## Silicon Coppélia and the Formalization of the Affective Process

Hoorn, Johan; Baier, Thomas; Van Maanen, Jeroen A. N.; Wester, Jeroen

***published in***
IEEE Transactions on Affective Computing
2023

***DOI (link to publisher)***
10.1109/taffc.2020.3048587

***document version***
Publisher's PDF, also known as Version of record

# Silicon Coppélia and the Formalization of the Affective Process

Johan F. Hoorn, Thomas Baier, Jeroen van Maanen, and Jeroen Wester

**Abstract**—After 20 years of testing a framework for affective user responses to artificial agents and robots, we compiled a full formalization of our findings so to make the agent respond affectively to its user. Silicon Coppélia as we dubbed our system works from the features of the observed other, appraises these in various domains (e.g., ethics and affordances), then compares them to goals and concerns of the agent, to finally reach a response that includes intentions to work with the user as well as a level of being engaged with the user. This ultimately results in an action that adds to or changes the situation both agencies are in. Unlike many other systems, Silicon Coppélia can deal with ambiguous emotions of its user and has ambiguous 'feelings' of its own, which makes its decisions quite human-like. In the current paper, we advance a fuzzy-sets approach and show the inner workings of our system through an elaborate example. We present a number of simulation experiments, one of which showed decision behaviors based on biases when agent goals had low priorities. Silicon Coppélia is open to scrutiny and experimentation by way of an open-source implementation in Ptolemy.

**Index Terms**—Affect, modeling, goal-driven robots, fuzzy algorithms

## 1 Introduction

For two decades, our group studies fictional characters in various audiences, ranging from movie heroes to robots to user-interface assistants. The main aim is to understand how humans can relate to non-existing others such that they become 'friends' with the system or 'blame' the computer. Anthropomorphization is the false attribution of human qualities to a non-human entity as if it were a real person. This affective attribution process with regard to fictional characters was laid down in [1] and empirically tested in [2], and [3]. This process was extended and tested with user-interface assistants and game characters (e.g., [4]) after which we dared to do a first formalization and have a software simulate affect with humans the way humans show affect for embodied agent systems [5]. We are now ready to attempt a more complete formalization of the empirically validated software model we coined 'Silicon Coppélia', named after dancing Coppélia who set her human lover's heart on fire, although she was a doll.

Previously, we studied how people respond to virtual others from which we know which variables are important and how they relate [2], [3]. That model was named Interactively Perceiving and Experiencing Fictional Characters or I-PEFiC for short [4]. However, I-PEFiC did not tell what functions we should write for the relationships between variables nor how a variable is built up.

The first objective of the current contribution, then, is scientific: To have more precision and a better understanding of how people may relate to non-existent others such as robots. Earlier attempts (e.g., [6], [7]) modeled certain aspects of the users' responses in I-PEFiC but not all that were laid down in [4]. Moreover, the implementation did not allow for different mathematical approaches. The unique position of the current paper in this line of inquiry is that we offer a formal account that is complete and is implemented such that algorithms can be replaced or relations between variables can be altered according to new empirical results or different theoretical insights. For example, if a researcher decides that normalized Hammings is a better estimate of similarity than our set-theoretical approach, the algorithm can be replaced without harming the integrity of the model as a whole.

The second objective, then, is technical: To deliver a tool that researchers may use to refine their hypotheses, simulate the results, and use those as test predictions for lab and field tests. Complementary to this article we also provide a software implementation [8] of the presented theory that provides an easy starting point for such experiments.

The third objective is in design. If we know how users come to like or dislike their agent systems and robots, we can use that information to simulate behaviors that the user might appreciate. If desired, it may make robot behaviors seem more human-like. The formal model may assist designers and application programmers in developing and fine-tuning the affective behavior of their agent systems (cf. [9]). Our system will be applicable in technical domains related to perceptual and/or behavior generation stages, seeking intermediate modules.

The general findings of I-PEFiC are as follows: There are nine relevant dimensions in user affect for a virtual being.

- *Johan F. Hoorn is with the Department of Computing and School of Design, The Hong Kong Polytechnic University, 11 Yuk Choi Rd, Hung Hom, Hong Kong. E-mail: csjfhoorn@comp.polyu.edu.hk.*
- *Thomas Baier is with the Faculty of Social Sciences, Department of Communication Sciences, Vrije Universiteit Amsterdam, De Boelelaan, 1105, 1081 HV Amsterdam, The Netherlands. E-mail: t.baier@vu.nl.*
- *Jeroen van Maanen is with The Future Group, Röntgenlaan 27, 2719 DX Zoetermeer, The Netherlands. E-mail: jeroen@leialearns.org.*
- *Jeroen Wester is with Wieswies-Information Technology and Services, Heirweg 34, 7841 AP Sleen, The Netherlands. E-mail: jeroen.wester@wieswies.nl.*

*Manuscript received 30 Apr. 2020; revised 3 Oct. 2020; accepted 23 Dec. 2020. Date of publication 8 Jan. 2021; date of current version 28 Feb. 2023.*
*(Corresponding author: Johan F. Hoorn.)*
*Recommended for acceptance by G. Castellano.*
*Digital Object Identifier no. 10.1109/TAFFC.2020.3048587*

We will explain those later in Section 3 but two important ones are the ethical conduct of an agent system (i.e., does it do harm or not) and its affordances (i.e., the things you can do with it such as 'conversation' or 'playing a game'). The same features of an agency are assessed on more dimensions. Features or clusters of features raise positive and negative responses concurrently. A robot may be appreciated for working effectively but simultaneously that effectiveness may induce fear of job loss. The relational and functional side of virtual others are intermingled (e.g., a good tool also is a better friend) (cf. [9]). Being involved with a character does not exclude feeling some distance in parallel. Involvement and distance are not mutually exclusive.

