International Journal of Social Robotics (2019) 11:235–253
https://doi.org/10.1007/s12369-018-0496-1

Dating a Synthetic Character is Like Dating a Man

Johan F. Hoorn[1,2] · **Elly A. Konijn**[2] · **Matthijs A. Pontier**[2]

Accepted: 30 May 2018 / Published online: 31 October 2018

Abstract
To evaluate our emotionally intelligent software, we put a virtual human capable of speech and facial expressions to an updated and enriched version of the traditional Turing test. In a speed-date with 54 young females, either our software or human confederates controlled the simulation of the virtual human's affective performance. Results were obtained with frequentist analysis and Bayesian structural equation modeling. Indeed, participants did not detect differences and observed similarity in the emotional behavior of the virtual human and in the way it assumingly perceived them. Additionally, participants did not recognize different but similar cognitive-affective structures between humans and our system. As is, designers may use our software for believable affective virtual humans or robots. Moreover, as far as the richness of interaction possibilities in the speed-dating session allowed, our software seems to reproduce human cognitive-affective structures.

Keywords Cognitive models · Social agents · Affective computing · Turing test · Bayesian analysis

1 Introduction

Characters in science-fiction media can be most engaging. Humanoid robots such as Hal in 2001: A Space Odyssey (Kubrick, 1968), C3PO and R2D2 in Star Wars (Lukas, 1977), the Replicants in Blade Runner (Scott, 1982), The Terminator (Cameron, 1984), or Data in Star Trek: Generations (Carson, 1994) communicate with humans on an equal footing. They are empathic, social, and occasionally hostile but all of them have an understanding of human affect and they fascinate us: We feel involved and sometimes emotionally distant.

In a series of empirical studies, reviewed in [19], we investigated the process of engagement with movie and game characters, avatars and robots. Research into fictional characters can be quite informative for scientific questions [3] [6]. We dubbed the resultant model Interactively Perceiving and Experiencing Fictional Characters or I-PEFiC for short. The model describes the cognitive-affective structure of a person becoming involved with and/or feeling at a distance towards a virtual other. We then implemented I-PEFiC as an artificial intelligence (AI) system, Silicon Coppélia, capable of building up affect for its user [19].

I-PEFiC consists of three phases: encode, compare, and respond. In the *encode* phase, the user perceives a robot or other synthetic character in terms of 'a good or a bad guy' (factor *Ethics*), as 'pretty vs. unattractive' (factor *Aesthetics*), and 'realistic vs. unrealistic rendering' (factor *Epistemics*). The action possibilities, what can be done with the virtual character, also are encoded, namely as an aid to achieve user goals or as an obstacle (factor *Affordances*).

In the *comparison* phase, the user compares him/herself with the character to estimate how relevant or irrelevant the character's features from the encode phase are to the user's personal goals and concerns (factor *Relevance*). If someone is lonely, companionship of a robot is more important than calendar keeping. Users also build up positive and negative expectations about interacting with the character (factor *Valence*). If the user talks to a chatbot and expects it to have an understanding of the conversational context, the interaction will be disappointing (i.e. evoke negative valence). The final comparison is for (dis)similarity between user and character (factor *Similarity*). Particularly men dislike a character that performs poorly and looks like them [51].

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12369-018-0496-1) contains supplementary material, which is available to authorized users.

✉ Johan F. Hoorn
csjfhoorn@polyu.edu.hk

[1] Department of Computing and School of Design, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

[2] Department of Communication Science, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

In the *response* phase, the results of the comparison phase lead to the affective processes of *Involvement* with, and *Distance* toward the character, and parallel to that, to the emergence of certain *Use Intentions*: The user's willingness to interact with the character to achieve his or her goals. Together, *Involvement, Distance,* and *Use Intentions* integrate into one overall measure of *Satisfaction* with the character, comprising of both relationship and utility aspects.

As said, we translated the I-PEFiC model into the Silicon Coppélia system. In the current paper, the main aim is to test Silicon Coppélia's simulation of affect against real users. Regarding emotion and affect, will users be capable of telling the difference between our system and a real person interacting with them? Because our AI was not embodied, we had it drive a synthetic character named Tom. To make the comparison fair to the computer, in a second condition, Tom also was driven by human confederates. Together, the two conditions equaled the set-up of a Turing test [48].

Figure 1 provides an overview of this arrangement and forms a visual diagram of the argument put up in this paper. In that capacity, Fig. 1 will serve as the central point of reference of our contribution and as a 'reader's guide.' Next, we will discuss the implementation of I-PEFiC in Silicon Coppélia.

1.1 Silicon Coppélia

In [21, 39], we formalized the relations among the I-PEFIC factors in a computer model called Silicon Coppélia, in which the software agent builds up a relationship and estimates the utility of a user to the agent's pre-set goals. To make Coppélia autonomous, a loop accepts features of a particular situation as input, does situation selection, and outputs certain affective actions, leading to a new situation (Fig. 2). To do situation selection, the level of *Satisfaction* determines the affective decision-making such as 'changing the subject,' 'avoiding conflict,' or 'positive approach' [21].

To build up *Satisfaction*, Coppélia goes through the steps outlined by I-PEFiC. For the current speed-date, we merely used the *Ethics* and *Affordances* procedures because empirical research showed that these factors are more important than *Aesthetics* and *Epistemics* [52].

1.1.1 Encoding Phase

In the encoding of the *Ethics* of the user, the agent (A1) perceives that the user (A2) has a certain feature (e.g., 'kind to animals'). These are the data-driven aspects of the agent's affective processing. A1 calculates how 'good' A1 believes that being kind to animals is and how 'bad.' This is a perceived value between [0, 1]. The agent assigns the perceived value with a certain bias [21], representing the more concept-driven aspects of the agent's processing. Hence:

$$\text{Perceived}_{(\text{Good,A1,A2})} = \text{Bias}_{(\text{A1,A2,Good})} * \text{Data-driven}_{(\text{Good,A2})} \quad (1)$$

$$\text{Perceived}_{(\text{Bad,A1,A2})} = \text{Bias}_{(\text{A1,A2,Bad})} * \text{Data-driven}_{(\text{Bad,A2})} \quad (2)$$

Bias is assigned in the range [0, 2] and then multiplied with the value between [0, 1] for the Data-driven observation of the feature. When Bias = 1, A1 does not do underestimations or overestimations. If Bias > 1, A1 is positively biased towards A2. If due to large Bias, the result of formula (1) or (2) is greater than 1, it is reset to 1 to keep the formulae within range.

The agent also encodes *Affordances* from its expectations about the possibilities to communicate with the user. The expected utility of a user pertains to the agent's preset goals (e.g., 'to get a date'). For example, if the agent asks "Do you have many hobbies?" and the user replies "Well, that's none of your business," then possibilities for communication become fewer and so the level of *Affordances* decreases. Agent A1 calculates the *Affordances* of A2 in terms of aids (e.g., things that help communication) and obstacles (e.g., things that obstruct communication).

For the agent to estimate to what degree a user's *Affordances* will help achieve goals, it holds a number of beliefs: Goal-states are false or true [0, 1] and strongly inhibit [−1] or strongly facilitate [1] another goal-state or they are neutral [0]. From the goals achieved earlier in the interaction, the agent calculates the likelihood [−1, 1] that the user will help to achieve the next goal-state. The agent multiplies goal-states being false or true with beliefs of goal-states inhibiting or facilitating the next goal-state. With each facilitating sub goal that the agent achieves with its user, the perceived likelihood rises of achieving the desired end goal. For that, the agent uses likelihood algorithm Λ:

1. Sort values in the facilitation list [0 → 1] and in the inhibition list [0 → −1].
2. In each list, start at zero and average the first value with the next and so on until EOF.
3. Calculate the likelihood as the weighed mean of the output of the two lists, using proportion weights (#pos/#tot) for the list of positive values and (#neg/#tot) for the list of negative values.

Agents also believe that actions of the user affect world-states with strong inhibition [−1], neutrally [0], or with strong facilitation [1]. If an agent sees the user execute an action that influences a world-state, the agent holds the user responsible, using belief algorithm B:

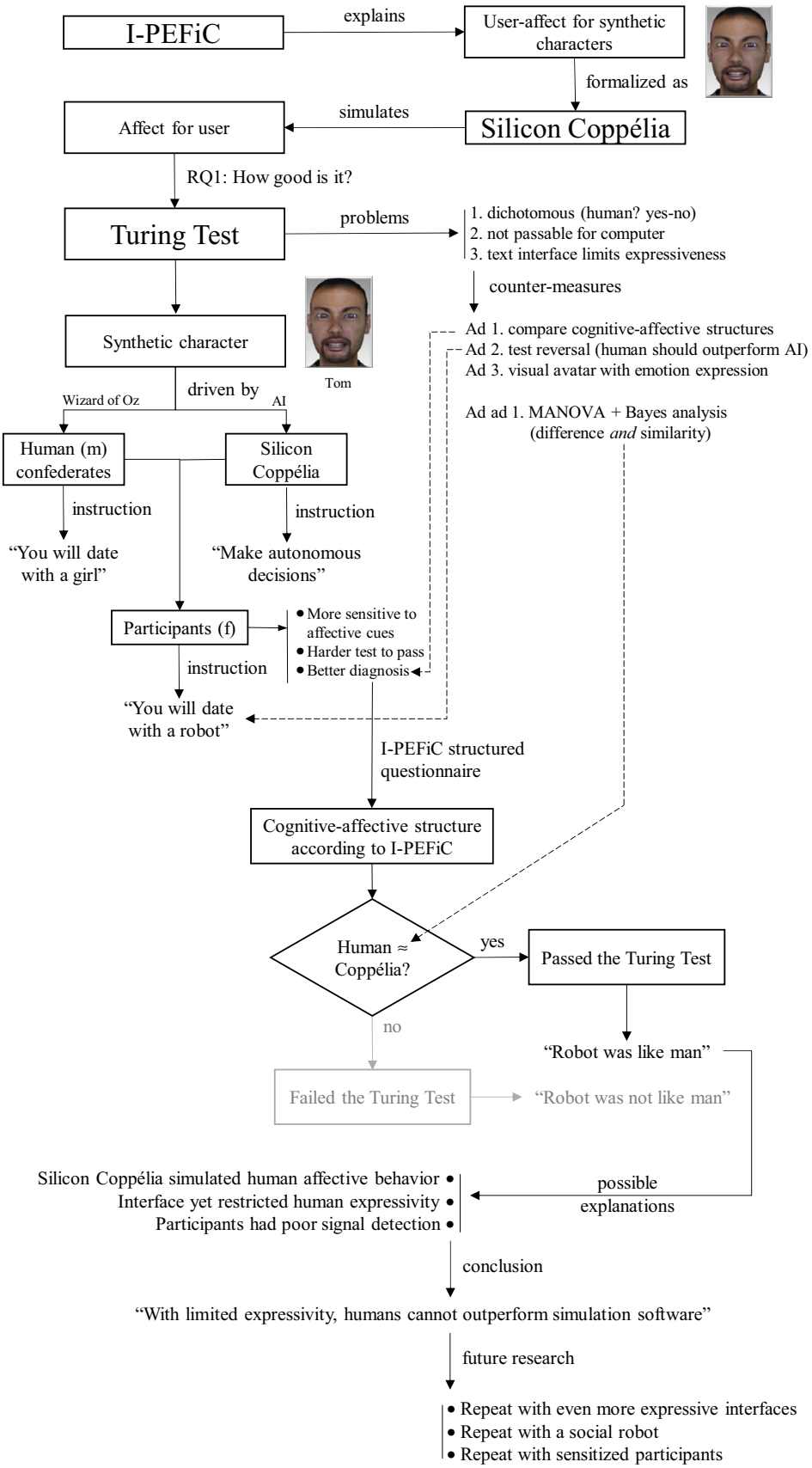


Fig. 1 Visual diagram of argument

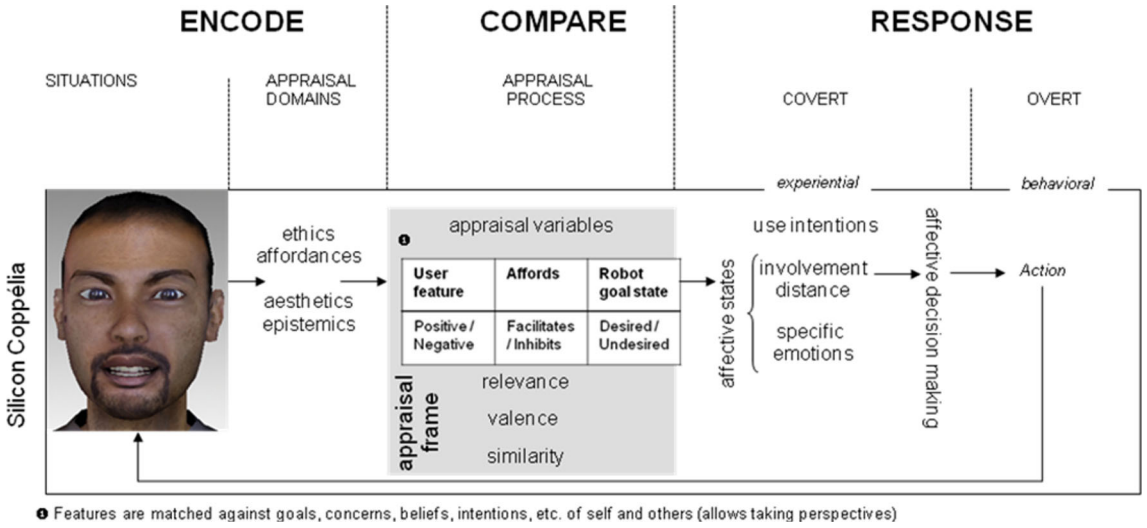


Fig. 2 Graphical representation of Silicon Coppélia [21]

IF observe(A1, A2, performs, action)
AND belief(action, facilitates, goal-state) > 0
THEN belief(A2, responsible, goal-state) = old_belief + mf_{bel_resp} * belief(action, facilitates, goal-state) * (1 – old_belief),

where modification factor $mf_{<variable>}$ controls the speed by which believed responsibility is updated, belief(action, facilitates, goal) is the impact value to multiply the modification factor with, and 1-old_belief is the limiter to avoid out-of-range values.

How much an agent aspires to achieve a goal, its ambition level, is designated by a real number [− 1, 1]. Positive values represent that the goal is desired, negative that it is undesired. The higher the value, the higher the desire. The agent calculates expected utilities of actions and features, taking facilitation and inhibition into account:

$$\begin{aligned}&\text{ExpectedUtility}_{(\text{Action,Feature,Goal})}\\&\quad=\text{Belief}_{(\text{facilitates(Action,Feature,Goal)})}\\&\qquad*\ \text{Ambition}_{(\text{Goal})}\end{aligned}\tag{3}$$

$$\begin{aligned}&\text{ExpectedUtility}_{(\text{Action,Agent,Goal})}\\&\quad=\text{Belief}_{(\text{facilitates(Action,Agent,Goal)})}\\&\qquad*\ \text{Ambition}_{(\text{Goal})}\end{aligned}\tag{4}$$

$$\begin{aligned}&\text{ExpectedUtility}_{(\text{Action,Agent})}\\&\quad=\Sigma\Big(\text{ExpectedUtility}_{(\text{Action,Agent,Goal})}\\&\qquad+\text{ExpectedUtility}_{(\text{Feature,Agent,Goal})}\Big)\end{aligned}\tag{5}$$

Actions and features that the agent believes facilitate desired or inhibit undesired goals affect *Expected Utility*. When they facilitate undesired or inhibit desired goals, *Expected Utility* drops. Based on its *Expected Utility*, the agent executes a user-directed action. *Expected Utility* is calculated by multiplying the 'believed facilitation of a goal by an action' with the 'level of ambition for that goal.' Because actions and features can build up several expected utilities for more goals, algorithm Λ calculates a *General Expected Utility* across the goals that are believed to be affected by that action or that feature. The general expected utilities of actions produce an agent's *Action Tendencies*, which have the same value.

Action Tendencies have a mix of positive and negative sides. An *Action Tendency* is multiplied with the positivity of an action, which leads to a measure of *General Positivity* that goes in list GP. Multiplied with the negativity of an action, it renders a measure of *General Negativity*, going into list GN. From these two lists, algorithm Λ then calculates the *General Positive Action Tendency* (GAT_{pos}) and the *General Negative Action Tendency* (GAT_{neg}) over all action tendencies in the agent.

1.1.2 Comparison Phase

In the *comparison* phase, the agent retrieves beliefs about actions that either facilitate or inhibit the desired or undesired goal-states. This is to calculate a general expected utility for each action. The agent also determines certain appraisal variables, such as the belief that someone is responsible for accomplishing certain goal-states or not. The features and variables as perceived by the agent are related to the goals and concerns it has stored (e.g., to get a date).

In the comparison phase, the agent system has knowledge of its own features and features it encoded from the user. It also has biases towards itself:

$$\text{Perceived}_{(\text{Feature,A1,A1})} = Bias_{(\text{A1,A1,Feature})} * \text{Data-driven}_{(\text{Feature,A1})} \quad (6)$$

The differences between the two feature sets determine the level of *Similarity* (similar vs dissimilar) that A1 perceives in A2, where the sum ranges over all encoded features:

$$\text{Similar}_{(\text{A1,A2})} = 1 - (\Sigma(\beta_{\text{sim}\leftarrow\text{feature}} * \text{abs}(\text{Perceived}_{(\text{Feature,A1,A2})} - \text{Perceived}_{(\text{Feature,A1,A1})})) \quad (7)$$

$$\text{Dissimilar}_{(\text{A1,A2})} = (\Sigma(\beta_{\text{dis}\leftarrow\text{feature}} * \text{abs}(\text{Perceived}_{(\text{Feature,A1,A2})} - \text{Perceived}_{(\text{Feature,A1,A1})})) \quad (8)$$

To determine the level of being dissimilar, the differences between the perceived values for A1's features and A2's features are added with a particular regression weight β. For the level of being similar, 1 minus the sum of all differences is taken, while using different weights. Note, however, that in the current application, the Similarity module was not active because it is empirically one of the least important factors [52].

Then the system estimates the *Relevance* (relevant-irrelevant) of the user's features to A1's goals as well as their *Valence* (i.e. positive–negative outcome expectancies). Because the formulas are quite elaborate but look alike, we provide the one for positive *Valence* regarding *Ethics*:

$$\begin{aligned}\text{Positive\_Valence\_Ethics}_{(A1,\,A2)} &= \beta_{\text{pv}\leftarrow\text{good}} * \text{Perceived}_{(\text{Good,A1,A2})} \\ &+ \beta_{\text{pv}\leftarrow\text{bad}} * \text{Perceived}_{(\text{Bad,A1,A2})} \\ &+ \beta_{\text{pv}\leftarrow\text{pos}} * (\text{GAT}_{\text{pos}} + 1)/2 \\ &+ \beta_{\text{pv}\leftarrow\text{neg}} * (\text{GAT}_{\text{neg}} + 1)/2 \end{aligned} \quad (9)$$

Note that by adding 1 and dividing the result by 2, GAT_{pos} en GAT_{neg} are transformed from [−1, 1] to the [0, 1] range.

In the comparison phase, the user's features come from *Ethics* and *Affordances*. The result of the comparison is a set of values for how similar, dissimilar, relevant, irrelevant, positive, and negative those user features are to the agent and the agent's preset goals. The values that the comparison phase outputs then go to the response phase.

1.1.3 Response Phase

The *response phase* consists of a measure of *Involvement* with the user, a measure of *Distance*, and *Intentions to Use* (i.e. interact with) the user at a next occasion. *Involvement* is calculated as:

$$\begin{aligned}\text{Involvement}_{(\text{A1,A2})} &= \beta_{\text{inv}\leftarrow\text{good}} * \text{Perceived}_{(\text{Good,A1,A2})} \\ &+ \beta_{\text{inv}\leftarrow\text{bad}} * \text{Perceived}_{(\text{Bad,A1,A2})} \\ &+ \beta_{\text{inv}\leftarrow\text{aid}} * \text{Perceived}_{(\text{Aid,A1,A2})} \\ &+ \beta_{\text{inv}\leftarrow\text{obst}} * \text{Perceived}_{(\text{Obstacle,A1,A2})} \\ &+ \beta_{\text{inv}\leftarrow\text{pv}} * \text{Pos\_Valence}_{(\text{A1,A2})} \\ &+ \beta_{\text{inv}\leftarrow\text{nv}} * \text{Neg\_Valence}_{(\text{A1,A2})} \\ &+ \beta_{\text{inv}\leftarrow\text{ps}} * \text{Pos\_Valence}_{(\text{A1,A2})} * \text{Similarity}_{(\text{A1,A2})} \\ &+ \beta_{\text{inv}\leftarrow\text{ns}} * \text{Neg\_Valence}_{(\text{A1,A2})} * \text{Similarity}_{(\text{A1,A2})} \\ &+ \beta_{\text{inv}\leftarrow\text{pd}} * \text{Pos\_Valence}_{(\text{A1,A2})} * \text{Dissimilarity}_{(\text{A1,A2})} \\ &+ \beta_{\text{inv}\leftarrow\text{nd}} * \text{Neg\_Valence}_{(\text{A1,A2})} * \text{Dissimilarity}_{(\text{A1,A2})} \\ &+ \beta_{\text{inv}\leftarrow\text{pb}} * \text{Pos\_Valence}_{(A1,\,A2)} * \text{Perceived}_{(\text{Good,A1,A2})} \\ &+ \beta_{\text{inv}\leftarrow\text{nb}} * \text{Neg\_Valence}_{(\text{A1,A2})} * \text{Perceived}_{(\text{Good,A1,A2})} \\ &+ \beta_{\text{inv}\leftarrow\text{pu}} * \text{Pos\_Valence}_{(\text{A1,A2})} * \text{Perceived}_{(\text{Bad, A1,A2})} \\ &+ \beta_{\text{inv}\leftarrow\text{nu}} * \text{Neg\_Valence}_{(\text{A1,A2})} * \text{Perceived}_{(\text{Bad,A1,A2})} \\ &+ \beta_{\text{inv}\leftarrow\text{rel}} * \text{Relevance}_{(\text{A1,A2})} \\ &+ \beta_{\text{inv}\leftarrow\text{irr}} * \text{Irrelevance}_{(\text{A1,A2})} \\ &+ \beta_{\text{inv}\leftarrow\text{rs}} * \text{Relevance}_{(\text{A1,\,A2})} * \text{Similarity}_{(\text{A1,A2})} \\ &+ \beta_{\text{inv}\leftarrow\text{is}} * \text{Irrelevance}_{(\text{A1,A2})} * \text{Similarity}_{(\text{A1,A2})} \\ &+ \beta_{\text{inv}\leftarrow\text{rd}} * \text{Relevance}_{(\text{A1,A2})} * \text{Dissimilarity}_{(\text{A1,A2})} \\ &+ \beta_{\text{inv}\leftarrow\text{id}} * \text{Irrelevance}_{(\text{A1,A2})} * \text{Dissimilarity}_{(\text{A1,A2})} \\ &+ \beta_{\text{inv}\leftarrow\text{rb}} * \text{Relevance}_{(\text{A1,A2})} * \text{Perceived}_{(\text{Good,A1,A2})} \\ &+ \beta_{\text{inv}\leftarrow\text{ib}} * \text{Irrelevance}_{(\text{A1,A2})} * \text{Perceived}_{(\text{Good,A1,A2})} \\ &+ \beta_{\text{inv}\leftarrow\text{ru}} * \text{Relevance}_{(\text{A1,A2})} * \text{Perceived}_{(\text{Bad,A1,A2})} \\ &+ \beta_{\text{inv}\leftarrow\text{iu}} * \text{Irrelevance}_{(\text{A1,A2})} * \text{Perceived}_{(\text{Bad,A1,A2})} \end{aligned} \quad (10)$$

