CHAPTER TEN

Psychological Modeling of Humans by Assistive Robots

A. Wagner, E. Briscoe
Georgia Institute of Technology, Atlanta, GA, United States

1. INTRODUCTION

Healthcare presents a large problem space to which novel robotic solutions can be applied. An important and evolving question is how healthcare-focused robots will interact with and assist the people they serve. Assisting a patient is, in many ways, fundamentally different from interacting with a healthy adult. Being a patient implies some type of ailment that could potentially debilitate or impair the person's judgment. Patient-robot interaction may therefore require a unique perspective on how and why a robot should interact with the person. As a caregiver, situations may arise in which the robot provides companionship, protects the patient's dignity, restrains a person to prevent self-harm, or discretely observes their behavior. As researchers in this space, it is imperative that we recognize that to be effective, robots must be accommodating and flexible, with the capability to interact with a wide range of personality types, disabilities, and levels of intelligence.

The dynamic nature of patient-robot interaction presents challenges and opportunities for novel research. For some cases, the possibility of creating a "one size fits all" robot-aided therapy is unrealistic. As we move from robots as assistive tools, to robots as active collaborators in maintaining and upgrading one's physical and mental health, we must develop systems that autonomously individualize their behavior over the course of treatment with respect to the needs of the patient. Creating systems that perceive and model a person's mental state is an important problem that is currently being investigated by the human-robot interaction (HRI) and social robotics communities. This chapter presents their progress and also examines an anticipated upcoming generation of robots that will socially interact with patients, model their moods, personality, likes, and dislikes, and use this information to guide the robot's assistance related decisions. These robots will need to recognize when a person trusts them, or when a person is being deceptive; and also decide whether to act deceptively in return for a beneficial patient outcome. We review the major research in areas related to cognitive, behavioral, and psychological modeling of a patient by an assistive robot.

The chapter begins by examining the dimensions or attributes by which people are commonly characterized, especially as relevant to a person needing assistance, focusing on qualities utilized in human-human interaction, such as age and emotional state. The person's condition is another important distinguishing factor that a more general purpose assistive robot might need to consider. For example, does the person have a neurological deficit that might influence how the robot should interact with the person? A hearing deficit might make verbal commands difficult or impossible to understand for some patients. Lessons learned while developing a medication and water delivery robot for older adults suggest that the mobility of the potential user has a critical impact on their view and likelihood of using the system [1]. Emeli et al. interviewed older adults asking about their interest in a robot tasked with delivery. Although some individuals expressed interest, the person's current mobility played an important factor in their assessment of the system (Fig. 1).

Next we present methods for behavioral modeling from more traditional robotics settings, such as from the field of HRI and social robots. As will be demonstrated, the use of psychological modeling by HRI researchers is still in its infancy.

Fig. 1 A water and medication delivery robot. Users requested service from the robot by pressing a button on their smart phone indicating their location. The robot then traveled to a medication and water station for pickup of the items to be delivered, traveled to the stated location, and stopped within an arm length of the person.

We then present the dominant approaches for psychologically modeling a person. Different categories of models exist related to the fields from which they arose. Economic models, for example, focus on the decisions people make and how these decisions impact their internal state or utility. Typically, different economic models of human behavior are assessed in terms of their ability to predict people's decisions in tightly controlled laboratory experiments. These approaches, although not directly related to assistive scenarios, are important because they offer a formal, grounded, computational means for modeling a person that can be implemented in the software that controls a robot.

Cognitive models are another important category of approach for modeling people. Cognitive models attempt to approximate the different types of processes that underlie cognition. Various architectures have been advanced as different visions of how animals and people think. Often these architectures are assessed in terms of the accuracy of their predictions of human behavior, and the timing of these behaviors. Next, we discuss how human characteristics may be measured, based on explicit signals, in order to be utilized in individual-based models. The chapter concludes by summarizing the work done in the area of human modeling by assistive robots, and suggesting new avenues for future work.

2. DIMENSIONS OF HUMAN CHARACTERIZATION

Human characterization serves a critical purpose, whether conducted by other humans or machines. Several of the social sciences, such as cognitive and social psychology, share a similar goal of determining characterizations of humans that serve to explain both their commonalities and individual differences. Often investigators seek to understand how and why people differ and how those differences may be measured, wherein by defining people using common dimensions, they can better explain, understand, and predict an individual's behavior. Developing computational methods for representing these dimensions also provides a means for artificial reasoners (eg, robots) to mimic and take advantage of known human categorization techniques.

Whether explicitly or implicitly, people constantly categorize the individuals with which they interact. As cognitive processes are optimized to compensate for limited processing power by maximizing resources, humans often form impressions and infer traits based on their immediate categorization of other individuals rather than waiting to experience each specific trait firsthand [78]. Research into how people categorize other people varies, for example, by spanning both social and individuated concepts [2,3]; focusing on either a person's salient characteristics (such as height) or behavioral characteristics (such as personality); or involving top-down (category-oriented) or data-driven (attribute-oriented) inferences [4]. It is also expected that when people evaluate one another, they do so in stages, where initial categorization, confirmatory categorization, recategorization, and piecemeal processing (as described by Fiske et al. [5]) are mediated by attention and interpretation and influenced by information and motivation.

While not exhaustive, many of the dimensions used to characterize humans that have been identified through social science research, and that are especially relevant to benefiting HRI, are described in the following paragraphs.

2.1 Personality

Perhaps the most common method, either explicitly or implicitly, of describing a person is by assigning them a range of personality traits. There are several primary theories describing personality [6], including the "Big Five" [7], Eysenck's PEN [8], and Myers-Briggs Type Indicator [9]. These theories "decompose" people into multiple dimensions. For example, using the Big Five, people are characterized along five factors: openness, conscientiousness, extraversion, agreeableness, and neuroticism (see Table 1 for a

Table 1 "Big Five" Personality Dimensions [7] and Their Associated Trait Ranges

Dimension	Trait Ranges		
Conscientiousness	Organized	↔	Disorganized
	Careful	↔	Careless
	Disciplined	↔	Impulsive
Agreeableness	Soft-hearted	↔	Ruthless
	Trusting	↔	Suspicious
	Helpful	↔	Uncooperative
Neuroticism	Calm	↔	Anxious
	Secure	↔	Insecure
	Self-satisfied	↔	Self-pitying
Openness	Imaginative	↔	Practical
	Prefers variety	↔	Prefers routine
	Independent	↔	Conforming
Extraversion	Sociable	↔	Retiring
	Fun-loving	↔	Sober
	Affectionate	↔	Reserved

more comprehensive description of the dimensions and their associated range of traits). Despite its potential to be dynamic (there is debate as to whether people have consistent personalities [10]), personality profiling has proven valuable in predicting a number of behaviors, such as job performance in occupational settings [11].