For modeling these results, we describe an agency by a set of qualitative features each of which leads to its own affective appraisal. Unlike neural networks and other correlational approaches, this approach keeps everything traceable during simulation (i.e. the causality remains observable). To account for features being assessed on more dimensions or to formalize the Involvement-Distance trade-off, it should be possible for one and the same feature to have membership functions in multiple sets, which fuzzy approaches allow for. When researchers beg to differ, they can replace ours with their approach and in a simulation experiment observe what different behaviors the agent system may exhibit with the alternative algorithm.

## 2 Related Work

The field of affective computing is growing rapidly; in extensiveness, completeness, and excellence of research. It is almost undoable to give a fair account of the state of the art here but luckily [7] did a tremendous job in this respect. In this section, nonetheless, we attempt to position Silicon Coppélia within the set of comparable affective systems.

Picard [10] defined three perspectives on affective computing, namely to recognize user emotions, to simulate emotions that humans would recognize, and to process emotions the way humans do. Silicon Coppélia acts in the latter two areas, mostly the third, however, Coppélia is not oriented on discrete emotions as such like, for example, Cathexis is [11], but rather on the cognitive-affective processes that lead to such emotions. If we follow the criteria that [12] identified in their excellent overview of affective systems, we can position Coppélia with respect to theory, operation, cognition and emotion, architecture, and agent technology, which we will pursue next.

In line with models such as Mamid [13] and Emotion and Adaptation (EMA) [14], Coppélia discerns an affective mode of processing that maps onto a cognitive-reflective mode of processing, running in parallel. In [13], however, the relations between affect and reflection are not stipulated in detail, which is something the current paper attempts to do by specifying the internal architecture of the agent system. Mamid does account for higher magnitudes of an emotion, impacting the reasoning more strongly. In Coppélia, magnitude would be represented by the notion of Relevance.

Different from, for example, Flame [15] (Fig. 1), Coppélia does not assume a strict taxonomy of emotions based on a hierarchy of conditions [16]. It also does not work from individual personality as Mamid [13] does. Instead, the theoretical

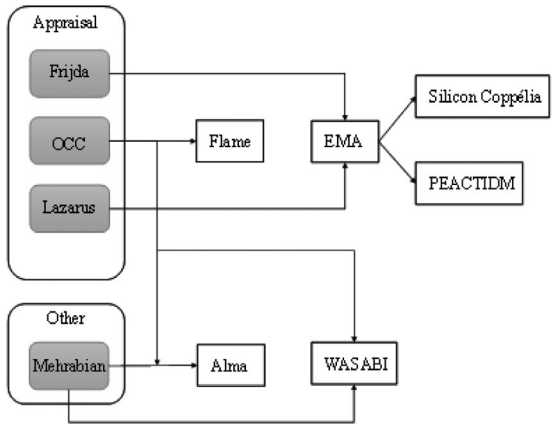


Fig. 1. Silicon Coppélia is partly based on EMA (abbreviated lineage of affective models adapted from Gratch and Marsella ([21], p. 60)).

foundations of Coppélia are found in Frijda [17], Smith and Lazarus [18] (Fig. 1), and Gross [19] and connected to user experiences of virtual agents by [4] and [20]. These are mainly dimensional theories that assume factors such as valence and relevance ('arousal') to underlie various distinct emotions. In that sense, Silicon Coppélia leans on EMA [14] (Fig. 1) for a subset of appraisal dimensions (e.g., relevance, current and future valence) and for the attempt to explain hope, joy, fear, sadness, anger, and guilt. Placed in the overview of Gratch and Marsella ([21], p. 60), Silicon Coppélia would be a descendant from EMA (Fig. 1). In EMA, the state of the surrounding world is observed and appraisals are continuously brought up-to-date. EMA appraises the utility of an agency or event as support or hindrance ('desirability'), estimates the probability that a given world-state actually occurs, and whether that world-state is predictable from cause-and-effect ('expectedness'). However, Coppélia's scope goes further than emotion appraisal and coping, which is EMA's main focus.

In being a goal-directed autonomous agent, Coppélia somewhat resembles PEACTIDM [22] for that matter. PEACTIDM (Fig. 1) produces appraisals in various ways and the values of appraisal dimensions may differ as well. Variables such as 'causal agent' are nominal (self, other, nature) whereas outcome probability, relevance to goals, and unexpectedness range between 0 and 1. PEACTIDM follows EMA by enclosing an agent or an event in a (sometimes incomplete) appraisal frame, which is used for focusing attention and selecting an interaction partner. Similar to PEACTIDM, Coppélia estimates appraisal information straightforwardly from its processes with a focus on delivering appraisals, selecting actions to perform on a targeted agency, while the associated emotions are supplementary.