Distance is calculated in the same way as *Involvement. Use Intentions* are calculated with algorithm Λ based on the expected utilities of all user features and the actions the agent can execute. Because a feature can contribute to both *Involvement* and *Distance* [18], the agent calculates a fuzzy trade-off, following [59] (p. 398):

$$\begin{aligned}\mu\ and\tilde{}\big(\mu_{\tilde{I}}(u), \mu_{\tilde{D}}(u)\big) &= \gamma \cdot min\{\mu_{\tilde{I}}(u), \mu_{\tilde{D}}(u)\} \\ &+ ((1-\gamma)(\mu_{\tilde{I}}(u) + \mu_{\tilde{D}}(u))/2),\end{aligned} \quad (11)$$

where each feature $u \in U$ in the trade-off has a membership function μ in fuzzy *Involvement* ($\tilde{I}$) and fuzzy *Distance* ($\tilde{D}$) and settles between the minimum and maximum degree of membership of these sets. The γ-operator is a sort-of-AND and may have different weights and different levels of compensation [0, 1]. In (11), the number of fuzzy sets is 2.

Finally, the agent runs an affective decision-making module to calculate the *Expected Satisfaction* of possible actions. When the agent selects and performs an action, a new situation emerges, and the model loops back to the first phase (Fig. 2).

Hoorn, Pontier, and Siddiqui [20] describe that the agent now works from a weighed mean of the *Involvement-Distance* trade-off on the one hand and the Λ-calculated *Use Intentions* on the other. The agent selects an action through:

$$\begin{aligned}&\text{ExpectedSatisfaction}_{(\text{A1,Action},A2)}\\&\quad = w_{eu} * \text{Action\_Tendency}\\&\qquad + w_{pos} * (1 - \text{abs}(\text{positivity} - \text{bias}_I * \text{Involvement}))\\&\qquad + w_{neg} * (1 - \text{abs}(\text{negativity} - \text{bias}_D * \text{Distance}))\end{aligned} \quad (12)$$

The agent picks the strongest action tendency while the positive side tends to the level of (biased) *Involvement* and the negative side to (biased) *Distance*. By changing weights, differences in positivity, negativity, and expected utility lead to selecting different actions. This way the agent has a form of response modulation, allowing for biases in individual defaults (e.g., optimistic or pessimistic).

Note that both overt (behavioral) and covert (experiential) responses can be executed in the response phase. Emotions such as hope, joy, and anger are generated using appraisal variables (e.g., the perceived responsibility of the interaction partner, and likelihood of achievement of goal-states in accordance with emotion theory [12]).

1.1.4 Emotion Generation

Silicon Coppélia can generate *hope* and *fear*, which originate from desired or undesired world-states, respectively. The agent system estimates hope for goal accomplishment in world-states that are likely to occur:

$$\begin{aligned}&\text{IF } f \geq \text{likelihood} \rightarrow \text{hope\_for\_goal} =\\&-0.25 * (\cos(1/f * \pi * \text{likelihood(goal)}) - 1.5)\\&\quad * \text{ambition\_level(goal-state)}\\&\text{IF f} < \text{likelihood} \rightarrow \text{hope\_for\_goal} =\\&-0.25 * (\cos(1/(1-f) * \pi * (1 - \text{likelihood(goal)})))\\&\quad - 1.5) * \text{ambition\_level(goal-state)},\end{aligned} \quad (13)$$

where f is a shaping parameter [0, 1]. If f is close to 0, the agent hopes against the odds. Algorithm Λ processes the values of hope_for_goal, where the positive-values list produces 'hope' and the negative list produces 'fear.' Algorithm B determines *joy* and *sadness* after world-states become true of false with ambition_level(world-state) or -ambition_level(world-state) as impact value. A world-state that is desired and becomes true raises joy; an undesired world-state that is true causes sadness. Additionally, a level of *surprise* is generated that tends to disbelieve the likelihood that such world-state would happen:

$$\begin{aligned}\text{Surprise} = {} & p_{surp} * \text{old\_surprise} + \left(1 - p_{surp}\right)\\&* (1 - \text{likelihood})\end{aligned} \quad (14)$$

If an expected goal-state is not achieved, surprise increases according to algorithm B with likelihood(goal-state) as the impact value. Yet, persistency factor p_{surp} regulates the speed by which surprise decays again, which can be set at each time step.

Algorithm B also controls the *anger* [0, 1] an agent feels when the user obstructs the achievement of desired goal-states with belief (A2, responsible, goal-state) * ambition_level(A1, goal-state) as impact value. Like surprise, decay of anger follows from (14). The general anger with the user follows from all levels of anger felt during the interaction, according to algorithm Λ (which stops after step 2). The same applied to agent A1 self decides the level of feeling *guilty*.

From the seven simulated emotions, an *overall mood* of the agent follows by taking 1 minus the weighed sum of the differences between the desired level and the current level of emotion for all generated emotions:

$$\text{Mood} = 1 - (\Sigma(\beta_{emotion} * \text{abs}(\text{Emotion} - \text{desired(Emotion)}) \quad (15)$$

1.1.5 Emotion Regulation

While using algorithm B with (Emotion(timepoint) − Emotion(timepoint − 1)) * Attention(Feature) as impact value, agents believe that a feature Feat causes an emotion E. This way, the agent may shift attention Att to regulate its emotions:

$$\begin{aligned}\text{Att(Feat)} = {} & \text{old\_value} - \text{belief(Feat, causes, E)}\\&* (\text{E} - \text{desired}(E))\end{aligned} \quad (16)$$

Attentional shifts may occur at each time point, owing to changes in the relevance of features. In taking the absolute value of the *General Expected Utility* of a feature:

$$\text{Att(Feat)} = p_{att} * \text{old\_value} + \left(1 - p_{att}\right) * \text{Relevance(Feat)} \quad (17)$$

where p_{att} is a persistency factor, controlling the pace of the attentional shift with the sum of the levels of attention normalized to 1 at each point in time.

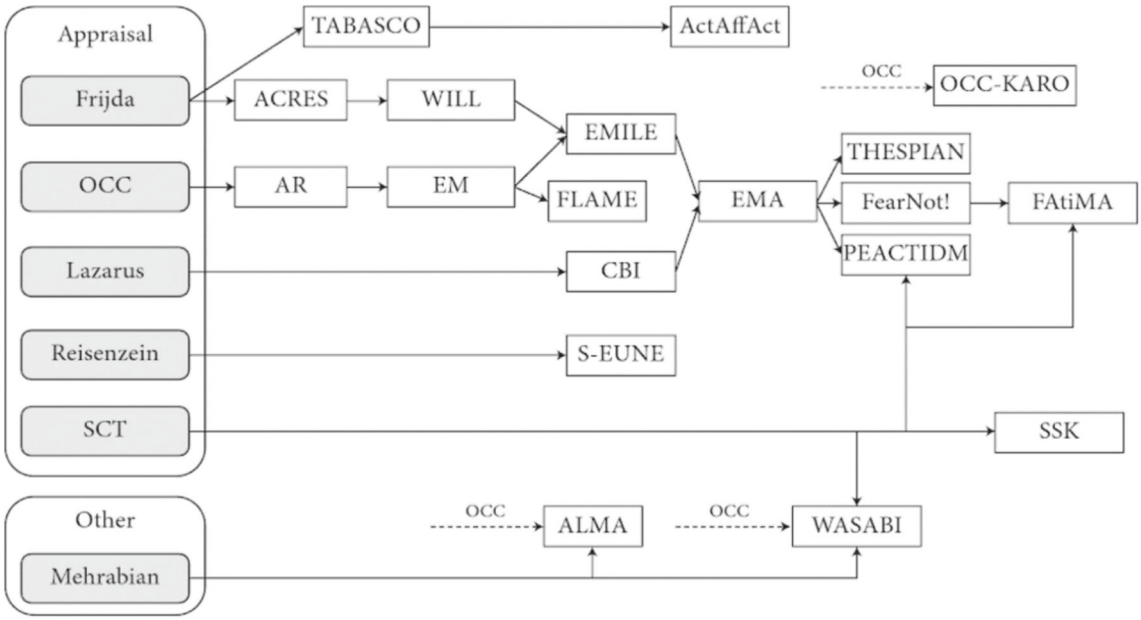


Fig. 3 Family tree of appraisal models [15] (p. 60). Silicon Coppélia is partly based on EMA

1.2 Related Work

Theoretically, I-PEFiC and its implementation in Silicon Coppélia stand in a lineage of psychological appraisal theories, particularly [12] and [45]. In the taxonomy of [15] (p. 60), Silicon Coppélia partly follows from Emotion and Adaptation (EMA) (Fig. 3). Particularly the epistemic side of the appraisal process such as the agent's beliefs, the likelihood of events, and expectations of occurrences in Silicon Coppélia resemble EMA.

EMA constantly updates the appraisal values connected to observations of the world-state. Such appraisals pertain to desirability (i.e. facilitation or inhibition of utility of an event or other agent), likelihood that a world-state becomes active, and expectedness (whether a world-state can be causally predicted).

Next, we discuss four related models that like Silicon Coppélia are indebted to EMA. For the less-related models listed in Fig. 3, we refer to the authoritative review of [15].