2.2 Emotions and Moods

Understanding a person's mood or emotional state is critically important, especially with regard to potential HRI. One primary approach in the study of emotions has been to view emotions as arising from a palette of "basic emotions." These perspectives tend to perceive basic emotions as "low level" feelings such as anger, disgust, fear, joy, sadness, surprise [79]. The identification of emotion "types" based on the theory of Ortony, Clore, and Collins [12], also known as the "OCC" model, assumes that emotions develop as a consequence of certain cognitions and interpretations, which are based on one's understanding of the world in terms of agents, objects, and events (see Table 2 for the 22 emotions described in the model). How emotions might emerge is very dependent on how individuals perceive and interpret events. One can be pleased or displeased about the consequences of an event (*pleased/displeased*); one can endorse or reject the actions of an agent (*approve/disapprove*) or one can like or not like aspects of an object (*like/dislike*).

A further differentiation consists of the fact that events can have consequences for others or for oneself, and that an acting agent can have different roles. The consequences of an event for another can be divided into *desirable* and *undesirable*; the consequences for oneself as relevant or irrelevant expectations. Relevant expectations for oneself finally can be differentiated again according to whether they actually occur or not (*confirmed/disconfirmed*).

Moods, in contrast, are not usually characterized by their direction at a person or event. While someone might show an emotion (anger) toward a specific object (eg, a colleague), as the specific emotion dissipates, they may feel a general bad mood. In some models of emotion (eg, [13]), an individual's perception of a generated emotion is considered the agent's *feelings*, which are modulated by both emotion, the individual's appraisal of the current situation, and mood, which is a memory of recent emotions [14].

2.3 Intelligence

Intelligence is a well-studied individual characteristic whose measurement has been used to predict a wide variety of human behavior, such as job

Table 2 The Types and Specifications of the 22 Emotions Specified by the OCC Model [12]

Emotion Type	Specification
Joy	(Pleased about) a desirable event
Distress	(Displeased about) an undesirable event
Happy-for	(Pleased about) an event presumed to be desirable for someone else
Pity	(Displeased about) an event presumed to be undesirable for someone else
Gloating	(Pleased about) an event presumed to be undesirable for someone else
Resentment	(Displeased about) an event presumed to be desirable for someone else
Hope	(Pleased about) the prospect of a desirable event
Fear	(Displeased about) the prospect of an undesirable event
Satisfaction	(Pleased about) the confirmation of the prospect of a desirable event
Fears-confirmed	(Displeased about) the confirmation of the prospect of an undesirable event
Relief	(Pleased about) the disconfirmation of the prospect of an undesirable event
Disappointment	(Displeased about) the disconfirmation of the prospect of a desirable event
Pride	(Approving of) one's own praiseworthy action
Shame	(Disapproving of) one's own blameworthy action
Admiration	(Approving of) someone else's praiseworthy action
Reproach	(Disapproving of) someone else's blameworthy action
Gratification	(Approving of) one's own praiseworthy action and (being pleased about) the related desirable event
Remorse	(Disapproving of) one's own blameworthy action and (being displeased about) the related undesirable event
Gratitude	(Approving of) someone else's praiseworthy action and (being pleased about) the related desirable event
Anger	(Disapproving of) someone else's blameworthy action and
Love	(Liking) an appealing object
Hate	(Disliking) an unappealing object

performance [80]. Most current factor models of intelligence typically represent cognitive abilities as a three-level hierarchy, where at the highest level is a single factor referred to as *g*, which represents the variance common to all cognitive tasks [81]. Intelligence has also been linked to personality traits, especially openness and intellect [7], where individuals who fall higher on the openness dimension generally score higher on measures of cognitive ability [15] and are more creative [16].

2.4 Social Intelligence

In addition to traditional measures of intelligence, *social intelligence* [82] is a quality that represents a person's ability to express, recognize, and manage social signals and social behaviors [17], such as politeness and empathy. As a subset of social intelligence, *emotional intelligence* involves a person's ability to monitor his own and others' feelings and emotions and to make informed decisions based on this assessment [18]. The five key factors of emotional intelligence are: self-awareness, self-regulation, motivation (passion for work and resiliency), empathy, and social skills. In order to be useful for a robot, these dimensions must be represented within a computational framework. The following sections present some of the most common and well-known computational frameworks for modeling people.

3. CONSTRUCTING BEHAVIORAL MODELS FOR HRI

Because a robot has a physical presence in the world, the ways that people interact with these machines can, in some ways, be unique. For instance, Bethel et al. compared robot interviews to human interviews in terms of the misinformation effect [19]. The misinformation effect occurs when a person's recall is influenced by postevent information in a way that makes recall less accurate. In criminal investigations, for example, questioning by the examiner influences the eyewitnesses' memories of the event. Bethel et al. found that use of a robot interviewer resulted in greater memory recall and accuracy in spite of misinformation. In an assistive setting, use of a robot might allow a patient to more accurately state the reason for their ailments.

Human-robot research has also shown that people are likely to heed the orders of a robot. McColl and Nejat found that 87% of older adults that they tested complied with a robot's prompts to eat [20]. Salem et al. found that people typically comply with a robot, even when its instructions are in error [21]. They found that two-thirds of subjects would even pour orange juice over a plant if asked to by the robot. Most rationalized the request in some way.

It is worth noting that anthropomorphism can strongly influence the ways that a person interacts with a robot. Anthropomorphism is the tendency to attribute human characteristics to nonhuman objects. When the object being anthropomorphized is a robot, this tendency can lead to inaccurate expectations related to the robot's behavior, intentions, or communication tendencies [22]. HRI researchers have, at times, resorted to developing robotic "creatures" in order to reduce anthropomorphism [23]. Others have embraced anthropomorphism by making their robots as humanlike as possible [24]. Gratch et al. have pioneered the use of humanlike virtual characters for the purpose of training, and for psychotherapeutic effectiveness [25]. The nature of a robot's interactions with the patient will likely determine whether anthropomorphism is embraced or avoided.

The field of HRI has begun to explore the possibility of a robot modeling a person and using this information to guide its behavior. This work has typically fallen under the rubric of mental modeling and the development of shared mental models. Unlike an economic model or a cognitive model, a mental model is simply one's understanding of another person's thought process [26]. As such, the notion of a mental model is rather vague and ill-defined. Also, unlike economic and cognitive models, mental models have no natural computational underpinnings. Hence, roboticists tend to create their own. Propositions are a common format for representing an individual's beliefs [27]; probabilistic statements are also often used [28].

Belief-desire-intention models (BDI) have also been used to model the psychological state of a person. Beliefs represent the information that the person is currently aware of, or in some way knows. Desires represent the person's goals or motivations, and intentions denote a person's deliberative course of action [29]. Although useful in simulation and virtual environments, BDI methods are limited in their ability to capture the richness and variability associated with human decision making.