In the operating cycle [12], the type of stimuli Silicon Coppélia can assess are agencies such as humans or humanoids (e.g., android robots) and their actions. The main factors involved in the processing of affect are ethical behaviors of the observed human(oid) (Ethics), its Aesthetics, levels of realism (Epistemics), and action possibilities (Affordances). Silicon Coppélia outputs affective states such as 'feeling involved' and 'at a distance' and selects actions to change or continue a situation (cf. approach or avoid). In terms of [23], visual cues, audio, gestures, eyes, etc. can be evaluated as long as their multi-modality is

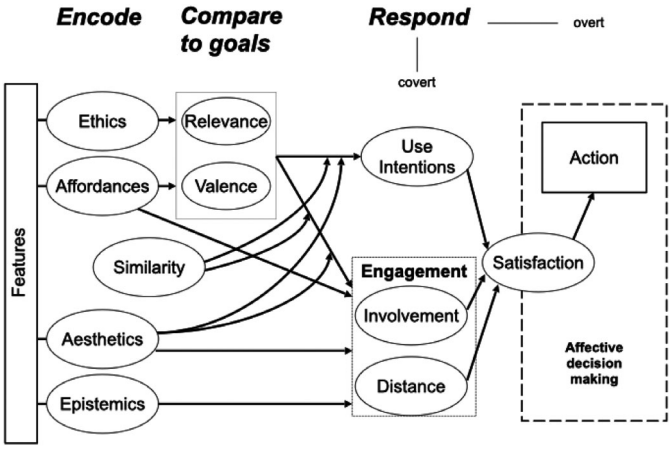


Fig. 2. Graphical Representation of I-PEFiC ([4], [30]).

represented by sets of fused features, translated into a set of general features [23]. The combined features are then used for further processing (ibid.). Those sets are assumed fuzzy (cf. [24]) to model simultaneously occurring affective processes (and hence, multiple emotions occurring at the same time). Coppélia has intensity and decay functions modeled as feature weights.

The interaction between cognition and affect [9], [12] takes place between two groups of factors. The agency under observation is approached by Silicon Coppélia as a potential friend (the affective component) but also as an instrument to achieve goals (the cognitive component). However, both sides inform and influence one another, which is accomplished by the use of fuzzy sets and fuzzy operators (cf. Flame, [15]).

Silicon Coppélia has been used in a speed-dating setting [25], in playing games with robots [26], and in agent-only soap stories [27]. Although Coppélia provides the information to act upon, in itself, it has no capabilities to speak or act, which have to be designed with each new application. This is different from WASABI [28] (Fig. 1), for instance, which does provide processes to express affect (e.g., facial expression) and different from Alma [29] (Fig. 1), which has dialog systems in place.

## 3 Sillicon Coppélia

The layout of the affective process can be retrieved from the aforementioned publications so this paper will limit itself to understand the basic principles that drive Silicon Coppélia. We will not delve deeply into the theoretical ins and outs of each variable; that was done in [1]. Illustrated with two realistic use cases, the current section provides a quick overview of the variables involved and how they relate to one another. Although Fig. 2 shows four dimensions for encoding an agency, the two most important are moral behavior (in our terms 'Ethics') and possibilities to act (what we call 'Affordances').

Next, we describe the variables of I-PEFiC that Coppélia uses to simulate an affective state. For illustration purposes, the use cases we discuss are friendship formation with a robot of older adults that feel lonely [31] and self-disclosure to a robot but not to human beings [32]. The first and foremost appraisal dimension in I-PEFiC is *Ethics* (Fig. 2). It is the evaluation of features of another agency as morally good (e.g., benign, not harmful, just, honest) and/or bad (e.g., vicious, mean, unjust, snooping). For example, a user may assume that a robot does not gossip, which is 'good' behavior. Therefore, s/he trusts the robot and discloses life secrets to it (cf. [32]). However, if the user also suspects monitoring functions in the robot, s/he may be concerned about privacy and so the 'spying' is regarded as 'bad' behavior.

*Aesthetics* pertains to the appearance of the other agency. Particularly in early friendship formation, how the other looks (pretty, beautiful) plays a role in being attracted or not. For all variables, it may be that features are classified as members of more classes (fuzziness); hence, a feature may be partially good, partially bad as well as pretty (cf. 'femme fatale').

*Epistemics* has multiple layers but the most researched one is perceived realism, meaning the degree of fidelity to which an agent system simulates human behavior or looks human-like. Designers usually maintain some balance between human-realistic features to stimulate interaction and some unrealistic features (e.g., cartoonish looks) to avoid uncanny effects. However, Epistemics can also relate to 'the validity of one's words,' meaning that the other agency may maintain an unrealistic outlook on life (which becomes germane when a robot should converse with an Alzheimer's patient).

*Affordances* relate to the instrumental side of the agency, particularly to the things a user can do with the agency, its action possibilities. If a robot has a mouth, the user will assume it can speak. In [31], one of the older women wanted to feed the robot cake, a mistaken affordance much to her disappointment. If a tabletop robot like iCat or Karotz has no legs, it does not afford to go for a stroll in the park. Certain affordances will help the user achieve goals; we call them 'aids'. Other will stand in the way; we call them 'obstacles.' Note that users may assess all variables, and particularly Affordances, for many different features, which may become a mixed conglomerate of aids and obstacles when the agency is assessed as a whole.