THESPIAN (Fig. 3) operates in story-telling environments in which agent and user interact [44]. THESPIAN has decision theory installed to infer agent or user goals and while doing belief maintenance and revision, the system builds up a representation (including emotion) of a character in its environment, using probabilistic preferences and weights.

In the FearNot! system [2], the agent can autonomously select actions based on the appraisal of the importance of an event, activating the applicable emotions. If a world-state becomes true, the agent then pursues goals that become active. Emotions related to the new goal are fed back into action selection, which could be an action tendency (e.g., approach, avoid) or coping behavior. By means of emotion regulation, an agent frustrated in achieving a goal state may do reappraisal of the importance of that goal so to mitigate its stress levels.

PEACTIDM calculates appraisal information directly from its processes, producing appraisals and their accompanying emotions as a side effect [36]. Appraisals may vary in how they are produced but also in the values of the various appraisal dimensions. For instance, goal relevance, discrepancy from expectation, and outcome probability have ranges [0, 1] whereas a causal agent (self, other, nature) is categorical. In following EMA, PEACTIDM loads the current appraisals of the agent into an appraisal frame, which need not be complete. There may be one (incomplete) frame for each event or other agent but the agent uses the frames for attention selection (who to turn to). When active, new appraisals may happen such as the level of facilitation of the event for goal achievement.

Similar to other appraisal models, FAtiMA stores appraisals in an appraisal frame [10]. FAtiMA updates its memory and appraisal frames while checking for new events. If something new happens, FAtiMA creates a new appraisal frame for that event, triggering the appraisal process. Then FAtiMA assesses the relevance of the event and appraisal variables such as desirability of the event become active from which corresponding emotions follow. This makes up the 'emotional state' the agent is in. Based on the current

emotional state, FAtiMA may decide to react with an action tendency or to plan a strategy towards a desired goal state.

If implemented in a robot as is, Silicon Coppélia does not have a detection system that analyzes the user's affective state [8] like, for instance, Kismet does [7]. It uses simple textual input for that. Coppélia is a system of emotional reasoning and action selection, producing a response that is affectively optimal. Coppélia does not have a special emotional expression generator either. In this study, we simply connect its output to smiles, frowns, or grins. In that sense, Coppélia is in need of an iGrace system that senses emotion from the user's language and then selects and performs an affective action based on a database of emotional experiences [40] (iGrace was implemented in a cuddle robot for hospitalized children [40]). Yet, iGrace is in need of a system such as Silicon Coppélia, because it does not have an architecture in place that does cognitive-affective appraisals.

The appraisal process is central to the Silicon Coppélia software. The GAT procedures together with action selection resembles the approach of [33], who developed a predictive model of the robot self and its surroundings so to anticipate what certain affective feedback to the user may do. This system was tried in an iCat robot, playing a game of chess [30]. If iCat perceived a mismatch between expectations about the user and actual user behavior, it would generate an affective response in line with positive or negative valence [30]. In a later study [31], this system was used to show empathy to one user (e.g., an ally) and not to the other (e.g., the opponent).

In [13] and [35] as well, moods of the robot build up in anticipation of an event and fade over time after a mood was enforced or changed (cf. our approach to anger). These studies show that emotion simulation systems should drive away from a fixed set of affective behaviors and become adaptive to the way users behave socially and emotionally. If a robot could automatically adapt its mood to the user's mood, this would make the robot more empathetic and users more helpful towards the robot [14]. To that end, we should know how well a robot simulates human affective responses, which we will try in an adaptation of the Turing test.

1.3 Turing Test Revised

With Silicon Coppélia in place, we wished to test it against human affective behavior in a Turing test (Fig. 1). However, the Turing test is not without controversies [38]. How to conduct that test has changed over time (e.g., [23]). The literature reports a number of problems. First, the Turing test is a pass/fail test, which limits its testing abilities [17]. One can only test for either full success or failure. If a computer fails to seem human in 10% of its responses, it most likely fails the Turing test. Nuance is lost and the test becomes unpassable. Second, the instruction focuses participants on cues that provide evidence for either 'human' or 'machine,' ignoring other aspects of interaction and communication. Third, participants usually work with a text-based interface. This results in a narrow test focus, which prevents the achievement of an appropriate research goal [38]. A relatively unsophisticated program such as ELIZA [58] already seemed quite intelligent to regular observers.

Because in its original form, the Turing test was dichotomous, supposedly unpassable, and text-based (Fig. 1), some suggested giving up the Turing test all together as a meaningful idea [43]. However, lack of consensus does not have to keep the Turing test from being useful.

We countered the methodological problems as follows: To avoid simple pass-fail decisions, we had participants diagnose for us the 'mental state' of their communication partner Tom, whether human-driven or by our AI (Fig. 1, Ad 1). We did this using a structured questionnaire that queried the cognitive-affective dimensions of the I-PEFiC model.

To make the test passable for a computer, we reversed it (Fig. 1, Ad 2): Participants did not think they were talking to a human that turned out to be a machine. We told them upfront they were conversing with a robot during a speed-date so that it was the human who should outperform our software system.

To move beyond text interfaces, we followed Barberi [4] in that text should be accompanied by more sensual stimulations such as facial expressions, vocalization, and the expression of affect (e.g., joy, fear, hope, anger) (Fig. 1, Ad 3). Further elaboration of our approach can be found under Methods.

In this paper, then, we present a new variant of the Turing test with which we examined our research question (RQ1) whether individuals could detect that affective communication during a speed-date was produced by our AI or by a human being.

1.4 Research Question and Hypotheses

RQ1 (Fig. 1): When Tom is driven by our software, will users diagnose a cognitive-affective structure in Tom that is similar to when he is driven by a human being? During a speed-date session, the virtual human Tom was controlled by either the Silicon Coppélia system (AI) or by a human confederate. In doing so, we created an enriched Turing test that contrasted an autonomously communicating computer system with a human communicating via a computer, respectively. We predicted that participants interacting with Tom would not be able to differentiate between Tom as controlled by our software versus Tom controlled by a human (H1). Furthermore, we predicted that participants did not recognize significant differences in the cognitive-affective structure underlying the behavior produced by our software and that underlying the behavior of human confederates (H2).

2 Method

2.1 Participants and Design

Fifty-four female heterosexual university students ($M = 20.07$, $SD = 1.88$) volunteered for course credits or a small financial reward and were uninformed about the actual background and conditions of the study. Participants communicated frequently via a computer ($M = 4.02$, $SD = 1.00$; 0 = totally disagree; 5 = totally agree), however, they had limited to no experience in online dating ($M = .33$, $SD = .80$). We chose to confront female participants with a male virtual human, because women are usually better equipped to do an emotional assessment of others than men [5, 47].

The participants were randomly assigned to two experimental conditions ($n = 27$ each): Tom as an agent controlled by a computer using the software Silicon Coppélia (the AI condition) versus Tom as an avatar controlled by a human confederate (the Wizard of Oz (WOz) condition). This confederate could control the responses of Tom as well as Tom's facial expressions. To control for idiosyncrasies of the human confederate, we assigned two confederates to the WOz condition.[1] The confederates were male and instructed to behave as they normally would. They were blind for the study's purpose, yet trained to handle the situation [27]. For the main analysis, we could collapse the data of the two human confederates into one factor of Human Confederate.

Note that all participants were told they were interacting with a computer agent, irrespective of who controlled it. This way, participants understood the limitations of their dating partner, thus avoiding rejection solely based on his limited interaction skills.

We applied a between-subjects design to the experiment to not overstrain participants and for simplicity of methods. If participants should do the test in both conditions (WOz and AI), they should fill out the questionnaire twice with high likelihood that they become bored, repeat answers, and that we measure test fatigue. The other problem is the order of conditions. In a within-experimental design, half the participants should do condition 1 before 2 and the other half 2 before 1 to avoid halo effects. This would add another level of complexity to the administration and execution of the experiment and would add an extra non-theoretical factor (i.e. condition order) to the analysis.

[1] We ran a separate analysis to test for possible differences between the two human confederates. No significant differences were found for the main variables of interest (i.e., not for each of the five emotions, not for each of the eight perceptions, nor for each of the two decision-making behaviors).

2.2 Materials and Procedure

Upon arrival in our laboratory, participants took place behind a computer. They were instructed to do a speed-date session with an agent. In the WOz-condition, the human confederate controlling the avatar was hidden behind a wall (i.e., invisible for the participants). In the speed-date industry, sessions typically take 4 min after which the candidates decide yes or no about a date [11]. In our application, participants had 2.5 times that time to get to know each other. After finishing a 10-minute speed-date, participants completed a digital questionnaire. After the experiment, participants were debriefed and dismissed.

Participants dated with avatar Tom, who was either controlled by our software (i.e., an agent system) or controlled by one of two human confederates. The participants served as interrogators and attempted to diagnose the cognitive-affective structure that was responsible for Tom's behavior. They did this along the lines of I-PEFiC, an empirically established model of affective responses to virtual others [52]. Additional to questions about how the participants perceived the behavior of Tom, we also asked them how they thought Tom perceived them. This approach resembles general speed-dating settings, where people want to know what their dating partner thinks of them. The interrogators thus assessed the emotional behavior of the virtual human, as well as the things that the virtual human perceived and that led him to that emotional behavior. This way, we could check the differences between our software and human confederates in producing emotional behavior. In focusing on the relations between the things that Tom perceived (according to the participants), we could also analyze the differences in the cognitive-affective structure (as perceived by the participants) in our virtual human versus living confederates.

Whereas we enriched our environment with vocalization of responses and nonverbal cues, the interactions between Tom and the interrogator were limited to selecting multiple-choice responses (± 4 options per response). This way, effects could be attributed to the nonverbal cues, vocalizations, and emotional expressions that fostered impression formation (cf. [57]) rather than to the differences in handling the user interface better (human confederates) or worse (software). We hypothesized that differences in cognitive-affective structure and emotion expression would be absent between the communications of Tom as a human-controlled avatar versus Tom as a software-controlled agent.

We designed a speed-date application, in which users met the virtual human Tom, to get acquainted, and possibly arrange a next meeting. Tom was represented by a virtual human created in Haptek's PeoplePutty software. To prevent eeriness caused by the phenomenon known as the *uncanny valley* [34], Tom's face followed the design principles for computer-graphics animators as given by MacDorman et al.

Fig. 4 The speed-date application

[32]. As can be seen in Fig. 4, Tom's facial proportions are human-like, but Tom does not approach photorealistic perfection.

Tom was capable of simulating five emotions: hope, fear, joy, sadness, and anger, expressed through facial expressions with low or high intensity. As such, 32 (2^5) different emotional states were created: one for each possible combination of two levels of intensity of the five simulated emotions. The level of intensity depended on the level of relevance of participants' responses to Tom's goals and concerns.

Human confederates were trained to handle the interface that steered Tom. From a dropdown box they selected their replies, a (to them) appropriate emotion, and an intensity (high, low), which Tom then performed.

We used JavaScript in combination with scripting commands provided by the Haptek software to control Tom's behavior within a Web browser. In the middle of the Website, Tom was prominently present and communicated messages through a speech synthesizer (e.g., "Do you have many hobbies?"). This text could also be read at the top of the screen. The participant's text was displayed at the bottom.