Partially-observable Markov decision processes (POMDPs) have also been suggested as a means for representing a model of the robot's interactive partner. POMDPs model the decision process as a chain of connected states [30]. At each time step, the robot observes information about the current state of the environment which is ultimately used to select an action during the next time step. Although the resulting behavior is provably optimal, exact POMDP solutions are computationally intractable for most nontrivial problems; approximate solutions are the norm [31].

In addition to modeling the person's mental state, it may also be useful for an assistive robot to model the risks to the patient associated with a therapy. Exoskeletons, for example, are poised to become an important therapeutic tool and perhaps even a long-term solution for some types of paralysis. It may be beneficial for these systems to model the risks associated with particular types of movements, such as climbing steps, and warn the user of the risk.

Finally, there are important ethical ramifications associated with a robot's creation and use of psychological models when interacting with a person.

Such machines may be empowered to recall an enormous amount of data associated with a person or type of person. Depending on the circumstances, this information could easily be used to manipulate or unjustly coerce a patient. Developers of robotic systems must consider the extent to which psychological modeling is appropriate and justified.

Given the limitations of current approaches, we can generate greater realism by incorporating elements of human psychology and the signals these elements produce.

4. ECONOMIC DECISION-MAKING MODELS

The earliest and most well-developed formal models of human decision making originate from economic theory. Jeremy Bentham developed the idea that economic exchange and the decision making that underlies it are driven by one's attempt to maximize pleasure and minimize pain, and can be measured as such [32]. A person's motivations and resulting behavior, Bentham argues, are a direct reflection of the utility of one's actions. At its core, utility theory claims that people use subjective assessments of the value of an action choice to make decisions. If an individual views a particular action or course of action as offering higher subjective value than some other course of action, that action is favored over others.

In contrast to many traditional psychological theories, utility theory is formalized mathematically. This mathematical grounding provides a computational framework which is implementable on an assistive robot attempting to model and predict a person's behavior. Utility theory assumes that a person receives a quantifiable pleasure, $u \in \mathbb{R}$, when making a decision. Utility functions, $U: X \to \mathbb{R}$, are used to describe an individual's preferences in relation to fixed set of choices. A preference is defined as a relation, $\preccurlyeq$, over X such that for every $x, y \in X$, $U(x) \leq U(y)$ implies $x \preccurlyeq y$. Rational behavior results when an individual selects actions which maximize their utility function. In theory, understanding a person's utility function allows one to predict the person's behavior.

Expected utility theory considers one's preferences when the utilities that result from an action choice are uncertain [33]. Uncertainty is typically modeled as a gamble resulting in outcome (x) with respect to a known probability distribution (p). The expected utility, EU, is then arrived at as the product of the outcome utilities and the probabilities. The outcome of several gambles, $x = \{x_1, x_2, \ldots, x_N\}$, is thus calculated as $EU[x] = \sum_i^N p_i x_i$.

John von Neumann and Oskar Morgenstern showed that an individual whose preferences satisfy certain axioms always prefer an action which maximizes their expected utility [34]. Alternatively, an individual whose preferences violate the von Neumann and Morgenstern axioms makes decisions which guarantee losses. Intentionally selecting an action which will result in losses is considered irrational. A rational individual is thus argued to be an individual who selects actions that maximize their own utility.

One of the primary criticisms of expected utility theory is that it falsely assumes that people act in a rational, utility-maximizing manner. In reality, well-known biases constantly influence human decision making. For instance, people tend to be averse to losses. A person's decisions are more strongly influenced by the desire to avoid losses (*loss aversion bias*) than the desire to seek gains. The manner in which a situation is described or framed also impacts decision making (*framing effects*). Framing a choice as a gain or a loss has been shown to bias decision making. Humans are also biased to prefer the status quo. The status quo is used as a point of reference for evaluating whether or not a choice will result in a loss or a gain. People tend to discount the value of additional gains or losses the further the gain or loss is from one's reference point. For example, the perceived difference in utility between receiving \$100,000 and \$100,100 is considered small, whereas the perceived difference in utility between receiving \$0 and \$100 is considered great, even though the value of the difference is the same. Finally, people tend to overestimate the likelihood of low-probability events and underestimate the probability of high-probability events.

Prospect theory was developed to better account for the behavior people actually exhibit when faced with choices under risk and uncertainty [35]. Prospect theory holds that the expected utility should be evaluated as $EU[x] = \sum_i^N w(p_i)v(x_i)$, where EU is the expected utility associated with making decisions $x_1, \; x_2, \; \ldots, \; x_N$ and $p_1, \; p_2, \; \ldots, \; p_N$ are their respective probabilities. The function v maps outcome values to utilities and the function w weighs probabilities in order to capture one's risk preferences. Fig. 2 depicts the typical shape of a prospect theory value function. The y-axis illustrates the individual's current reference point. Losses and gains are evaluated with respect to the reference point. The steeper slope on the loss portion of the graph models loss aversion. The curved tails indicate the lessening impact of additional gains or losses, a reflection of gain/loss satiation.

Fig. 3 depicts a probability weight function from prospect theory. The curved line used for the weight function mimics certainty effects by

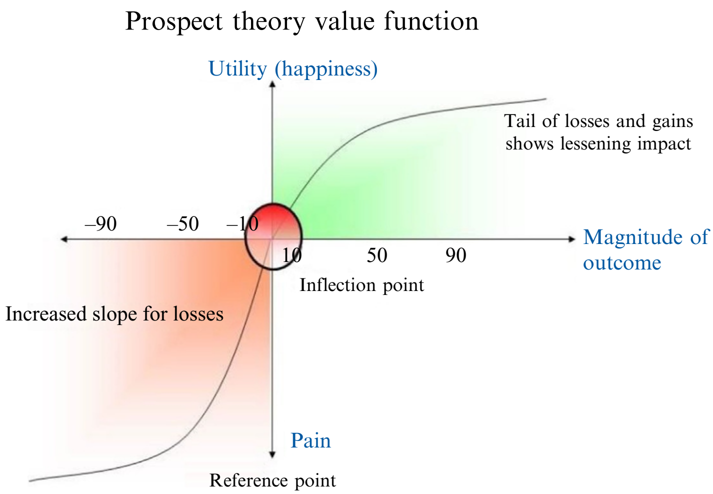

Fig. 2 An example weighted utility function from prospect theory.

