inductive framework CARE which can drive runtime situation-appropriate empathetic behaviors after training and accurately assess social contexts and generate parallel and reactive empathy (McQuiggan, Robison, Phillips, & Lester, 2008). Boukricha proposed a computational model to simulate the theory of mind of a virtual character, which is based on mimicry and role-taking mechanisms to perform empathetic behaviors. Empathic behavior has been utilized in conducting more proper and comfortable dialogues. Bee et.al developed an empathic listener Alfred which can give three level of empathic feedback (Boukricha, 2008). McRorie et.al investigated how behaviors are influenced by personalities and constructed a Sensitive Artificial listener which can reflect its personality credibly as prescribed (McRorie, Sneddon, de Sevin, Bevacqua, & Pelachaud, 2009).

V. Conclusions and Future Research Directions

In this paper, we define the characteristics of an ideal Humanized Artificial Intelligence (HAI) which possess human-like traits and can establish long-term close ties with humans. We also propose a framework for guiding the design and development of HAI. The HAI framework broadly encompasses two main parts: 1) a learning module which observes how human beings solve a given problem and imitates the techniques, and 2) an anticipative common sense knowledge model which represents the knowledge about how a given environment works and simulate multiple subsequent steps of actions for the AI agent to evaluate the long-term implications of various decision options and perform causal reasoning.

Based on the HAI framework, we further discussed the key techniques that can infuse human-like traits into Home AI and Personal AI. The emphasis of the Home AI is the ability to understand the living context of the inhabitant and responding accordingly. Hence, our review focused on sensing techniques which can be employed the Home AI. To build an AI that communicates like a human

one need to consider its behavior and its ability to interpret and respond verbally and non-verbally to the human user. In order to do so, an AI should be able to generate its own behavior according to its perception of the interaction with the human user and/or the environment. Consequently, an AI needs to have a computational model that allows itself to generate understandable and convincing expressions, emotions and credible behavior (Romano, 2007). Hence, we reviewed the research efforts in modelling the following human-like traits in AI: 1) curiosity, 2) persuasion, 3) explainablity and 4) emotion.

Based on the research reviewed, we envision several possible future research directions which can impact this field (HAI) going forward.

A. Curiosity

Besides learning, curiosity can also be used to guide better recommendations. Currently, social information has mainly been utilized for enhancing rating prediction accuracy, which may not be enough to satisfy user needs. Items with high prediction accuracy tend to be the ones that users are familiar with and may not interest them to explore (Wu, Liu, & Miao, 2017). By adding curiosity into AI recommendations, AI can explore more alternative that may be interesting but not familiar to the human user, hence improving the coverage and diversity of AI recommendations.

B. Explainability

Improving transparency can enhance the acceptance of AI recommendations, suggestions and decisions by human. On one hand, Explainable AI allows an AI system to automatically explain why certain decisions / recommendations are being made. However, complete transparency may be impractical if the objective is to persuade a user to follow a time-critical recommendation but can be useful for a posterior analysis of the AI decision process. On the other hand, inadequate transparency may limit users' trust in AI. Further research to gain insight into users' satisfaction, mental models of the AI system, and task performance is required in order to improve the interpretability of AI systems. Human factors design should be used to determine how and when to offer explanations to seniors about the decisions made by the AI system. To develop effective

senior friendly AI tools to support productive aging, we need to bring together seniors, caregivers, healthcare professionals, scientists and policymakers. This will help to create the inclusive and respectful AI technology ecosystem for the aging societies of tomorrow.

C. Persuasion and Explainability

According the state of the person being persuaded, effective persuasion techniques may vary. The Elaboration Likelihood Model (ELM) states that, when the person is highly motivated and has cognitive ability to comprehend the persuasion message, he/ she is likely to scrutinize the persuasion message carefully. In this case, logical and well-grounded persuasive arguments are useful persuasion cues. When an AI is trying to persuade its user to adopt its decisions/ recommendations, a good explanation can be used as a strong argument. On the other hand, good explanations should also be persuasive. Hence, persuasion and explainability can be considered together to improve the modelling of both.