After encoding the features, comparisons start between the observing agency and the agency that is observed. Does he look like me, does she have a different character? *Similarity* is a measure of the degree to which one agency thinks s/he is like (or different from) the other. This could be appearance as well as social background or personality. It is not necessarily so that more similarity leads to better liking. Not everybody wants to talk to a lookalike robot and sometimes distinct but desirable qualities in the other may lead to attraction. In the current paper, we model similarity and dissimilarity separately as the reflection of both intersecting and distinctive features between two agencies following the line of [33]. In our approach, neither symmetry between the two agencies nor between similarity and dissimilarity is required. This is a theoretical position that may be solved otherwise, for instance, through correlation. We will later discuss why we opt for a (fuzzy) feature set approach (see also Section 5).

Another comparison is about the *Relevance* of a feature to the goals, concerns, and needs of the observing agency. If legs are irrelevant to the goal of having a conversation, malfunctioning legs have no meaning for affect (see [34]). If the

goal is to walk in the park, bad legs become relevant and disappointment may occur. Rooted in the importance of the concern, Relevance regulates the intensity of affect ('great' or 'little'), not so much its direction. *Valence* determines the direction of affect (e.g., 'sad' or 'happy'). Valence pertains to what the agency expects a (set of) feature(s) may do in achieving goals - will it facilitate a positive outcome or a negative one? Both Relevance and Valence are practically always involved in the assessment of an agency: They are the 'emotion engine.' If privacy is a big concern, then spyware on the robot is assessed as ethically 'bad;' it has high relevance (thus, exerting intense emotions), and negative valence (angrily, the user will not self-disclose).

After comparison, the response phase commences: Three variables absorb the outputs from the comparison phase. *Use Intentions* are mainly fed by Affordances (as mediated by Relevance and Valence). Agencies may be willing or not to continue interacting with the other agency. In [31], one of the older women was disappointed that after the experiment was over, the robot would not return to her home and she had to 'keep her mouth shut again.' Apart from this instrumental approach to others, one may feel friendship as well. *Involvement* is to have warm and friendly feelings for an agency and *Distance* is to have cold and unfriendly feelings. Involvement also may include affect that is related to empathy, flow, and being challenged. Distance also refers to being aloof or bored. Involvement is the felt tendency to approach an agency, whereas Distance is the felt tendency to avoid it. Involvement and Distance occur concurrently and compensate one another. For example, a robot may break down once in a while but also may give valid health advice. The combination of local involvement-distance conflicts into a more global trade-off is established with a *fuzzy-or* operator. This internally experienced Involvement-Distance trade-off feeds into a measure of *Satisfaction* with the agency. The most interesting characters awaken both feelings. From a viewpoint of bonding and relationships, the highest Satisfaction comes from an optimal Involvement-Distance trade-off. Thus, the processes of Involvement ('you will be my new friend') or to feel at a Distance ('you bore me') occur in parallel. For instance, the need for self-disclosure may foster involving tendencies whereas the suspicion of monitoring functionality may strengthen distancing trends, the outcome being an ambiguous state of 'a friend that you keep at a distance.' Together with Use Intentions, Involvement and Distance result in a container judgment of Satisfaction ('He is Okay', 'She is so so').

The level of Satisfaction with that agency in that particular situation governs affective decision making. Frijda [17], Smith and Lazarus [18], and Gross [19] discern a number of action tendencies, which Coppélia may execute to change a given situation: positive approach (e.g., to compliment the other), negative approach (e.g., criticize the other), avoid (e.g., turning away from the conversation), or do nothing. In our implementation, we substituted doing nothing by 'changing the other' so that Coppélia may be used for teaching or therapeutic purposes as well. Under consideration of the action tendencies an action is chosen and executed. The result loops back into a novel situation at hand and the encode phase may start again with a new assessment of the other agency.[1]

The purpose of this framework is to simulate how an agency relates to another agency and what actions to perform based on its affective states. Like this, we can model ambiguous behaviors. An agency could choose to take the action of asking another agency for help while also expressing an affective aversion to that action (e.g., embarrassment, reluctance). Our framework is meant for determining affective state as well as for action selection.

Features of humans or other agent systems (together 'agencies') that enter Coppélia are indexed for several variables that are important for the appraisal of someone else (either organic or artificial). The remainder of this paper offers the calculation steps for the affective process and decision making. It merely addresses the interaction with one single other agency, not more. In principle, Coppélia does nothing but throughput, output, and feedback loops. As is, it has no means to express itself.

In sum, Fig. 2 shows a graphical representation of Interactively Perceiving and Experiencing Fictional Characters (I-PEFiC, [4]). On the input side, features are observed and encoded with respect to the appraisal domains Ethics, Aesthetics, Affordances, and Epistemcs (encoding phase). These observations are matched against the beliefs that the agent holds to determine the Relevance and Valence of the features. Also, the agent compares the other agency for Similarity to its perceived self (compare phase). Based on the previous two phases and the actions that the agent believes are at its disposal, the agent evaluates its level of Involvement and Distance towards the other agency as well as its Intentions to make Use of the observed features of the other agency. These values are combined into the Satisfaction level that the agent expects, which is the basis for the agent's decision about which action to take (response phase).

In the following, we give a detailed formalization of the calculation steps involved in the affective decision process. The formalization is to a large extent based on previous work, mainly [27], [35] and references therein. Here we give, however, a rigorous mathematical formulation of the steps in the affective process. In particular we give a description of feature encoding and the variables involved in the affective process based on fuzzy sets. We formalize Similarity in this setting based on Tversky's ratio index and express Relevance and Valence in terms of fuzzy rules that are based on an empirical model. During the individual steps, we give references were we adopt specific parts from previous work.