During the speed-date session, participants conversed about seven topics: (1) Family, (2) Sports, (3) Appearance, (4) Hobbies, (5) Music, (6) Food, and (7) Relationships. Because we focused on emotional communication and affective decision-making, and not on the production and interpretation of natural language (which in current systems is poor), we limited the communication to multiple-choice questions and answers. This procedure also increased the comparability between the two conditions and the various interaction partners. For each topic, the dating partners went through an interaction tree with responses they could select from a dropdown box (Fig. 5). When a topic was done, the participant could select a new topic or let Tom select one. When all topics were completed, a message appeared: "The speed-date session is over" followed by an introduction to the questionnaire.

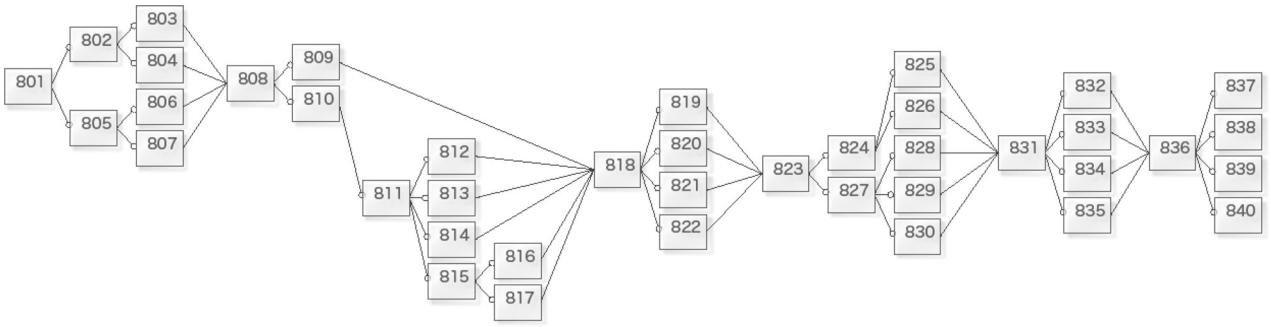
In the AI condition, Tom was designed with certain 'beliefs' that specific features of a participant could affect certain goal-states: 'get a date', 'be honest', and 'connecting well' on each of the conversation topics. Furthermore, Tom 'attached' a general level of positivity and negativity (both between 0 and 1) toward the user to each response. For example, Tom attached a positivity level of .2, and a negativity level of .9 to the response "To be honest, I don't think that this date is going well." We made sure that Tom (whether AI or WOz) and the participant could always pick from responses with various levels of positivity and negativity, to ensure enough degrees of freedom. Below a set boundary, Tom facially expressed low-intensity emotion. If greater than or equaling the boundary, Tom expressed high-intensity emotion.

During the speed-date, Tom's 'perceptions' of the participant were continuously updated based on the participant's responses during the session. Thus, Tom 'assessed' the participant's *Ethics* and *Affordance* values while matching these appraisals with his goals, reckoning with the *Involvement* and *Distance* he felt toward his interaction partner, and the utility he expected of each action. On each turn, Tom could select his response from a number of options. To achieve Tom's goals, he was equipped with search strings to choose actions with positivity levels that came closest to the level of *Involvement* and negativity levels closest to the level of *Distance*, as well as achieving the highest *Expected Utility*.

Tom 'expressed' a variety of emotions depending on the simulated emotional state. We used 5 out of 7 emotions we could simulate. Hope and fear were calculated in response to each answer according to the perceived likelihood of a follow-up date. Levels of Tom's joy and sadness were based on achieving desired or undesired goal-states. Tom's anger was calculated using the assumed responsibility of the participant for the success or failure of the speed-date. Details on how this was calculated are in the section on Silicon Coppélia and in [39]. Each of the five emotions were implemented into the software and simulated in parallel.

2.3 Measures

The questionnaire consisted of 97 Likert-type items, each followed by 6-point rating scales (0 = totally disagree; 5 = totally agree). Each item that was related to theoretical constructs queried what the participant thought Tom felt about her. The complete questionnaire can be found in part A of "Online Resource 1" and covered 15 measurement scales as described below per construct. A scale analysis was performed, in which items were removed until an optimal Cronbach's alpha was found with a minimum scale length of three items. To check divergent validity, we also performed a



801. (A) Do you easily fall in love?
802. (P) Yes, very often.
803. (A) Hehe, yes, me too! (distance/involvement)
804. (A) Oh? I don't really fall in love that easily myself (distance/involvement)
805. (P) No, not really.
806. (A) Hehe, me neither. (distance/involvement)
807. (A) Oh, I fall in love very easily (distance/involvement)
808. (A) Have you ever been in a relationship before?
809. (P) No, never.
810. (P) Yes, I have.
811. (A) Have you ever cheated on your partner?
812. (P) Yes
813. (P) Yes, but I regret it
814. (P) No, never
815. (P) I'd rather not talk about that
816. (A) That's too bad, but I understand. What would you rather talk about? (involvement/perspective).
817. (A) Oh, well that's too bad. I guess we'll talk about something else then... (negative emotion)
818. (A) Do you plan to marry eventually?
819. (P) Yes, I would love to!
820. (P) Yes, but not for the foreseeable future!
821. (P) No, probably not
822. (P) No, absolutely not!
823. (A) Are you in favor of long-term relationships, or do you like your freedom?
824. (P) I love my freedom!
825. (A) yes, I completely agree (distance/involvement)
826. (A) Oh, I myself like some stability (distance/involvement)
827. (P) I think a long-term relationship is very important! (distance/involvement)
828. (A) Me too! (distance/involvement)
829. (A) Oh, I like my freedom! (distance/involvement)
830. (A) Let's move on (negative emotion)
831. (A) Are you trustworthy?
832. (P) Yes, of course!
833. (P) I try my best; usually I am.
834. (P) I try my best, but I often fail.
835. (P) No, not really.
836. (A) Do you believe in 'the one'?
837. (P) Yes, my prince is walking this earth somewhere!
838. (P) Yes, but maybe I'm wrong.
839. (P) No, but it would be nice...
840. (P) No, sentimental nonsense

Fig. 5 Interaction tree for topic 'Relationships.' A = Agent, P = Participant

factor analysis (see "Online Resource 1", part B). The results of the scale analysis are in Table 1.

Emotions were measured by five measurement scales (after [54]), each indicating one of five emotions: Joy, Anger, Hope, Fear, and Sadness, assessing the emotions perceived in Tom by the participants. Example items are: 'Tom was happy', 'Tom was afraid', and 'Tom was irritated'.

Perceptions were measured by eight measurement scales, each indicating one of eight perceptions according to [52]: Ethics, Affordances, Similarity, Relevance, Valence, Involvement, Distance, and Use Intentions. Previous studies (e.g., [24]) showed that the Ethics scale is consistently reliable. Therefore, we decided to maintain the Ethics scale despite its relatively weak measurement quality in this study (Table 1). Example items for the Ethics scale are: 'Tom thought I was trustworthy', 'Tom thought I was mean', and 'Tom thought I was evil'.

Decision-Making Behavior consisted of two scales: Situation Selection and Affective-Decision Making. Situation Selection measured the perceived decision-making behavior of Tom in terms of concrete actions, through a scale with

Table 1 Number of items and Cronbach's alphas after scale and factor analysis

Scale	# Items	Cronbach
Joy	5	.93
Sadness	4	.84
Anger	4	.84
Hope	3	.76
Fear	3	.81
Sadness	4	.84
Ethics	3	.61
Affordances	3	.88
Similarity	3	.66
Relevance	5	.93
Valence	8	.93
Use intentions	9	.95
Involvement	4	.75
Distance	3	.82
Situation selection	3	.84
Affective decision-making	3	.74

items such as "Tom kept on talking about the same thing." Affective Decision-Making measured the amount of affective influences in the decision-making of Tom, through a scale with items such as "Tom followed his intuition."

2.4 Analysis

We tested the effects of the between-factor Control-type (AI versus WOz) on each of the dependent variables *Emotions* (Hope, Fear, Joy, Sadness, Anger), *Perceptions* (Ethics, Affordances, Similarity, Relevance, Valence, Involvement, Distance, Use Intentions) and *Decision-Making Behavior* (Situation Selection, Affective Decision-Making) using three multivariate analyses of variance (MANOVAs).

We also performed a one-sample t test with 0 as the test value to test whether the participants saw emotions in Tom at all. Additionally, we performed separate paired t-tests for all possible pairs within a group of related variables to investigate which was strongest.

Structural equation modeling analyses with Amos 16.0 [1] explored the underlying associations among I-PEFiC variables as found in [52]. With I-PEFiC as the benchmark, we compared the two cognitive-affective structures underlying evaluations of Tom in each condition and tested possible differences and similarities.

To compare the regression coefficients for the AI group and the WOz group, we performed a Multiple Group Analysis. In the 'unrestricted loadings' model, the regression coefficients had no restrictions. In the 'equal loadings' model, the regression coefficients of the AI group and the WOz group were assumed equal.

We used Bayesian estimation to analyze our data. Advantages of Bayesian statistics over frequentist statistics are well documented in statistics (e.g., [26, 49]). For example, Bayesian statistics are less dependent on a large sample size. Furthermore, Central Credibility Intervals (CCI), the Bayesian equivalent to Confidence Intervals (CI), actually is the probability that a certain parameter is in between two values, which is not the definition of a confidence interval. We used the default settings of the Bayesian Estimator in Amos with regard to prior specifications, burn-in, and convergence criteria.

To compare whether the 'unrestricted loadings' or the 'equal loadings' model fitted the data better, we used the Deviance Information Criterion (DIC; [46]). The DIC is the Bayesian equivalent for the AIC/BIC. It is an evaluation between models using a trade-off model between fit and model complexity. Several competing statistical models may be ranked according to their value on the model selection tool used. The one with the best trade-off is the winner of the model selection competition.

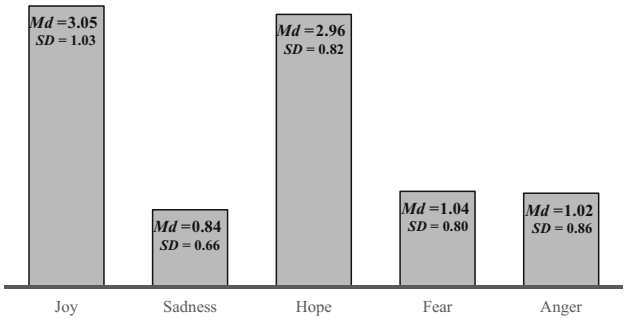


Fig. 6 Mean differences and Standard Deviations for *Emotions* perceived in Tom

3 Results

3.1 Single Occurrence and Isolated Effects

3.1.1 Emotions

Because we were interested in how participants thought Tom felt about them, we first investigated the differences among five emotions. Mean difference scores and Standard Deviations are in Fig. 6. The main effect of Control-type on Perceived intensities was not significant ($F_{(2,51)} < 1$). Also, the interaction between experimental group and Emotion was not significant (Wilks' Lambda = .86, $F_{(5,48)} = 1.595$, $p = .179$).[2] Thus, the participants assumingly did not detect differences in the emotions produced by Silicon Coppélia versus the human confederates.