D. Rising trend – ethical AI

As AI technologies enter many areas of our daily lives, the problem of ethical decision making, which has long been a grand challenge in AI, has caught increasing public attention. It is important that AI decisions and recommendations be made under an ethical framework. Future research in ethical AI could focus on four major directions, namely: 1) understanding how people from different backgrounds make decisions under ethical considerations, 2) enabling individual AI agents to assess the ethical impact of decisions made by itself and other agents, 3) enabling a group of AI agents to collectively evolve towards ethical decision-making, and 4) enabling AI agents to communicate the ethical aspects of their decisions to humans through appropriate emotional cues. Such a general ethical decision-making framework can be a useful starting point for AI to provide a supportive environment in which its recommendations prioritize human wellbeing.

References

Armony, J. (2005). Computational models of emotion. *IEEE International Joint Conference on Neural Networks(IJCNN)*, 1598-1602.

Benyon, D., & Mival, O. (2007). *Introducing the companions project: intelligent, persistent, personalised interfaces to the internet*. Retrieved 12 28, 2018, from http://researchrepository.napier.ac.uk/3784

Berlyne, D. E. (1960). *Conflict, Arousal, and Curiosity.* McGraw-Hill. Retrieved 12 28, 2018, from https://books.google.com/books?id=Z1x9AAAAMAAJ&pg=PP1

Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: a survey. *IJCAI-17 Workshop on Explainable AI (XAI)*, (pp. 8-13).

Borowske, K. (2005). Curiosity and Motivation-to-Learn. *Proceedings of The Twelfth National Conference of the Association of College and Research Libraries*, 346-350.

Boukricha, H. (2008). *A first approach for simulating affective theory of mind through mimicry and role-taking*. Retrieved 12 31, 2018, from http://techfak.uni-bielefeld.de/~hboukric/papers/afirstapproachforaffectivetom.pdf

Brinol, P., & Petty, R. (2008). Embodied persuasion: Fundamental processes by which bodily responses can impact attitudes. *Embodiment grounding: Social, cognitive, affective, and neuroscientific approaches*, 184-207.

Burleson, W., & Picard, R. (2007). Gender-specific approaches to developing emotionally intelligent learning companions. *Intelligent Systems*, 62-69.

Cavazza, M., Camara, R. S., & Turunen, M. (2010). How was your day?: a companion ECA. *Autonomous Agents and Multi-Agent Systems*, 1629-1630. Retrieved 12 31, 2018, from http://dblp.uni-trier.de/db/conf/atal/aamas2010.html

Cavazza, M., Smith, C. G., Charlton, D., Crook, N., Boye, J., Pulman, S., . . . Turunen, M. (2010). *Persuasive dialogue based on a narrative theory: an ECA implementation*. Retrieved 12 28, 2018, from https://link.springer.com/chapter/10.1007/978-3-642-13226-1_25

Consolvo, S., McDonald, D. W., & Landay, J. A. (2009). *Theory-driven design strategies for technologies that support behavior change in everyday life*. Retrieved 12 28, 2018, from http://dub.washington.edu/djangosite/media/papers/paper0552-consolvo.pdf

Firpo, D., Kasemvilas, S., Ractham, P., & Zhang, X. (2009). Generating a sense of community in a graduate educational setting through persuasive technology. *Proceedings of the 4th International Conference on Persuasive Technology*, 41-51.

Fogg, B. J. (1998). *Persuasive computers: perspectives and research directions*. Retrieved 12 28, 2018, from http://cse.chalmers.se/research/group/idc/ituniv/kurser/07/idproj/papers/fogg.pdf

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4*, 26. Retrieved 12 31, 2018

Griskevicius, V., Shiota, M., & Neufeld, S. (2010). Influence of different positive emotions on persuasion processing: A functional evolutionary approach. *Emotion*, 190-206.