## 4 Features and Encoding

In the encode phase (see Fig. 2), the artificial agent observes features of another agency in a given situation. Those features receive a series of numerical values that describe various appraisal dimensions of the features. These values act as the primary input to the affective process and are used

1. In the current form, Silicon Coppélia is an attempt to model the static problem of deciding on the best action in a given situation. To apply this model in a dynamic context, it would need to be extended with a way to prevent instabilities that could arise from the discrete nature of the selected actions.

for further processing of the agent's observations, an approach that to a degree is indebted to the work of Gratch and Marsella in ([7], pp. 54-67).

### 4.1 Features

The features the agent observes are the product of a complex observation process (cf. [30], [33]) and at this point, we are not concerned how this process looks like in detail, this is a task that needs to be solved outside of the theory presented here. We assume that the agent, into which the affective decision process is embedded, has a mental model (or belief system) of the world that also has a model of the current situation and tasks to be solved. The observed features, as well as other inputs to Silicon Coppélia, are derived in the context of this belief system, and the systems that provide these inputs are aware of the current situation. They can derive the available actions and their relations to features, determine the goals of the agent, and have an estimate of what effect actions might have. As such, the agent is capable of extracting a (typically small) set of qualitative features that are relevant in the given situation as input to the affective decision process.

The observed features can be simple attributes like *blue* or *small*, but are not restricted to that and can be *"[...] generally rich in content and complex in form. It includes appearance, function, relation to other objects, and any other property of the object that can be deduced from our general knowledge of the world."* [33].[2] In the context of our example of a lonely plastic surgeon looking for a date (Section 11), *"she is an attractive woman in her middle ages"* would count as a 'feature'. The features are elements of a space of possible features $\Omega$ and the observation of the agent results in a set of features[3]

$$F = \{f^{(1)}, f^{(2)}, \ldots, f^{(N)}\} \subset \Omega. \tag{1}$$

Again, in a simple case, $\Omega$ could be a set of discrete labels. In a more complex case, features could be composed of other features or attributes, i.e. $\Omega$ is a subset of the power set over some other set $\Omega'$. For example, a feature could be described by a set of synonyms in a thesaurus. Let us, however, for the sake of simplicity assume that $\Omega$ is finite.[4]

During the affective decision process, each feature is treated individually and the results are combined at the end in the response phase to form a decision. Note that if the belief system treats a subset of features identically, i.e., if the inputs to the affective process as described in the following sections are the same for all of those features, then all features in this subset will also result in the same outputs and we can treat them as equivalent. In this case, we can group them and replace them in (1) with a single $f^{(i)}$ representing the equivalence class they form.

2. In fact, the exact form of the features will usually not be determined by the true nature of the involved entities rather than by the model the belief system has about them and the task to solve.

3. For convenience, we will sometimes identify $F$ with the index set $\{1, 2, \ldots, N\}$.

4. This is not a necessary assumption but releases us from mathematical complications. In digital data processing, $\Omega$ is always finite, even though a mathematical model may assume infinite $\Omega$, for instance, for continuous quantities.

### 4.2 Labels and Weights

Before tackling the mathematics of encoding the observations of the agent, we need to introduce the overall framework of representations that will be used in this paper.

First, we found that for our purposes we do not have a need for any structure or constraints on the encoding of the characteristics that make up features. For surrounding systems, they could be logical categories, segments of a statistical space, outputs of neural networks, or any other complex structure. Nevertheless, to us they are just labels.

The magnitude of the significance of a label for a specific calculation is more complicated. In that respect, our search for a way of modeling affective behavior has been an exercise in letting go of constraints. We started out with sets of labels. Each characteristic is either inside or outside the relevant set. It is immediately clear, however, that this is insufficient to model affective behavior. Every day we encounter many situations where we would have acted otherwise.

The next approach we considered was to assign probabilities to labels, but then we would have to construct a model where every term has a probabilistic interpretation. For example, if our model calculates with a value of 0.65 for the probability that a feature is ugly, we should be able to specify what we should measure to arrive at a judgment of ugly 65 times out of a hundred. We fear that for many sensory systems that we would like to consider to provide the inputs for Sillicon Coppélia, this would be impractical. It also makes it very hard to balance magnitudes of different kinds (e.g., compare how desirable a characteristic is with how beautiful it is). This is also something humans do all the time.

Therefore we decided to build our model around *weights*. A weight is a number that represents the significance of a label in a given context without any constraints on the interpretation of that number in terms of probabilities or logical entailment. Like probabilities, we take most weights from the interval [0,1]. In some cases, when one extreme has the connotation of 'very negative' rather than 'not significant' it is more natural to use weights in the interval $[-1, 1]$. Precisely because of the lack of constraints on weights, they can be used to represent subjective judgments. Although we only address the static setting of how an agency decides how to respond to a given situation in this paper, we envision that a dynamic system that continually guides behavior will have to be described by a model that states how these weights change with the unfolding of events. We will make extensive use of the fact that a set of weighted labels maps very naturally to a fuzzy set [36]. In the next section we introduce appraisal weights. In subsequent sections other types of weights will be introduced to model various aspects of Silicon Coppélia.