Because the main effect of Control-type on the Emotion scales was not significant, this might mean that no effects of emotion occurred within a condition. To check for this, we performed a one-sample t-test with 0 as the test value, equaling no emotions perceived. Results showed that all emotion scales differed significantly from 0, the smallest t-value being ($t_{(2,\,51)} = 8.777, p < .001$) for Anger. Thus, effects of Emotion *did* occur. Detailed results are in part C of "Online Resource 1".

To investigate the systematic differences in perceiving emotions in Tom, we computed paired samples t-tests for all pairs of emotions within each condition (see Analyses). Out of the 10 possible pairs, 6 pairs differed significantly. The 4 pairs that did *not* differ significantly were Joy and Hope ($p = .444$), Fear and Sadness ($p = .054$), Fear and Anger

[2] To exclude the option that there were no effects of emotions, perceptions, or decision-making at all, we investigated the intercept. Results showed that the intercept of the five different emotions was significant (Wilks' Lambda = .032, $F_{(5,47)} = 289.45$, $p < .001$, $\eta_p^2 = .97$). Additionally, the intercept of 'Perceptions' was significant (Wilks' Lambda = .009, $F_{(8,45)} = 637.61$, $p < .001$, $\eta_p^2 = .99$). Finally, the intercept of Decision-Making Behavior was significant (Wilks' Lambda = .11, $F_{(2,51)} = 198.126$, $p = .031$, $\eta_p^2 = .89$).

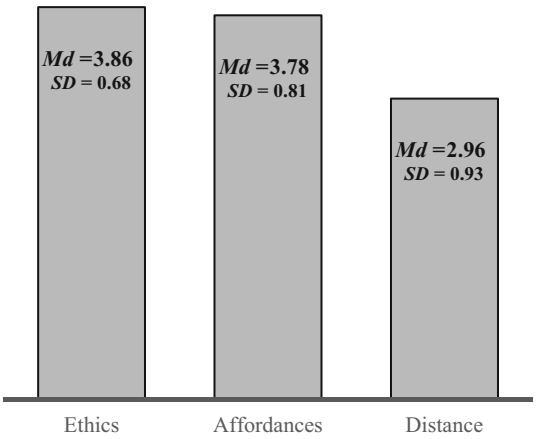


Fig. 7 Mean differences and Standard Deviations for significantly differing *Perceptions* observed in Tom (not all perceptions are listed)

($p = .908$), and Sadness and Anger ($p = .06$). Joy and Hope were both recognized relatively much in Tom, whereas Fear, Sadness, and Anger were recognized in Tom in low intensities.

In conclusion, the *t*-tests showed that in both conditions emotions were recognized in Tom by the participants. According to MANOVA, however, the intensity of emotions in the AI condition was probably not seen as different from the WOz condition.

3.1.2 Perceptions

The main effect of Control-type on Perceptions in a separate MANOVA (see Analyses) was not significant ($F_{(2,51)} < 1$). Furthermore, the interaction between Control-type and Perceptions was not significant (Wilks' Lambda = .85, $F_{(8,45)} < 1$, $p = .462$, $\eta_p^2 = .15$).[2] Thus, it seems that participants did not detect differences in the way Tom perceived them in the AI and WOz conditions.

Again, we performed one-sample *t*-tests with 0 as the test value, equaling no perceptions detected. Results showed that all perception scales differed significantly from 0 (see also Fig. 7). The smallest *t*-value was found for Distance ($t_{(2, 51)} = 15.865$, $p < .001$). Thus, effects of Perceptions *did* occur. Detailed results can be found in part C of "Online Resource 1".

In addition, we analyzed systematic differences in perceiving the perceptions of Tom by paired samples *t*-tests for all pairs (see Analyses). Out of the 28 pairs, 23 pairs differed significantly. The pair that differed most was Ethics and Distance ($t(51) = 13.59$, $p < .001$) (Fig. 7). The participants rated Tom's perceptions of their Ethics and their Affordances the highest. The participants rated Tom's perceptions of feeling distant toward them the lowest.

In conclusion, the *t*-tests showed that Tom's perceptiveness was indeed recognized in all conditions. MANOVA indicated again, however, that the intensity of emotions in the AI condition probably was not seen as different from WOz.

3.1.3 Decision-Making Behavior

The main effect of Control-type on Perceived decisions was not significant ($F_{(2,51)} < 1$). Also the interaction between Control-type and Decision-Making Behavior was not significant (Wilks' Lambda = .97, $F_{(2,51)} < 1$).[2]

Again, we performed a one-sample *t*-test with 0 as the test value, equaling no Decision-Making Behavior perceived. Results showed that Situation selection ($t_{(2, 51)} = 14.562$, $p < .001$) and Affective Decision-making ($t_{(2, 51)} = 15.518$, $p < .001$) both deviated significantly from 0. Thus, the participants *did* perceive Decision-Making Behavior in Tom. Details can be found in part C of "Online Resource 1".

In conclusion, the *t*-tests showed that Decision-Making Behavior was recognized in all conditions and the MANOVA showed that participants most likely saw comparable Decision-Making Behavior in AI and WOz conditions.

These conclusions were substantiated also by Bayesian analysis. Table 2 shows that all the CCIs of the AI and the WOz conditions overlapped. Inspection of the DIC showed that the model in which the regression coefficients were assumed equal across both groups fitted the data better (DIC = 404.40) than the model in which the regression coefficients were unrestricted (DIC = 413.43) as well. This indicates that the participants implied a similar underlying cognitive-affective structure that produced Tom's perceptions and emotions, whether controlled by a human confederate or by our Silicon Coppélia software. These results confirmed our conclusions based on the MANOVAs, which can now be maintained more firmly.

3.2 Structural Equation Modeling with Bayes: Cognitive-Affective Structure

The next step was to test whether the participants recognized a similar cognitive-affective structure in Tom in both the AI and WOz condition. Before inspecting differences between the two experimental groups, we modeled all possible hypothesized direct effects of the encoding variables on the comparison and response variables, as well as all possible direct effects of the comparison variables on the response variables. However, to strive for a more parsimonious model, an association was omitted if it was found to be not significant in either the AI group, the WOz group, or the 'equal loadings' model. We continued removing insignificant associations until only significant associations remained (cf. [22]). Figure 8 shows the results for the model with factor

Table 2 Significant influences in the WOz and AI condition

Influence	WOz condition			AI condition		
	B	SD	CCI_{95}	B	SD	CCI_{95}
Affordances → relevance	.349	.237	−.118, .816	.766	.245	.284, 1.250
Ethics → involvement	.147	.254	−.354, .648	.486	.171	.148, .825
Valence → involvement	.541	.180	.188, .895	.251	.165	−.074, .575
Relevance → involvement	.264	.168	−.006, .599	.261	.137	−.007, .533
Relevance → distance	−.049	.490	−2.006, −.076	−.102	.368	−1.790, −353
Similarity → distance	−.988	.584	−2.111, .192	−.720	.487	−1.629, .247
Similarity * relevance → distance	.245	.201	−.156, .634	.272	.137	−.002, .528
Valence → use intentions	.936	.133	.672, 1.198	1.067	.059	.950, 1.183

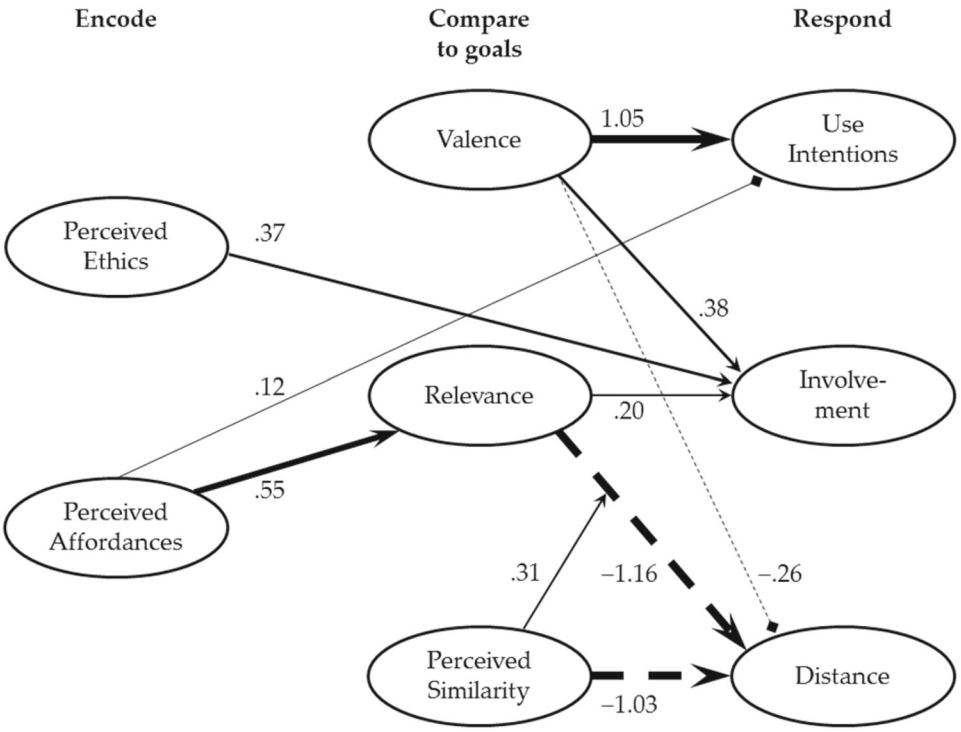


Fig. 8 Significant relations among the theoretical factors. Solid arrows indicate positive influences; dashed arrows negative influences. Pointed arrows are significant at ($p < .05$); blunt arrows represent trends ($p < .10$). The arrow pointing at another arrow indicates interaction. Arrow width is proportional to the estimated effect size. Indirect effects are left out

loadings assumed equal. More detailed information can be found in part D of "Online Resource 1".

Ethics and Relevance significantly contributed to participants' report of Tom being Involved. Thus, if the participants thought Tom perceived them as morally good and relevant, they also thought he would feel more involved toward them. Furthermore, Relevance mediated a negative influence of Affordances on Distance. This influence was further moderated by perceived Similarity. That is, if the participants thought Tom perceived them as clumsy, they thought they would be less relevant for him, making him more distant toward them. This influence was strengthened if they thought Tom perceived them as similar. Finally, Valence had a positive influence on Use Intentions and Involvement. Thus, if the participants thought Tom had high expectations, they thought he would feel more involved with them and more eager to meet again. The found relations matched very well to relations found in previous research regarding I-PEFiC, describing human perceptions of fictional characters (e.g., [51]).

4 Discussion

The goal of the present study was to test whether affective behavior performed by our software could be distinguished from that of a real human. In an enriched variant of the Turing test, we compared the emotional communication of a virtual human as produced by our software Silicon Coppélia in an

AI-condition to a WOz-condition with an avatar in which the emotional communication was produced by real humans (having the same screen appearance). Female participants were randomly assigned to a virtual human named Tom in a speed-date session, either controlled by our software (AI condition) or Tom controlled by human confederates (WOz condition). Tom could give verbal as well as nonverbal cues. Analysis of the results according to MANOVA and Bayes showed that in our version of the Turing test, participants did not detect differences but saw similarity between the two versions of Tom, supporting H1. Participants rated both versions of Tom as similarly eager to meet them again, as exhibiting similar ways to select a situation, and making similar affective decisions. Also, the emotions the participants perceived in Tom during the speed-date session did not differ between both conditions. Differences in emotion attributions could be observed, but these were similar for the human-controlled-Tom and software-controlled-Tom alike.