Gunning, D. (2016). *Explainable Artificial Intelligence (XAI)*. Retrieved 12 31, 2018, from DARPA: https://www.darpa.mil/program/explainable-artificial-intelligence

Heerink, M., Kröse, B. J., Evers, V., & Wielinga, B. J. (2008). *The Influence of Social Presence on Acceptance of a Companion Robot by Older People*. Retrieved 12 31, 2018, from https://pure.uva.nl/ws/files/4248389/58668_285120.pdf

Hendricks, L. A., Akata, Z., Rohrbach, M., Rohrbach, M., Donahue, J., Schiele, B., & Darrell, T. (2016). Generating Visual Explanations. *arXiv: Computer Vision and Pattern Recognition*, 3-19. Retrieved 12 31, 2018, from https://arxiv.org/abs/1603.08507

James, W. (1890). *The principles of psychology.* New York: Camelot Press.

Jennings, N., Faratin, P., Lomuscio, A., Parsons, S., Wooldridge, M., & Sierra, C. (n.d.). Automated negotiation: prospects, methods and challenges. *Group Decision and Negotiation, 10*(2), 199–215. Retrieved 12 28, 2018

Kakade, S., & Dayan, P. (2002). Dopamine: generalization and bonuses. *Neural Networks*, 549-559.

Kang, Y., Tan, A.-H., & Miao, C. (2015). *An adaptive computational model for personalized persuasion*. Retrieved 12 31, 2018, from http://dblp.uni-trier.de/db/conf/ijcai/ijcai2015.html

Kashdan, T. B., & Steger, M. F. (2007). Curiosity and pathways to well-being and meaning in life: Traits, states, and everyday behaviors. *Motivation and Emotion, 31*(3), 159-173. Retrieved 12 28, 2018, from http://mason.gmu.edu/~tkashdan/publications/moem_curiosity_wb_and_meaning.pdf

Katagiri, Y. a. (2001). Social persuasion in human-agent interaction. *Second IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems.* Seattle.

Kidd, C. D., Taggart, W., & Turkle, S. (2006). *A sociable robot to encourage social interaction among the elderly*. Retrieved 12 31, 2018, from http://dblp.uni-trier.de/db/conf/icra/icra2006.html

Kriglstein, S., & Wallner, G. (2005). *HOMIE: an artificial companion for elderly people*. Retrieved 12 31, 2018, from http://dblp.uni-trier.de/db/conf/chi/chi2005a.html

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, 1332–1338.

Landecker, W., Thomure, M. D., Bettencourt, L. M., Mitchell, M., Kenyon, G. T., & and Brumby, S. P. (2013). Interpreting individual classifications of hierarchical networks. *2013 IEEE Symposium on Computational Intelligence and Data Mining*, (pp. 32-38).

Langley, P., Meadows, B., Sridharan, M., & Choi, D. (2017). Explainable agency for intelligent autonomous systems. *Proceedings of the 29th AAAI Conference on Innovative Applications of Artificial Intelligence (IAAI-17)*, (pp. 4762–4763).

Martens, D., Huysmans, J., Setiono, R., Vanthienen, J., & and Baesens, B. (n.d.). Rule extraction from support vector machines: an overview of issues and application in credit scoring. *Rule extraction from support vector machines*, 33-63.

McDaniel, M. A., & Einstein, G. (2007). Prospective memory: An overview and synthesis of an emerging field. Sage Publications Ltd.

McQuiggan, S., Robison, J. L., Phillips, R., & Lester, J. C. (2008). Modeling parallel and reactive empathy in virtual agents: an inductive approach. *Autonomous Agents and Multi-Agent Systems*, 167-174. Retrieved 12 31, 2018, from https://intellimedia.ncsu.edu/wp-content/uploads/empathy-aamas-2008.pdf

McRorie, M., Sneddon, I., de Sevin, E., Bevacqua, E., & Pelachaud, C. (2009). A model of personality and emotional traits. *Intelligent Virtual Agents*, 27-33.