### 4.3 Encoding

During encoding, each feature is evaluated with respect to a set $D$ of indicative (+) and counter-indicative (-) appraisal variables. Ethics relates to good (indicative) and bad (counter-indicative), Aesthetics to beautiful and ugly, Epistemics to realistic and unrealistic, and Affordances to aid and obstacle

$$D = \{eth, aff, aest, ep\} \times \{(+), (-)\}. \tag{2}$$

Each feature $k \in F$ receives[5] a weight value between 0 and 1 for each appraisal variable, which leads to a vector

$$\vec{d}^{(k)} = (d_i^{(k)})_{i \in D}, \qquad \text{where} \qquad d_i^{(k)} \in [0,1], \tag{3}$$

of *appraisal weights*. In terms of fuzzy sets, the weight values $\vec{d}^{(k)}$ can be considered as grades of membership of feature $k$ in $D$. For example, a feature to some degree may be regarded as good ($eth_{(+)}$) as well as bad ($eth_{(-)}$) at the same time.

In addition to the appraisal weights, which encode how relevant a feature is to the agent with respect to the appraisal variables, the agent can also consider certain features to be more salient than others, because of personal focus, bias, earlier observations, because the feature is partially occluded or because of other observational restrictions. All these possibilities are caught by additional *feature weights* $w^{(k)}$ for each feature $k \in F$, where $w^{(k)} \in [0,1]$. The magnitude of a feature weight could be derived from the strength of a signal (amplitude, loudness), a probabilistic judgment or the frequency by which a feature is observed by the agent, but, following the reasoning of the previous section, the weight itself needs no objective interpretation. Feature weights together with the appraisal weights lead to *perceived weights*

$$\vec{p}^{(k)} = (p_i^{(k)})_{i \in D} = (w^{(k)}\, d_i^{(k)})_{i \in D}, \tag{4}$$

for each feature $k \in F$.

## 5 Comparing for Similarity

Looking, feeling, and thinking alike or not; sharing the same background or being different bare relevance on social relationships, virtual agents and robots included. This is measured by Similarity which is a further input variable that enters the affective process during the compare phase in Fig. 2. A popular way of modeling similarity between two entities, particularly in machine learning, is to use a symmetric distance metric. However, the choice of a metric (e.g., Euclidian, Hammings) may depend on the (psychological) theory one adheres to or the distribution of the data sets in the comparison (e.g., 'Gaussian' may require sum of squared differences and 'exponential' perhaps sum of absolute differences). Our system is designed such that researchers can insert different algorithms to test different performance.

In the following, however, we assume that underlying the perceived (dis)similarity are the intersection and differences between the observed set of qualitative features and a feature set representing the agent itself. The first to consider intersection in relation to distinctive sets was Amos Tversky [33] in his critique on similarity estimates being symmetric. Unlike in geometric distances, empirically, people make asymmetric similarity judgments, stating that the son looks more like his father than vice versa. According to Tversky, asymmetry is explained from the different sizes of the distinctive sets, which are indicative of the different amounts of knowledge one has about the compared entities.[6]

Nonetheless, classic set theory is insufficient as feature detection is driven by appraisal dimensions such as ethics and aesthetics, including individual biases, contextual setting, and situation. This is why in our approach, we make use of weighted features and indices that specify the biases. More importantly, one single feature may be assessed on many dimensions: A robot's eyes may be designed attractively (Aesthetics: beautiful) but the cameras are broken (Affordances: obstacle). The most straightforward way to extend set theory such that features have membership functions in multiple sets is Zadeh's fuzzy logic (e.g., [36]). Not only do fuzzy sets allow for concurrency, uncertainty, and non-linearity (i.e. asymmetric judgments), they can deal with incompleteness and imprecision as well, which are typical for empirical data sets. Additionally, fuzzy sets do not have to be purely numerical, symbolic and linguistic elements can be mixed in as well. Fuzzy algorithms remain stable under situational change, even when rules that are applied to the situation are wrong or ignored.

To evaluate Similarity between an agent and another agent, the observing agent will retrieve its own features $F_{\text{self}}$ from its belief system [30] and encode them by the same procedure that is used for assessing the features of the other agent as described in Section 4.

As a measure $s$ of Similarity between the resulting two feature sets $A$, $B \subset \Omega$, we use Tversky's feature-based ratio model [33], also called the Tversky index

$$s(A,B,\mu,\alpha,\beta) = \frac{\mu(A \cap B)}{\mu(A \cap B) + \alpha\mu(A \setminus B) + \beta\mu(B \setminus A)}, \tag{5}$$

where $\alpha, \beta > 0$, $\mu$ is a non-negative, increasing function that is a measure of the weight of the sets[7] and $s$ takes values in [0,1].

The Tversky index is based on the intersecting and distinctive features of the sets $A$ and $B$. The magnitudes of the coefficients $\alpha$ and $\beta$ determine on the one hand the strength with which the intersecting features of $A$ and $B$ enter the Similarity index compared to the features on which the sets differ. The Tversky index is, on the other hand, intentionally not symmetric in $A$ and $B$ and the ratio between $\alpha$ and $\beta$ controls the influences of the difference $A \setminus B$ compared to $B \setminus A$ on the similarity. The measure $\mu$ weights the importance of a feature from the viewpoint of Similarity: Congruence or disparity of features with high weights has a larger impact on Similarity than of features of low weight.

Note that the measure $\mu$ used here is different from the weights $w^{(k)}$ used in (4) as the importance of a feature for the affective process can differ from its importance in the context of similarity estimation. For instance, in the context of our example in Section 11, the disease of a patient is a key feature for a surgeon when considering whether to operate on the patient or not, but can be of minor importance when the surgeon compares herself to the patient in terms of similarity.