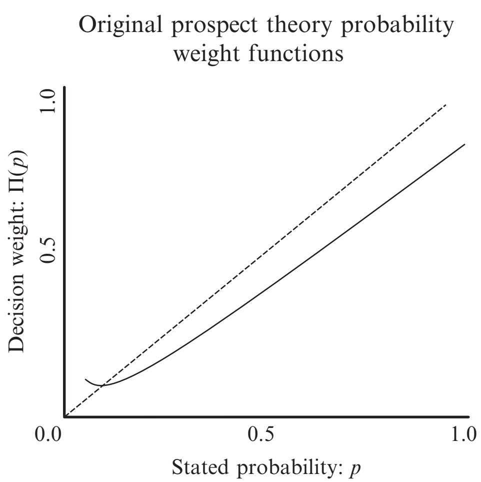

Fig. 3 Tversky and Kahneman's original prospect theory probability weight function [35].

overestimating the likelihood of low-probability events and overestimating the likelihood of high-probability events. Overall, the addition of a probability weighting and utility valuation function and the shape that these functions take allow prospect theory to better model a person's economic decision making as well as their biases.

Although prospect theory was a major advancement, a problem remained [36]. Let x and y be the outcomes that result from gambles

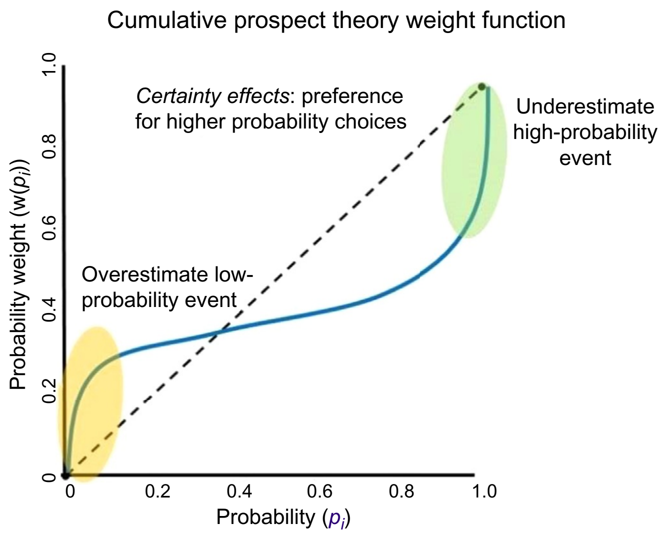

Fig. 4 The probability weight from function from cumulative prospect theory.

A_1 and A_2. If $P(x > t|A_1)$ is the probability that an outcome of Gamble A_1 exceeds t, then the decision (A_1) is said to "stochastically dominate" decision (A_2) if and only if $P(x_1 > t|A_1) \geq P(x_2 > t|A_2)$ for all t, assuming $A_1 \neq A_2$. Unfortunately, because of the shape of the weighted probability function, prospect theory predicts situations in which A_2 is chosen over A_1 in spite of the fact that A_1 is stochastically dominated by A_2. *Cumulative prospect theory* is a theoretical and practical refinement of prospect theory which does not violate stochastic dominance [37]. Fig. 4 depicts a weighted probability function that maintains stochastic dominance.

Many extensions and refinements of these behavioral decision theories have been proposed. In contrast to utility theory, regret theory focuses an individual's motivation to minimize the negative feelings associated with regret [38–40]. Regret theory holds that the preference over choices A_1 can be represented formally as:

$$A_1 \succeq A_2 \Leftrightarrow \sum_i^N p_i(v(A_1(x_i)) - v(A_2(x_i)))$$

where v is the utility function. Regret theory explains certain decisions better than prospect theory. For a more formal and detailed treatment of regret theory, see Ref. [40].

The preceding theories are based on rational models of human behavior. Decision field theory, on the other hand, is a dynamic theory that does not assume a fixed, rational set of preferences [41]. Decision field theory includes a model of preference evolution that allows it to generate a better account of decision regularity and make predictions about decision time under certain conditions. For more detailed information related to decision field theory, Busemeyer and Diederich provide a survey of the field [42].

For robots trying to model a person, decision theories offer a computational starting point. They are easily implemented on a robot and share some common conceptual traits with more traditional robotics frameworks such as reinforcement learning. Yet decision theories also tend to assume rational or semirational behavior and underestimate the idiosyncratic, emotional, and automatic nature of human behavior. Moreover, some argue that a single scalar utility value is too impoverished a model to represent a person's state. Nevertheless, this approach has been employed successfully for more than 60 years by economists and others. With respect to robotics, Wagner showed that a robot could learn categories of models related to different types of people [43]. These stereotyped models could then be used to make predictions about interactions with newly encountered people. These techniques could presumably be used by an assistive robot as a strategy for bootstrapping its early interactions with a new patient. Later, with time and experience, the system could tailor its interactions to needs of the particular patient.

4.1 Neuroeconomics

Given the limitations of utility-based economic models of human decision making, researchers began to investigate methods that augment traditional economic models with results and data from neuroscience. They called this new direction *neuroeconomics*. Neuroeconomics is an interdisciplinary field which seeks to build from utility-based theories while also including evidence from neuroscience. As a field, neuroeconomics makes a concerted effort to include automatic and emotional processing in their models of human behavior [44]. There is a great deal of evidence that human decision making is strongly influenced by automatic, unconscious processes [45,46] as well as emotion [47]. Emotion, in particular, has a tendency to hijack a person's decision-making faculties and generate behavior which is typically viewed as irrational. Emotion also impacts learning which, in turn, influences decision making [13]. Significant evidence suggests that memories are imbued with affective information, which upon recall, generates similar emotions to the time the memory was created.

Neuroeconomics embraces the use of brain imaging as a means to better understand the mind's processing while making decisions. Typically, brain images are used to compare brain activity while people are engaged in an experimental or control task. Often the tasks used involve economic games such as the prisoner's dilemma or the investment game. King-Casas et al., for example, used a hyperscan functional Magnetic Resonance Imager (fMRI) to monitor participant's neural responses while playing a two-person investment game [48]. In this game one person assumes the role of investor and the other of trustee. The investor is given $20 and may invest any portion of the money with the trustee. Money that is invested with the trustee then appreciates (is multiplied by 3). The trustee must then decide how much to return to the investor. King-Casas et al. monitored subjects over 10 rounds of play. They found neural correlates which indicated a person's intention to trust their partner.

These types of studies provide insight into the areas of the brain that are activated in specific decision-making situations as well as the temporal process underlying that decision. For instance, it has been shown that situations involving distrust activate the parts of the brain associated with the fear emotion [49], whereas situations involving trust do not activate any specific area of the brain [50].

Significant work in this area has also shown that decisions are made when the brain integrates inputs from multiple systems to generate a utility-like value. This multiple systems model claims that information from these different systems is processed in a qualitatively individualized manner and weighted accordingly [51]. For example, the emotional or affective system quickly and unconsciously processes information to generate a reflexive response. The analytic system, on the other hand, is a slower, conscious effort which influences behavioral decisions. The behavior that results is a combination of signals from these different sources (Fig. 5).