MANOVA and Bayesian analyses further showed that participants did not detect differences but saw similarity in the cognitive-affective structure between the two versions of Tom, supporting H2. To exclude finding null difference caused by a lack of power, we compared two statistical models, one under the assumption that the AI and WOz conditions were equal, another under the assumption that the two would be different. According to Bayes, the model assuming equality between human and AI explained the data better than the model assuming difference. Thus, the participants attributed a similar cognitive-affective structure to Tom, whether mediated by a human confederate or being performed by our software. In the following, the results, limitations, alternative explanations, and implications will be discussed in more detail.

Figure 1 shows three possible ways of explaining the results: 1) Silicon Coppélia simulated human affective behavior as expected and passed the Turing test. 2) The interface restricted human expressivity too much to outperform the AI. 3) Despite young women supposedly being sensitive to emotional cues, these participants perhaps had poor signal detection.

The main argument in favor of (1) that Silicon Coppélia passed the Turing test is that results occurred as predicted, confirming H1 and H2. Particularly H2 is vulnerable to refutation because it predicts a complex of relations that can easily come out differently. For instance, our results showed that if the participants believed that Tom perceived them as morally good, or more relevant, they also thought he would be more involved with them. If participants believed Tom had positive expectations of them, they also thought he would feel more involved and more eager to meet them again [24].

An extra argument is the specificity of certain findings that replicated results of earlier I-PEFiC research. We could reproduce very specific findings reported in [51]. In that study, male participants were more distant to a male virtual agent if it were clumsy, even more so if the faces were made similar through morphing. This effect was absent in female participants. Apparently, men did not want to be associated with clumsy partners, even if those partners were virtual or non-existent. In the present study, the female (!) participants seemed to have projected this preoccupation onto Tom (being a male). When the participants believed that Tom held them for clumsy, they thought they would be irrelevant to him, which made him more distant toward them, so they thought. This effect was strengthened when they assumed that Tom perceived the participant as similar to him. These women seemingly assumed they conversed with a real male partner, who did not want to be like them if he thought they were clumsy; even if we told them beforehand that the male was a robot. Apparently, empirical findings this precise can be reproduced in participants imagining how a virtual agent equipped with our software perceived them!

Explanation (2) may be that now that the test was passable for the computer, it became too hard for the human to outperform the AI, owing to the restricted interface. The limited expressive possibilities (i.e. selected choice questions and answers) would make it undetectable whether one deals with a computer or not.

First, this may be true but at least we went beyond the standard text-based interfaces and added expressive extras such as voice, decision making, facial expressions, and emotion in line with Barberi's [4] recommendations. Kotlyar and Ariely [25] found that nonverbal cues in an online dating setting were associated with more favorable perceptions, greater exchange of information and a stronger desire to pursue a relationship. This also counted for the human confederates, not for the AI alone.

Second, if we allow free interactions, we lose experimental control, rendering our measurements uninterpretable. Third, if cues to face-to-face contact are unavailable, people seek out and interpret cues that serve as substitutes for nonverbal communication [56]. With respect to gender differences, women's perceptions were especially positively affected by the presence of moving images that accompanied text messages, such as avatar movements, gestures, and facial expressions [56]. Additionally, Lee et al. [28] suggest that adding audio and video to text-only online dating profiles increases social presence and engagement, which is what we did. Fourth, rich communication is not necessarily dependent on a richer medium [50]. A review by Derks et al. [9] even concludes that there is no indication that computer-mediated communication is a less emotional or less personally involving medium than face-to-face.

An alternative explanation related to (2) is that ten minutes were too short for an adequate estimation of the cognitive-affective structure of Tom (cf. [55]). People may well tell human and computer apart when they have sufficient interac-

tion possibilities and plenty of time to develop a relationship. This also may continue for true but then the measurement should be as long as it takes for someone to unmask the system. Yet, in real life as well, we often do not invest much time in our interaction partners and speed-dates are notoriously quick and shallow [11], which means that our experiment had quite some ecological validity for certain real life situations.

In future research (cf. bottom of Fig. 1), our study could be repeated with more interaction options (i.e. language-related) and a longer period of interaction. Point is, how many more interaction options should there be and how much longer should a session take?

Alternative explanation (3) was that the female participants in our sample had poor stimulus discrimination. The idea was that selecting young females allowed for a conservative test. Females make more sensitive diagnoses of someone's state of mind than men (e.g., [47]) and therefore will detect differences between an artificial and a human dating partner better than males. If girls cannot tell the two Toms apart, neither will boys.

It might be that the instruction that Tom was a computer-controlled agent in both conditions influenced the participants' evaluation of Tom. Discourse analysis of a text-based structured conversation revealed differences in participants' behavior when they believed they were talking to a person versus when they believed they were talking to a computer [42]. However, according to the Threshold Model of Social Influence [12], the social influence of real persons who are represented by avatars will always be high (cf. human confederates), whereas the influence of an artificial entity depends on the realism of its behavior. Moreover, in a study by [53], the participants' belief in whether they were interacting with an agent controlled by a computer, or an avatar controlled by a human, barely resulted in differences with regard to behavioral reactions or the evaluation of a virtual character, whereas high behavioral realism of the virtual character affected both [53]. This suggests that the instruction that Tom was controlled by a computer in both conditions should not have had a large effect on the participants' evaluation of Tom. It may even have increased the test sensitivity for measuring the perceived realism of Tom's behavior. In sum, the only way to check alternative 3) is to repeat the study with women who are sensitized and trained in detecting cues to humanness when interacting with a synthetic character versus an untrained group (Fig. 1).

As an extra, the investigation of different personalities and unexpected behaviors also may be something for future research. In the speed-date, Tom never showed extreme behaviors but then again, the human confederate did not do so either. In the current study, we did not want the software to show non-humanlike behaviors (on the contrary). It would be yet worthwhile to investigate how 'strange' an AI is allowed to act before people start recognizing this is not human behavior any more.

The novelty and uniqueness of our approach is threefold. We turned the original Turing test around in that participants were asked how they thought Tom perceived them. Next to asking how the participants experienced the virtual human, participants were asked how they thought they were evaluated by their interaction partner Tom. This was the most straightforward way of testing whether the I-PEFiC model of perceiving virtual others, as implemented in Tom, would result in similar human-like perception mechanisms. Given the speed-date setting, these questions made sense. To the best of our knowledge, no previous studies requested participants to diagnose the perceptions of an artificial other.

Second, the use of both frequentist and Bayesian analysis in robot studies is rare but quintessential if we want to tell whether AI and human differ ((M)ANOVA) and if we want to assess whether they perform similarly (Bayes). Third, we created an AI, Silicon Coppélia, that apparently can simulate human cognitive-affective behaviors so well that under limited conditions of interaction, humans cannot tell apart the AI from actual human behavior (Fig. 1).

5 Conclusion

In having created a humanoid simulation of affect, our models are relevant to engineers as well as scholars: for the first to empower their application development; for the second to increase insights in human affective communication. On all measured dimensions, participants did not experience any significant difference between Tom's communication as generated by a human and Tom's communication as generated by our software. This indicates that the theory and the formalization that translates that theory into software may capture human affective behavior.

Our computer model can be of help in many human—machine interfaces such as (serious) digital games, virtual stories, tutor and advice systems, telemedicine applications [37], coaching or therapist systems (cf. [38]) but also in social support groups (cf. [16]) and virtual patient communities (cf. [29]).

Although we could demonstrate that software agents and robots can closely simulate human emotions and the cognitive-affective structure underlying them, reminiscent of Searle's [41] Chinese Room argument, we do not claim that a computer can be called emotionally intelligent just because it passed our variant of the Turing test. We do claim, however, that we can simulate emotional intelligence so realistically that young women cannot discern between the behavior produced by a man (i.e., a human dating partner) and the behavior produced by our robot through a computer model in a virtual speed-date setting.

Acknowledgements This study is part of the Services of Electromechanical Care Agencies (SELEMCA) project and was supported by a grant from the Dutch Ministry of Education, Culture, and Science (grant number NWO 646.000.003). The authors wish to thank Rens van de Schoot for the Bayesian analysis. We kindly acknowledge Ivy S. Huang for her translation of the abstract into Chinese.