During encoding, the set of nominal features is augmented by cardinal values in the interval [0,1] for each appraisal variable in $D$. As (5) is based on a set-theoretical description of the features, we need to define a set-theoretical

5. Again, as for the observation process, we are not concerned here about how the encoding takes place in detail.

6. Tversky shows in [33] that if a similarity order on feature sets fulfills the axioms *matching, monotonicity, independence, solvability and invariance* (as described there), it can always be represented by a function of a weight of the intersection and the relative complements (set differences) of the feature sets.

7. In mathematical terms, $\mu$ is a *measure* on the set $\Omega$.

representation of these real values to include them into the Similarity calculation. One way to achieve this is to identify each value with the corresponding interval $x \to [0, x] \subset [0, 1]$. Then we can represent the features together with their encoding as sets in the product space[8] $\Omega \times D \times [0, 1]$. Let us denote those weighted feature sets by[9]

$$\widehat{F} = \bigcup_{k \in F} \hat{f}^{(k)}, \text{ where } \hat{f}^{(k)} = \Big\{ (k, i, x) \,|\, i \in D,\, x \in [0, d_i^{(k)}] \Big\}. \tag{6}$$

An existing measure $\mu_\Omega$ on $\Omega$ (weights of the features) can be extended to the product measure of $\mu_\Omega$ with a measure $\mu_D$ on $D$ (weights of the appraisal variables) and the ordinary interval length as a measure on [0,1] (Lebesgue measure)

$$\mu\big(\{(k, i, x) \,|\, x \in [x_1, x_2])\}\big) = \mu_\Omega(k)\, \mu_D(i)\, |x_1 - x_2|, \tag{7}$$

where $k \in F$, $i \in D$ and $x_1, x_2 \in [0, 1]$. Then the weight of a feature $\hat{f}^{(k)}$ becomes

$$\mu(\hat{f}_k) = \mu_\Omega(k) \sum_{i \in D} \mu_D(i) |d_i^{(k)}|. \tag{8}$$

in this setting.[10] With this construction of the weighted features, the weights of the individual appraisal domains add up for each feature. Further discussions about calculating Similarity as based on the Tversky index in this setting are in the remarks at the end of this section.

Just as similarity is not necessarily symmetric with respect to the compared entities, dissimilarity is not necessarily directly derived from similarity. The agent can evaluate the degree to which it is similar or dissimilar to another agent with different emphasis on the commonalities and disparities between the features as well as with different weights for the importance of a feature or appraisal variable (cf. [33]). This leads us to the following definitions for an agent being similar or dissimilar to its counterpart:[11]

$$\begin{aligned} sim^{(k)} &= s(\widehat{F}_{self}, \widehat{F}_{other}, \mu^\Omega_{sim}, \mu^D_{sim}, \alpha_{sim}, \beta_{sim}), \\ dis^{(k)} &= 1 - s(\widehat{F}_{self}, \widehat{F}_{other}, \mu^\Omega_{dis}, \mu^D_{dis}, \alpha_{dis}, \beta_{dis}), \end{aligned} \tag{9}$$

where the agent can use different parameters $\alpha$ and $\beta$, as well as different measures $\mu$ in the evaluation of being similar and/or dissimilar.

### 5.1 Remarks on Calculating Similarity

- The ratio model provides a normalized relative distance measure, i.e. Similarity depends not on the absolute difference between the feature sets but their relative, or perceived, difference. That means for example that for a feature that is encoded with a low weight in an appraisal variable for both the agent itself and its counterpart, Similarity reflected by that feature can be still low if the encoded weights differ relative to each other (i.e., their ratio is large).
- To illustrate the properties of the above definition of Similarity between cardinal values $x_1$ and $x_2$ from the interval [0,1], let us look at the case of a one dimensional variable and let $s'$ denote the Similarity between the two values $x_1$ and $x_2$. Then $s'$ is given as the Tversky index of the corresponding intervals and if we assume $\alpha = \beta = 1$, this takes the form

$$\begin{aligned} s'(x_1, x_2) &= s([0, x_1], [0, x_2]) \\ &= \frac{\mu([0, \min(x_1, x_2)])}{\mu([0, \max(x_1, x_2)])} \\ &= \frac{\min(x_1, x_2)}{\max(x_1, x_2)}, \end{aligned} \tag{10}$$

  i.e., in this special case $s'$ is the ratio between the two values $x_1$ and $x_2$.
- Let us examine the influence of the feature encoding on the Similarity between two agents. Let us assume that the same feature $k$ is encoded in two different ways $\hat{f}^{(k)}$ and $\hat{f}'^{(k)}$, where

$$\begin{aligned} \hat{f}^{(k)} &= \{(k, i, x) \,|\, i \in D,\, x \in [0, d_i^{(k)}]\}, \\ \hat{f}'^{(k)} &= \{(k, i, x) \,|\, i \in D,\, x \in [0, d'^{(k)}_i]\}. \end{aligned}$$

  Again we look at the case $\alpha = \beta = 1$ and assume all appraisal variables are weighted equally, i.e., $\mu_D$ is uniform, then the Similarity between these differently encoded features is given by

$$s(\hat{f}^{(k)}, \hat{f}'^{(k)}) = \frac{\sum_{i \in D} \min(d_i^{(k)}, d'^{(k)}_i)}{\sum_{i \in D} \max(d_i^{(k)}, d'^{(k)}_i)}. \tag{11}$$