The multiple systems model of cognition can be used to explain several aspects of human behavior. For instance, Shiv and Fedorikhin [52] showed that self-regulation decreases with increased cognitive load by asking people to choose between cake or fruit salad while remembering either a 2-digit or 7-digit number. Significantly more people choose cake when asked to remember the 7-digit number. Presumably remembering the larger number marshals resources away from the executive functions of the mind, allowing one's lower functions to play a larger role in decision making.

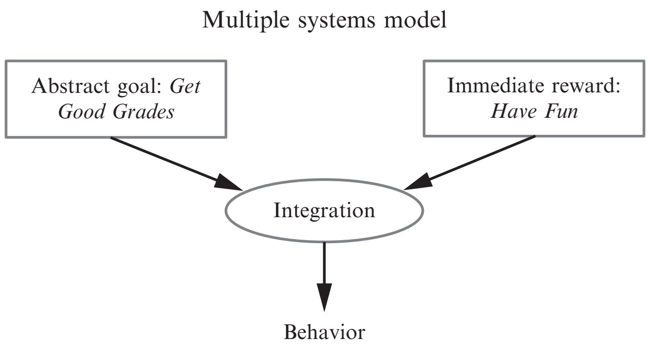

Fig. 5 An example of the multiple system's model.

Integration of these signals is not necessarily limited to behavior. Perceptual processes, for example, are intimately influenced by motor activity. Jessica Witt has found that motor experience derived from practice actually increases a person's perceived size of the target [53]. For instance, better baseball batters perceive pitched baseballs as larger. The same has been demonstrated with field goal kickers and golfers. Witt even found that, when compared with nonthreatening objects, people perceived spiders as moving faster [54].

This interplay of different, competing systems influences people's decisions during interactions with a robot. Robinette et al. examined a situation in which a robot guided subjects to a meeting room [55]. Later, an emergency evacuation occurred and the same robot offered to guide them out of the building. In virtual environments subjects that were initially led by a robot that navigated poorly generally choose not to use the robot. In a real instantiation of the experiment, subjects universally chose to follow the robot regardless of how poorly it had navigated earlier. The multiple systems hypothesis provides a possible explanation for this behavior. In the virtual environment, people's high executive functions consider the robot's reputation when making a decision. In the real environment, when alarms are sounding, their lower more automatic cognitive functions dominate and the robot's reputation is no longer considered.

4.2 Cognitive Architectures

The challenge of evaluating potential interaction outcomes may be addressed by representing features as components in a human cognitive architecture, such as those proposed by Anderson and Lebiere [83], Revelle et al. [15], or Newell [84], which consists of computational components that, in coordination, are argued to produce human-level intelligence. These architectures allow a robot to "understand" and predict human behavior on some level [56]. Several of these architectures, such as H-CogAff, ACT-R, and Soar, operate at a high level of comprehensiveness, where they include key psychological moderators, such as personality.

By utilizing the cues provided by both humans and contextual information, cognitive architecture(s) may greatly enhance a robot's ability to interact with an individual. Trafton et al. [56] specifically discuss an expansion of ACT-R, ACT-R/E to provide robots with theory of mind (knowledge of others' cognition) so as to improve their functionality. With an understanding of how people might perform in different situations, the robot can better achieve its own goals. For example, reasoning about how people are likely to react during emergencies (such as a fire) in order to protect themselves may impart a robot with knowledge that will aid it to locate victims. Likewise, the ability to detect and reason about emotion may be a critical component of robotic caregivers [57]. These types of models can be used as a source of information by the robot for decision making.

In regard to assistive robots, one should expect that the person's mental state will strongly influence their decision making relative to the robot. For instance, emotional patients, younger and older patients, and those with traumatic brain injuries may not consider the longer-term benefits of the robot. It is important that an assistive robot respond to the person it is trying to help. Failing to monitor and react to changes in the person's mood or behavior is likely to lead to little desire by the person to use the robot and may even result in patients injuring themselves to prevent assistance.

Overall, economic models of human decision making may have a large role to play with respect to assistive robotics technologies. These models offer a formal, computational means for the systems to predict and possibly understand the person's behavior. While these models have limitations, there are currently few well-developed alternatives. Perhaps the most well-developed alternatives are the cognitive models presented in the next section.

5. INFERRING PSYCHOLOGICAL MODELS

Implementing social science research allows an artificially intelligent machine to better represent, understand, and ultimately interact with a human. Better machine understanding of human behavior may result in more natural and beneficial HRI (eg, [58]). In order to take advantage of the knowledge that has been provided by social science research, machines must be able to perceive and interpret information about the human with which it is interacting. Efforts in social signal processing [59,60] have determined that social signals commonly identified by psychologists and sociologists can be recognized and captured by machinery, such as microphones and cameras, and processed into intelligence through algorithmic techniques, such as machine learning. The ability to utilize these signals allows a machine (robot) to create "mental" models of people in a way very similar to humans. These models then allow for reasoning about human behavior, which can be optimized towards tailoring successful machine interactions.

Nonverbal behavior is a continuous source of signals that convey information about the traits of people, such as their emotions and personality [85]. Ekman and Friesen [61] notably categorized communication types that result from nonverbal cues. These include: affective/attitudinal/cognitive states (eg, fear, joy, stress, disagreement, ambivalence, and inattention), emblems (culture-specific interactive signals like wink or thumbs up), manipulators (eg, touching objects in the environment), illustrators (eg, finger pointing and raised eyebrows), and regulators (eg, head nods and smiles).

Using nonverbal behaviors is not necessarily sufficient, as they carry a great deal of ambiguity [85]. For example, an awkward posture may indicate aggression or an injury. Culture is also a contributing factor to individual differences [86], and must be considered during nonverbal cue interpretation. To overcome these complications, it is advisable to marry features across modalities, though attention to the context in which the behavior arises is also critically important.

In order to model the behavior of a person, it is necessary to discover the subset of features relevant to a specific signal [17]. Often, the best approach is to select the most relevant features from all available data; however, this may result in only selecting features that are not relevant to any specific individual but only to an average model.

5.1 Detecting and Modeling Psychological Characteristics

5.1.1 Personality

Humans exhibit a number of cues that can be used to infer personality (eg, [62,63]). These cues may take many forms, such as the language that a person uses [64], a person's environment [65], or the tone in which they speak [66].

Personality cues may be used to create profiles to understand and, in some cases, predict an individual's behavior. For example, Walters et al. [87]

investigated whether personality, as characterized by certain personality traits, could be used to predict the likely approach distance preferred by the human subjects in robot interaction experiments. They found that their variable "proactiveness," primarily made up of creativity and impulsiveness, was positively correlated with the preferred human-to-robot approach distance. A person's personality may also be utilized to help optimize robotic interactions, where extroverted individuals may prefer language that emphasizes friendliness and warmth, and introverted individuals may prefer slow movements and more silence.