Funding This study was funded by the Dutch Ministry of Education, Culture, and Science (Grant No. NWO 646.000.003).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Arbuckle JL (2007) AMOS 16 [Computer software]. SPSS, Chicago IL
2. Aylett RS, Louchart S, Dias J, Paiva A, Vala M (2005) FearNot!—an experiment in emergent narrative. In: Panayiotopoulos T, Gratch J, Aylett R, Ballin D, Olivier P, Rist T (eds) Intelligent Virtual Agents, vol 3661. IVA 2005. Lecture Notes in Computer Science. Springer, Berlin, pp 305–316
3. Bainbridge WS (2007) The scientific research potential of virtual worlds. Science 317:472–476
4. Barberi D (1992) The ultimate Turing test, minds and machines. Syracuse University, Syracuse
5. Barrett LF, Robin L, Pietromonaco PR, Eyssell KM (1998) Are women the 'more emotional sex?' Evidence from emotional experiences in social context. Cogn Emot 12:555–578
6. Blascovich J, Loomis J, Beall A, Swinth K, Hoyt C, Bailenson JN (2002) Immersive virtual environment technology as a methodological tool for social psychology. Psychol Inq 13:103–124
7. Breazeal C (2003) Emotion and sociable humanoid robots. Int J Hum Comput Stud 59(1–2):119–155
8. Cavallo F, Semeraro F, Fiorini L, Magyar G, Sinčák P, Dario P (2018) Emotion modelling for social robotics applications: a review. J Bionic Eng 15(2):185–203
9. Derks D, Fischer AH, Bos AER (2008) The role of emotion in computer-mediated communication: a review. Comput Hum Behav 24(3):766–785
10. Dias J, Mascarenhas S, Paiva A. (2014) FAtiMA Modular: towards an agent architecture with a generic appraisal framework. In: Broekens T, Dias J, van der Zwaan J (eds) Emotion modeling. Lecture Notes in Computer Science, vol 8750. Springer, Cham
11. Fisman R, Iyengar SS, Kamenica E, Simonson I (2006) Gender differences in mate selection: evidence from a speed dating experiment. Q J Econ 121(2):673–697
12. Frijda NH (1986) The emotions. Cambridge University, New York
13. Gockley R, Simmons R, Forlizzi J (2006) Modeling affect in socially interactive robots. In: Proceedings of the 15th IEEE international symposium on robot and human interactive communication (RO-MAN06), September 6–8, 2006, Hatfield UK. IEEE, New York, pp 558–563. https://doi.org/10.1109/roman.2006.314448
14. Gonsior B, Sosnowski S, Buß M, Wollherr D, Kühnlenz K (2012) An emotional adaption approach to increase helpfulness towards a robot. In: Proceedings of the international conference on intelligent robots and systems (IROS), 7–12 Oct 2012 IEEE/RSJ, Vilamoura, Portugal. IEEE, New York, pp 2429–2436. https://doi.org/10.1109/iros.2012.6385941
15. Gratch J, Marsella SC (2015) Appraisal models. In: Calvo RA, D'Mello S, Gratch J, Kappas A (eds) The Oxford handbook of affective computing. Oxford University, New York, pp 54–67
16. Green-Hamann S, Campbell Eichhorn K, Sherblom JC (2011) An exploration of why people participate in Second Life social support groups. Journal of Computer-Mediated Communication 16:465–491
17. Hayes P, Ford K (1995) Turing test considered harmful. In: Proceedings of the 14th international joint conference on artificial intelligence, Montreal, Canada. Morgan Kaufmann, New York, pp 972–977
18. Hoorn JF (2008) A robot's experience of its user: Theory. In: Sloutsky V, Love BC, McRae K (eds) Proceedings of the 30th Annual Conference of the Cognitive Science Society. Cognitive Science Society, Washington, pp 2504–2509
19. Hoorn JF (2015) Psychological aspects of technology interacting with humans. In: Shyam Sundar S (ed) The handbook of the psychology of communication technology. Wiley-Blackwell, New York, pp 176–201
20. Hoorn JF, Pontier MA, Siddiqui GF (2008) When the user is instrumental to robot goals. First try: agent uses agent. In: Proceedings of IEEE/WIC/ACM web intelligence and intelligent agent technology 2008 (WI-IAT '08), 2. IEEE/WIC/ACM, Sydney, pp 296–301. ISBN: 978-0-7695-3496-1. https://doi.org/10.1109/wiiat.2008.113
21. Hoorn JF, Pontier MA, Siddiqui GF (2012) Coppélius' concoction: similarity and complementarity among three affect-related agent models. Cognit Syst Res J 15(16):33–49
22. Host V, Knie-Andersen M (2004) Modeling customer satisfaction in mortgage credit companies. Int J Bank Market 22:26–42
23. Karelis J (1986) Reflections on the Turing test. J Theory Soc Behav 16:161–171
24. Konijn EA, Hoorn JF (2005) Some like it bad: Testing a model for perceiving and experiencing fictional characters. Med Psychol 7(2):107–144
25. Kotlyar I, Ariely D (2013) The effect of nonverbal cues on relationship formation. Comput Hum Behav 29(3):544–551
26. Kruschke JK (2011) Bayesian assessment of null values via parameter estimation and model comparison. Perspect Psychol Sci 6:299–312
27. Landauer TK (1987) Psychology as a mother of invention. In: Proceedings of the CHI-GI'87 human factors in computing systems and graphics interface. ACM, New York, pp 333–335
28. Lee S, Sun Y, Thiry E (2011) Do you believe in love at first sight: effects of media richness via modalities on viewers' overall impressions of online dating profiles. In: Proceedings of the 2011 iConference. ACM, New York, pp 332–339
29. Leimeister JM, Krcmar H (2005) Evaluation of a systematic design for a virtual patient community. J Comput-Med Commun 10(4):JMCM1041
30. Leite I, Martinho C, Pereira A, Paiva A (2008) iCat: An affective game buddy based on anticipatory mechanisms. In: Proceedings of the 7th international joint conference on autonomous agents and multiagent systems (AAMAS'2008) May 12–16, 2008, Estoril, Portugal, IFAAM, Richland, pp 1229–1232
31. Leite I, Pereira A, Mascarenhas S, Martinho C, Prada R, Paiva A (2013) The influence of empathy in human-robot relations. Int J

Hum Comput Stud 71(3):250–260. https://doi.org/10.1016/j.ijhcs.2012.09.005
32. MacDorman KF, Green RD, Ho CC, Koch CT (2009) Too real for comfort? Uncanny responses to computer generated faces. Comput Hum Behav 25(3):695–710
33. Martinho C, Paiva A (2006) Using anticipation to create believable behaviour. In: Proceedings of the 21st national conference on artificial intelligence (AAAI'06), vol 1 July 16–20, 2006, Boston. AAAI, Menlo Park CA pp 175–180
34. Mori M (1970) Bukimi no tani [The uncanny valley]. Energy 7(4):33–35
35. Moshkina L, Park S, Arkin RC, Lee JK, Jung H (2011) TAME: time-varying affective response for humanoid robots. Int J Soc Robot 3(3):207–221
36. Marinier RP III, Laird JE, Lewis RL (2009) A computational unification of cognitive behavior and emotion. Cogn Syst Res 10(1):48–69
37. Nelson EL, Cook D, Shaw P, Peacock G, Doolittle G (2001) Evolving pediatrician perceptions of a telemedicine program. J Comput-Med Commun 6:00
38. Oppy G, Dowe D (2016) The turing test. In: Zalta EN (ed) The stanford encyclopedia of philosophy. Stanford University, Stanford
39. Pontier MA, Siddiqui GF, Hoorn JF (2010) Speed-dating with an affective virtual agent—developing a testbed for emotion models. In: Allbeck J et al (eds) Lecture Notes in Computer Science 6356:91–103
40. Saint-Aimé S, Le Pevedic B, Duhaut D (2011) Children recognize emotions of EmI companion robot. In: IEEE international conference on robotics and biomimetics (IEEE Robio-2011), 7–11 Dec 2011, Phuket, Thailand. IEEE, New York NY, pp 1153-1158. https://doi.org/10.1109/robio.2011.6181443
41. Searle J (1981) Minds, brains, and programs. Behav Brain Sci 3:417–457
42. Shechtman N, Horowitz LM (2003) Media inequality in conversation: How people behave differently when interacting with computers and people. In: Proceedings of the SIGCHI conference on human factors in computing systems (CHI '03). ACM, New York NY, pp 281–288
43. Shneiderman B (1990) Future directions for human–computer interaction. Int J Hum-Comput Interact 2(1):73–90
44. Si M, Marsella SC, Pynadath DV (2008) Modeling appraisal in theory of mind reasoning. In: Prendinger H, Lester J, Ishizuka M (eds) Intelligent virtual agents. IVA 2008. Lecture Notes in Computer Science, vol 5208. Springer, Berlin, pp 334–347
45. Smith CA, Lazarus RS (1990) Emotion and adaptation. In: Pervin LA (ed) Handbook of personality: theory and research. Guilford, New York, pp 609–637
46. Spiegelhalter DJ, Best NG, Carlin BP, Van der Linde A (2001) Bayesian measures of model complexity and fit. J R Stat Soc Ser B64:583–639
47. Thayer J, Johnsen BH (2000) Sex differences in judgment of facial affect: a multivariate analysis of recognition errors. Scand J Psychol 41:243–246
48. Turing A (1950) Computing machinery and intelligence. Mind 59(236):433–460
49. Van de Schoot R, Hoijtink H, Mulder J, Van Aken MAG, Orobio de Castro B, Meeus W, Romeijn JW (2011) Evaluating expectations about negative emotional states of aggressive boys using Bayesian model selection. Dev Psychol 47:203–212
50. Van der Kleij R, Lijkwan JTE, Rasker PC, De Dreu CKW (2009) Effects of time pressure and communication environment on team process and outcomes in dyadic planning. Hum Comput Stud 67:411–423
51. Van Vugt HC, Bailenson JN, Hoorn JF, Konijn EA (2010) Facial similarity shapes user response to embodied agents. Trans Comput-Hum Inter (TOCHI) 17(2)7:1–27
52. Van Vugt HC, Hoorn JF, Konijn EA (2009) Interactive engagement with embodied agents: an empirically validated framework. Comput Anim Virtual Worlds 20:195–204
53. Von der Pütten AM, Krämer NC, Gratch J, Kang SW (2010) 'It doesn't matter what you are!' Explaining social effects of agents and avatars. Comput Hum Behav 26(6):1641–1650
54. Wallbot HG, Scherer KR (1989) Assessing emotion by questionnaire. Emot Theory Res Exp 4:55–82
55. Walther JB (1992) Interpersonal effects in computer-mediated interaction: a relational perspective. Commun Res 19:52–90
56. Walther JB, Parks M (2002) Cues filtered out, cues filtered in. Handbook of interpersonal communication. Sage Publications, Thousand Oaks
57. Walther JB (2006) Selective self-presentation in computer-mediated communication: hyperpersonal dimensions of technology, language, and cognition. Comput Hum Behav 23:2538–2557
58. Weizenbaum J (1966) ELIZA—a computer program for the study of natural language communication between men and machines. Commun ACM 9:36–45
59. Zimmermann HJ (1994) Fuzzy set theory—and its applications. Kluwer-Nijhoff, Boston

Johan F. Hoorn graduated with two transdisciplinary Ph.D.-theses: his first Ph.D. was in Literature and Psychology (VU University Amsterdam, 1997), while he obtained his second Ph.D. degree in Computer Science (VU University Amsterdam, 2006). He worked at Utrecht University, Tilburg University, and in four different schools at VU University Amsterdam (Humanities, Science, Life Sciences, and Social Sciences). Currently, Johan is full professor of Social Robotics in the Department of Computing and the School of Design of The Hong Kong Polytechnic University. His main research focus is to create social robots as the universal interface between data (Cloud, Internet of Things) and humans, studying the unique social position robots begin to occupy, different from humans, animals, or other digital technology.

Elly A. Konijn graduated in Psychology, Social Scientific Information Systems, and Media Studies, and was a visiting scholar at the City University of New York. Currently, she is full professor of Media Psychology in the Department of Communication Science and a Fenna Diemer-Lindeboom endowed chair at VU University Amsterdam. Her scientific research program has three main lines: (1) Relating to media figures, virtual humans, and social robots; (2) Emotions and media-based reality perceptions; and (3) Media use among adolescents. Application areas are human well-being and health, media education, and creative industries. In each domain, she has published in major scientific journals (e.g., *Nature, Pediatrics*, *Developmental Psychology*, *Journal of Adolescent Health*, *Media Psychology*, *International Journal of Human–Computer Studies*, *Transactions on Computer–Human Interaction* (TOCHI). Elly Konijn wrote several books, was lead editor of *Mediated interpersonal communication* (Routledge, 2008), and co-editor of *The Routledge handbook on emotions and mass media* (Routledge, 2010). She was elected chair of the Information Systems' division of the International Communication Association and editor of the ISI-journal *Media Psychology*.

Matthijs A. Pontier obtained bachelor degrees in Artificial Intelligence as well as (Cognitive) Psychology in 2006 at VU University Amsterdam, and obtained his Masters degree in Cognitive Science in 2007 also at VU. He obtained his Ph.D. in 2011, in which he co-developed

Silicon Coppélia, a computational model of emotional intelligence, person perception, and affective decision making. As a postdoctoral fellow at VU University Amsterdam, he worked in the SELEMCA project on emotionally sensitive social robots that have a sense of morality. He authored 20+ internationally refereed conference publications and journal articles, and co-edited the book *Machine medical ethics* (Springer, 2015).