Okonkwo, C., & Vassileva, J. (2001). Affective pedagogical agents and user persuasion. *Universal access in HCI: Towards and information society for all Proceedings*, 397-401.

Reed, P., Mitchell, C. J., & Nokes, T. (1996). Intrinsic reinforcing properties of putatively neutral stimuli in an instrumental two-lever discrimination task. *Animal Learning & Behavior, 24*(1), 38-45. Retrieved 12 28, 2018, from https://link.springer.com/article/10.3758/bf03198952

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 1135-1144).

Roa, S., Kruijff, G.-J. M., & Jacobsson, H. (2009). *Curiosity-driven acquisition of sensorimotor concepts using memory-based active learning*. Retrieved 12 28, 2018, from http://yadda.icm.edu.pl/yadda/element/bwmeta1.element.ieee-000004913080

Robnik-Sikonja, M., & Kononenko, I. (2008). Explaining classifications for individual instances. *IEEE Transactions on Knowledge Data Engineering*, 589-600.

Romano, D. (2007). The Look, the Emotion, the Language and the Behaviour of a Companion at Real-Time. *IEEE International Joint Conference on Neural Networks(IJCNN)*, 38-40.

Salvi, G., Tesser, F., Zovato, E., & Cosi, P. (2010). *Cluster Analysis of Differential Spectral Envelopes on Emotional Speech*. Retrieved 12 31, 2018, from http://speech.kth.se/prod/publications/files/3502.pdf

Schelling, T. (1980). *The strategy of conflict.* Harvard University Press.

Schmidhuber, J. (1991). Curious model-building control systems. *IEEE International Joint Conference on Neural Networks*, 1458-1463.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of neurophysiology*, 1-27.

Si, Z., & Zhu, S.-C. (2013). Learning AND-OR Templates for Object Recognition and Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(9), 2189-2205. Retrieved 12 31, 2018, from http://ieeexplore.ieee.org/document/6425379

Stiehl, W. D., Breazeal, C., Han, K.-h., Lieberman, J., Lalla, L., Maymin, A. Z., . . . Kishore, A. (2006). *The huggable: a new type of therapeutic robotic companion*. Retrieved 12 31, 2018, from https://doi.acm.org/10.1145/1179849.1179866

Ugur, E., Dogar, M. R., Cakmak, M., & Sahin, E. (2007). *Curiosity-driven learning of traversability affordance on a mobile robot*. Retrieved 12 28, 2018, from http://ieeexplore.ieee.org/document/4354044

Wu, Q., & Miao, C. (2015). *C-ELM: A Curious Extreme Learning Machine for Classification Problems*. Retrieved 12 28, 2018, from https://link.springer.com/content/pdf/10.1007/978-3-319-14063-6_30.pdf

Wu, Q., Liu, S., & Miao, C. (2017). *Modeling uncertainty driven curiosity for social recommendation*. Retrieved 12 31, 2018, from https://dl.acm.org/citation.cfm?doid=3106426.3106475

Yu, H., Shen, Z., Miao, C., Leung, C., Leung, C., Lesser, V. R., & Yang, Q. (2018). *Building Ethics into Artificial Intelligence*. Retrieved 12 31, 2018, from https://ijcai.org/proceedings/2018/779

Zeng, Z., Fan, X., Miao, C., Leung, C., Jih, C. J., & Soon, O. Y. (2018). Context-based and Explainable Decision Making with Argumentation. *Autonomous Agents and Multi-Agent Systems*, 1114-1122. Retrieved 12 31, 2018, from http://celweb.vuse.vanderbilt.edu/aamas18/acceptedpapers

Zeng, Z., Miao, C., Leung, C., & J., C. J. (2018). Building More Explainable Artificial Intelligence with Argumentation. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence.*