5.1.2 Emotion and Mood

The previously described OCC model, which defines 22 emotions (described in Table 1), is extremely useful for modeling emotional agents as the authors explicitly constructed the model to allow for "reasoning about emotion" [12], by assuming that individuals perform a subjective assessment of their relationship to the environment. Objects, events, and actions are evaluated in an appraisal process based on specific criteria, and result in multiple emotions of different intensities. As a person interacts with objects or agents, they evaluate the benefits or potential harm that they may cause, based on concerns such as goals, standards, or tastes. If those concerns are satisfied, which may be detected through recognition of cues such as facial recognition [67], the individual experiences a positive emotion (eg, admiration, joy); otherwise, a negative emotion (eg, frustration) is elicited. Past research has found many cues to emotion detection, much of it arising from research in *affective computing* [68]. Past research has also determined other methods for automatic emotion detection, such as facial expression analysis using artificial machine vision [69], voice analysis to detect the emotion of the interlocutor [88], or multimodal analysis [70]. The detection efficacy of modalities differs across emotions, where some emotions are better identified by voice, such as sadness and fear, but others are better detected through facial expression analysis, such as happiness and anger [89]. By representing the generation of emotions, and, over a longer temporal scale, moods, artificial agents can utilize small cues, such as changes in facial expressions, to provide additional meaning to communications and detect subtle changes in an individual.

5.1.3 Measuring General Intelligence

Though intelligence is traditionally measured through standardized testing (eg, [71]), other methods infer intelligence through its relation to other observable traits, such as personality and expressed interests (eg, [72,73, 90]) and even physical characteristics (eg, [74]). Understanding a person's intelligence may also aid in HRIs. For example, when interacting with highly intelligent persons, interactions may benefit from presenting ideas in more technical depth, using words that are more difficult and asking challenging questions.

5.1.4 Measuring Social Intelligence

Inferring social intelligence is a bit more straightforward than general intelligence, as relevant behaviors may be expressed through various cues in reaction to interactions, such as gestures and facial expressions. The use and incorporation of human social signals is often referred to as *socially aware computing* [59]. Vinciarelli et al. [17] describe the primary social cues as falling into the following classes: physical appearance, gestures and posture, face and eyes behavior, vocal behavior, and space and environment. Relevant for social intelligence, Salovey and Mayer [18] proposed a model that identified four different factors of emotional intelligence: the perception of emotion, the ability to reason using emotions, the ability to understand emotion, and the ability to manage emotions. An individual's place at each of these dimensions may be used to determine optimal interaction methods [75].

5.2 Utilizing Context

Modeling behavior is extremely dependent on the situation and context in which that behavior is exhibited. Human behavior is highly variable, changing and adapting according to the situation. What may be construed in one situation (eg, a smile upon greeting another, indicating happiness) can be representative of something completely different in another (eg, a smile after seeing another get hurt, indicating malice). In order to utilize a behavioral cue in one context, it is necessary to understand that context—potentially by determining the 5 W+ questions (who, what, when, where, why, and how) [17].

Similar to categorizing individuals, categorizing the situation may provide meaningful knowledge for optimizing a robot's response. For example, in situations where a decision must be made immediately, such as in emergencies, individuals may be forced to rely on instinct- or experience-based processes, which may be viewed as irrational [76]. Decision making in these situations differs from nonemergency reasoning, where there are higher stakes, higher uncertainty, and increased time pressure [77]. These considerations support the need for a comprehensive means to understand and represent human behavior, especially as it is likely to change in different contexts and situations.

6. CONCLUSIONS

Robots are becoming an important facet of physical and rehabilitative therapy. Exoskeletons and autonomous social robots may soon assist people with daily tasks. We argue that as robots become prevalent, it will be important that they model the people and patients that they interact with in a way that includes the many different psychological facets that make humans human. We have explored many of the major research directions by which these models are realized and developed. The inclusion of methods for behavioral and psychological modeling as a part of a robot's decision making is a new and exciting area of research. As this avenue of research grows, one should expect to see robots that are less aligned with our traditional notion of robots. These systems will appear more human and interactive. Although these models will likely be beneficial for a number of tasks, researchers and the community at large must give serious consideration to the ethical implications of creating robots which psychologically model humans.

REFERENCES

[1] V. Emeli, A.R. Wagner, C.C. Kemp, A robotic system for autonomous medication and water delivery, Technical report GT-IC-12-01, College of Computing, Georgia Institute of Technology, 2012.
[2] M.B. Brewer, A dual process model of impression formation, in: Advances in Social Cognition, Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, vol. 1, 1988.
[3] S.T. Fiske, S.L. Neuberg, A continuum of impression formation, from category-based to individuating processes: influences of information and motivation on attention and interpretation, Adv. Exp. Soc. Psychol. 23 (1990) 1–74.
[4] T.M. Ostrom, Between-theory and within-theory conflict in explaining context effects in impression formation, J. Exp. Soc. Psychol. 13 (5) (1977) 492–503.
[5] S.T. Fiske, M. Lin, S.L. Neuberg, The continuum model: ten years later, in: Dual-Process Theories in Social Psychology, Guilford Press, New York, NY, 1999, pp. 231–254.
[6] R. Ewen, Personality: A Topical Approach, Lawrence Erlbaum Associates, London, 1998.
[7] P.T. Costa, R.R. McCrae, NEO PI-R Professional Manual, Psychological Assessment Resources, Odessa, FL, 1992.
[8] H.J. Eysenck, The Biological Basis of Personality, Charles C. Thomas, Springfield, IL, 1967.
[9] I.B. Myers, The Myers-Briggs Type Indicator: Manual, Consulting Psychologists Press, Palo Alto, CA, 1962.
[10] A. Bandura, Social cognitive theory of personality, in: Handbook of Personality: Theory and Research, Guilford Press, New York, 1999, pp. 154–196.
[11] M.R. Barrick, M.K. Mount, The big five personality dimensions and job performance: a meta-analysis, Pers. Psychol. 44 (1) (1991) 1–26.
[12] A. Ortony, G.L. Clore, A. Collins, The Cognitive Structure of Emotions, Cambridge University Press, Cambridge, 1990.

[13] A. Damasio, Descartes' Error: Emotion, Reason, and the Human Brain, Avon Books, New York, 1994.

[14] R.P. Marinier, J.E. Laird, Computational modeling of mood and feeling from emotion, Power 1 (1) (2007) 461–466.

[15] W. Revelle, J. Wilt, A. Rosenthal, Personality and cognition: the personality-cognition link, in: A. Gruszka, G. Matthews, B. Szymura (Eds.), Handbook of Individual Differences in Cognition: Attention, Memory and Executive Control, Springer, New York, 2010, pp. 27–49.

[16] R.R. McCrae, Creativity, divergent thinking, and openness to experience, J. Pers. Soc. Psychol. 52 (6) (1987) 1258–1265.

[17] A. Vinciarelli, M. Pantic, H. Bourlard, Social signal processing: survey of an emerging domain, Image Vis. Comput. 27 (12) (2009) 1743–1759.

[18] P. Salovey, J. Mayer, Emotional intelligence, Imagin. Cognit. Personal. 9 (3) (1990) 185–211.

[19] C.L. Bethel, D. Eakin, S. Anreddy, J.K. Stuart, D. Carruth, Eyewitnesses are misled by human but not robot interviewers, in: 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE, 2013.

[20] D. McColl, G. Nejat, A socially assistive robot that can monitor affect of the elderly during meal-time assistance, J. Med. Devices 8 (3) (2014) 030941.

[21] M. Salem, G. Lakatos, F. Amirabdollahian, K. Dautenhahn, Would you trust a (faulty) robot? Effects of error, task type, and personality on human-robot cooperation and trust, in: 10th ACM/IEEE International Conference on Human-Robot Interaction, 2015.

[22] B.R. Duffy, Anthropomorphism and the social robot, Robot. Auton. Syst. 42 (3) (2003) 177–190.

[23] C. Breazeal, Designing Sociable Robots, MIT Press, Cambridge, MA, 2002.

[24] H. Ishiguro, S. Nishio, Building artificial humans to understand humans, J. Artif. Organs 10 (3) (2007) 133–142.

[25] G. Stratou, S. Scherer, J. Gratch, L.-P. Morency, Automatic nonverbal behavior indicators of depression and PTSD: exploring gender differences, in: International Conferences on Affective Computing and Intelligent Interaction, Geneva, Switzerland, 2013.

[26] P.N. Johnson-Laird, Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness, Cambridge University Press, Cambridge, 1983.

[27] P. Felli, T. Miller, C. Muise, A.R. Pearce, L. Sonenberg, Artificial social reasoning: computational mechanisms for reasoning about others, in: The International Conference on Social Robotics, {ICSR}, 2014.

[28] M.A. Goodrich, D. Yi, Toward task-based mental models of human-robot teaming: a Bayesian approach, in: Proceedings of HCI International, Las Vegas, Nevada, 2013.

[29] X. Zhao, J. Venkateswaran, Y.J. Son, Modeling human operator decision making in manufacturing systems using BDI agent paradigm, in: Proceedings of Annual Industrial Engineering Research Conference, 2005, pp. 14–18.

[30] S. Nikolaidis, J. Shah, Human-robot teaming using shared mental models, in: ACM/IEEE HRI, 2012.

[31] M. Hauskrecht, Value function approximations for partially observable Markov decision processes. J. Artif. Intell. Res. 13 (2000) 33–94, http://dx.doi.org/10.1613/jair.678.

[32] J. Bentham, The Principles of Morals and Legislation, Clarendon Press, Oxford, 1789. p. 1 (Chapter I).

[33] D. Bernoulli, Exposition of a new theory on the measurement of risk. Econometrica 22 (1) (1954) 22–36, http://dx.doi.org/10.2307/190982.

[34] J. von Neumann, O. Morgenstern, Theory of Games and Economic Behavior, Princeton University Press, Princeton, NJ, 1953.

[35] D. Kahneman, A. Tversky, Prospect theory: an analysis of decision under risk, Econometrica 0012-9682, 47 (2) (1979) 263, http://dx.doi.org/10.2307/1914185.

[36] M.H. Birnbaum, J.B. Navarrete, Testing descriptive utility theories: violations of stochastic dominance and cumulative independence, J. Risk Uncertain. 17 (1998) 49–78.
[37] A. Tversky, D. Kahneman, Advances in prospect theory: cumulative representation of uncertainty, J. Risk Uncertain. 5 (4) (1992) 297–323, http://dx.doi.org/10.1007/BF00122574.
[38] D.E. Bell, Regret in decision making under uncertainty, Oper. Res. 30 (5) (1982) 961–981.
[39] P.C. Fishburn, The Foundations of Expected Utility, Theory & Decision Library, Springer, New York, 1982.
[40] G. Loomes, R. Sugden, Regret theory: an alternative theory of rational choice under uncertainty, Econ. J. 92 (4) (1982) 805–824.
[41] J.R. Busemeyer, J.T. Townsend, Decision field theory: a dynamic cognition approach to decision making, Psychol. Rev. 100 (1993) 432–459.
[42] J.R. Busemeyer, A. Diederich, Survey of decision field theory, Math. Soc. Sci. 43 (3) (2002) 345–370.
[43] A.R. Wagner, A representation for interaction, in: Proceedings of the ICRA 2008 Workshop: Social Interaction with Intelligent Indoor Robots (SI3R), Pasadena, CA, USA, 2008.
[44] C. Camerer, G. Loewenstein, D. Prelec, Neuroeconomics: how neuroscience can inform economics, J. Econ. Lit. XLIII (2005) 9–64.
[45] J.A. Bargh, T.L. Chartrand, The unbearable automaticity of being, Am. Psychol. 54 (7) (1999) 462–479.
[46] R.M. Shiffrin, W. Schneider, Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory, Psychol. Rev. 84 (2) (1977) 127–190.
[47] J.E. LeDoux, The Emotional Brain: The Mysterious Underpinnings of Emotional Life, Simon and Schuster, New York, 1996.
[48] B. King-Casas, D. Tomlin, C. Anen, C.F. Camerer, S.R. Quartz, P.R. Montague, Getting to know you: reputation and trust in two-person economic exchange, Science 308 (2005) 78–83.
[49] J.S. Winston, B.A. Strange, J. O'Doherty, R.J. Dolan, Automatic and intentional brain responses during evaluation of trustworthiness of faces, Nat. Neurosci. 5 (2002) 277–283.
[50] F. Krueger, K. McCabe, J. Moll, N. Kriegeskorte, R. Zahn, Neural correlates of trust, Proc. Natl. Acad. Sci. 104 (2007) 20084–20089.
[51] D. Laibson, Neuroeconomics and the Multiple Systems Hypothesis, Mannheim Summer School, Mannheim, 2009.
[52] B. Shiv, A. Fedorikhin, Heart and mind in conflict: the interplay of affect and cognition in consumer decision making, J. Consum. Res. 26 (3) (1999) 278–292.
[53] J.K. Witt, S.A. Linkenauger, J.Z. Bakdash, D.R. Proffitt, Putting to a bigger hole: golf performance relates to perceived size, Psychon. Bull. Rev. 15 (2008) 581–585.
[54] J.K. Witt, M. Sugovic, Spiders appear to move faster than non-threatening objects regardless of one's ability to block them, Acta Psychol. 143 (2013) 284–291.
[55] P. Robinette, A. Wagner, A.M. Howard, The effect of robot performance on human robot trust, IEEE Trans. Hum. Mach. Syst., forthcoming.
[56] G. Trafton, L. Hiatt, A. Harrison, F. Tamborello, S. Khemlani, A. Schultz, Act-r/e: an embodied cognitive architecture for human-robot interaction, J. Hum.-Robot Interact. 2 (1) (2013) 30–55.
[57] M. Swangnetr, D.B. Kaber, Emotional state classification in patient–robot interaction using wavelet analysis and statistics-based feature selection, IEEE Trans. Hum. Mach. Syst. 43 (1) (2013) 63–75.

[58] L. Hiatt, A. Harrison, J. Trafton, Accommodating human variability in human-robot teams through theory of mind, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2011.

[59] A. Pentland, Socially aware computation and communication, IEEE Comput. 38 (3) (2005) 33–40.

[60] A. Pentland, Social dynamics: signals and behavior, in: International Conference on Developmental Learning, 2004.

[61] P. Ekman, W.V. Friesen, The repertoire of nonverbal behavior: categories, origins, usage, and coding, Semiotica 1 (1) (1969) 49–98.

[62] B. Lepri, N. Mana, A. Cappelletti, F. Pianesi, M. Zancanaro, Modeling the personality of participants during group interactions, in: User Modeling, Adaptation, and Personalization, Springer, Berlin, 2009, pp. 114–125.

[63] J.S. Uleman, Spontaneous versus intentional inferences in impression formation, in: S. Chaiken, Y. Trope (Eds.), Dual Process Theories in Social Psychology, Guilford Press, New York, 1999.

[64] Y.R. Tausczik, J.W. Pennebaker, The psychological meaning of words: LIWC and computerized text analysis methods, J. Lang. Soc. Psychol. 29 (1) (2010) 24–54.

[65] S.D. Gosling, S.J. Ko, T. Mannarelli, M.E. Morris, A room with a cue: personality judgments based on offices and bedrooms, J. Pers. Soc. Psychol. 82 (3) (2002) 379.

[66] D.A. Kenny, C. Horner, D.A. Kashy, L.C. Chu, Consensus at zero acquaintance: replication, behavioral cues, and stability, J. Pers. Soc. Psychol. 62 (1) (1992) 88.

[67] P. Ekman, E.L. Rosenberg, What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS), Oxford University Press, Oxford, 1997.

[68] R.W. Picard, Affective computing for HCI, in: HCI (1), 1999, pp. 829–833.

[69] A. Chakraborty, A. Konar, U.K. Chakraborty, A. Chatterjee, Emotion recognition from facial expressions and its control using fuzzy logic, IEEE Trans. Syst. Man Cybern. Syst. Hum. 39 (4) (2009) 726–743.

[70] T. Bänziger, D. Grandjean, K.R. Scherer, Emotion recognition from expressions in face, voice, and body: the Multimodal Emotion Recognition Test (MERT), Emotion 9 (5) (2009) 691–704.

[71] R.J. Sternberg, M.K. Gardner, A componential interpretation of the general factor in human intelligence, in: H.J. Eysenck (Ed.), A Model for Intelligence, Springer-Verlag, New York, 1982, pp. 231–254.

[72] P.L. Ackerman, E.D. Heggestad, Intelligence, personality, and interests: evidence for overlapping traits, Psychol. Bull. 121 (2) (1997) 219.

[73] K. Kleisner, V. Chvátalová, J. Flegr, Perceived intelligence is associated with measured intelligence in men but not women, PLoS One 9 (3) (2014) e81237.

[74] M.D. Prokosch, R.A. Yeo, G.F. Miller, Intelligence tests with higher g-loadings show higher correlations with body symmetry: evidence for a general fitness factor mediated by developmental stability, Intelligence 33 (2) (2005) 203–213.

[75] J. Chan, G. Nejat, Social intelligence for a robot engaging people in cognitive training activities, Int. J. Adv. Robot. Syst. 9 (2012) 1–13.

[76] X. Pan, C.S. Han, K. Dauber, K.H. Law, A multi-agent based framework for the simulation of human and social behaviors during emergency evacuations, AI Soc. 22 (2) (2007) 113–132.

[77] G. Proulx, Understanding human behaviour in stressful situations, in: Workshop to Identify Innovative Research Needs to Foster Improved Fire Safety in the United States, Delegate Binder Section 7, April 15–16, National Academy of Sciences, Washington, DC, 2002, pp. 1–5.

[78] S.T. Fiske, S.E. Taylor, Social Cognition, Addison-Wesley Pub. Co., Reading, MA, 1984.

[79] P. Ekman, W.V. Friesen, P. Ellsworth, What emotion categories or dimensions can observers judge from facial behaviour? in: P. Ekman (Ed.), Emotion in the Human Face, second ed., Cambridge University Press, Cambridge, 1982, pp. 98–110.
[80] I.J. Deary, M.C. Whiteman, J.M. Starr, L.J. Whalley, H.C. Fox, The impact of childhood intelligence on later life: following up the Scottish mental surveys of 1932 and 1947, J. Pers. Soc. Psychol. 86 (1) (2004) 130.
[81] A.R. Jensen, The G Factor: The Science of Mental Ability, Praeger, Wesport, 1998.
[82] M. Argyle, The Psychology of Interpersonal Behaviour, Penguin, Harmondsworth, 1967.
[83] J.R. Anderson, C. Lebiere, The Atomic Components of Thought, Erlbaum, Mahwah, NJ, 1998.
[84] A. Newell, Unified Theories of Cognition, Harvard University Press, Cambridge, MA, 1994.
[85] V.P. Richmond, J.C. McCroskey, S.K. Payne, Nonverbal Behavior in Interpersonal Relations, Prentice Hall, Englewood Cliffs, NJ, 1991.
[86] U.C.E. Segerstråle, P.E. Molnár, Nonverbal Communication: Where Nature Meets Culture, Lawrence Erlbaum Associates, Inc., Mahwah, NJ, 1997.
[87] M.L. Walters, K. Dautenhahn, K.L. Koay, C. Kaouri, S. N. Woods, C.L. Nehaniv, I. Werry, The influence of subjects' personality traits on predicting comfortable human-robot approach distances', in: Proceedings of Cog Sci 2005 Workshop: Toward Social Mechanisms of Android Science, 2005, pp. 29–37.
[88] S. Dar, Z. Khaki, Emotion recognition based on audio speech, IOSR J. Comput. Eng. 1 (2013) 46–50.
[89] L.C. De Silva, T. Miyasato, R. Nakatsu, Facial emotion recognition using multi-modal information, in: Information, Communications and Signal Processing, 1997, in: Proceedings of 1997 International Conference on ICICS, Vol. 1, IEEE, 1997, pp. 397–401.
[90] H.G. Gough, A nonintellectual intelligence test, J. Consult. Psychol. 17 (4) (1953) 242.