- Let us also examine the Similarity between two different features sets $\widehat{F}$ and $\widehat{F}'$, where the encoding of each feature $k$ that is present in both $\widehat{F}$ and $\widehat{F}'$ is the same for both feature sets. In this case the Similarity of the encoded feature sets $\widehat{F}$ and $\widehat{F}'$ resembles the Tversky index on the original feature sets $F$ and $F'$ with the difference that the measure $\mu_\Omega$ on the feature space is modified by the measure of the encoded appraisal variables

$$\mu'_\Omega(k) = \mu_\Omega(k) \left( \sum_{i \in D} \mu_D(i)\, |d_i^{(k)}| \right), \tag{12}$$

  where $k \in \Omega$. As a consequence, features with higher weights in the encoding have a higher impact on the evaluation of Similarity. If all features are encoded identically with values $d_i^{(k)} = c_i\ \forall k \in \Omega,\, i \in D$, we resemble the Tversky index on the original feature sets $F$ and $F'$.
- The encoding of a feature in $D \times [0, 1]$ can be visualized by a histogram over the appraisal variables in $D$, as displayed in Fig. 3. The second factor in (8) corresponds to the area of the histogram if the measure $\mu_D$ of the appraisal variables is represented by

8. A similar approach is suggested by [33] and investigated in [37].
9. See Fig. 3 for a visualization of the $\hat{f}^{(k)}$.
10. This setting is different from Equation 4, and hence defines a different kind of weight.
11. For the Cartesian product of sets, the intersection and set difference are given as

$$\begin{aligned} (S_1 \times T_1) \cap (S_2 \times T_2) &= (S_1 \cap S_2) \times (T_1 \cap T_2) \\ (S_1 \times T_1) \setminus (S_2 \times T_2) &= (S_1 \times (T_1 \setminus T_2)) \cup ((S_1 \setminus S_2) \times T_1) \end{aligned}$$

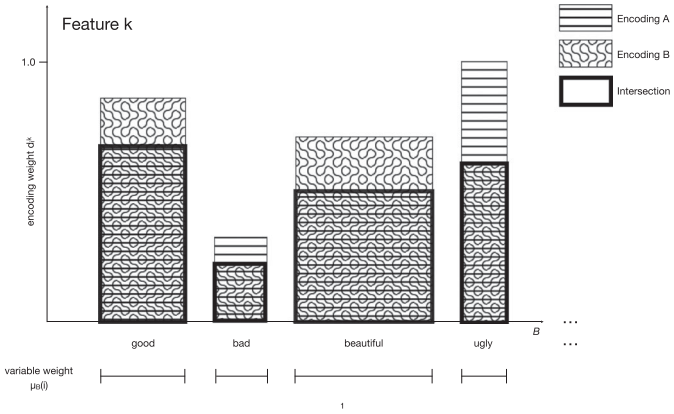


Fig. 3. Visualization of the feature encoding for Similarity calculation.

varying bar widths - more important variables get a larger width. The intersection (set difference) between two encodings A and B of the same feature corresponds to the area where two histogram plots overlap (differ).

- For simplicity, we could assume that the importance of a feature for the evaluation of Similarity does not differ from the importance of a feature in the decision process and is thus already incorporated via the weight values $w^{(k)}$ in (4). In this case, we can simply replace the appraisal weights $d_i^{(k)}$ in the above formulas with the perceived weights $p_i^{(k)}$ and choose the measures $\mu_\Omega$ on the feature space to be uniform.

## 6 Goals and Actions

As stated in Section 4.1, the belief system of the agent has a world model that, based on the current situation and past experience, provides the agent with possible actions and their relations to features,[12] maintains goals[13] of the agent with ambitions to achieve them, it has beliefs that actions affect certain goals and has a sense of realism. In the following, we will give a detailed description of these inputs to the decision process.

As part of this belief system, the agent has a given set $G$ of goal states it wants to achieve or avoid. Each goal $j \in G$ has an associated value $a_j \in [-1, 1]$ that determines the ambition of the agent towards that goal.

Each feature $k \in F$ provides the agent with certain actions $A^{(k)}$. For the decision process, the belief system categorizes each action in $A^{(k)}$ into one of the four action tendencies

$$T = \{positive, negative, change, avoid\}. \tag{13}$$

These action tendencies allow Silicon Coppélia to relate the available actions to her affective state in the response phase (see (43)). Further, the agent has beliefs

$$ag_{ij}^{(k)} \in [-1, 1], \qquad i \in A, j \in G, k \in F, \tag{14}$$

that the choice of the action $i$ for feature $k$ facilitates or inhibits goal $j$. The $ag_{ij}^{(k)}$ can incorporate knowledge about the current situation as well as past experience of the agent, i.e., they can vary for different observations or in different execution cycles of the affective process.

## 7 Comparing for Relevance and Valence

Relevance and Valence express the expected effect of a feature on the goals of the agent. Relevance expresses the intensity of the effect, Valence the direction (positive or negative) of the effect. These two values are obtained by evaluating the beliefs of the agent about actions and goals and by comparing them to the encoding of the features in the Ethics and Affordance domain, as drawn in Fig. 2. To calculate indicative and counter-indicative values for Relevance and Valence of a feature, we evaluate a set of fuzzy rules[14] with conditions based on the available actions and affected goals as well as the encoded values of the appraisal variables for that feature.

Before we formulate these rules, let us give some definitions. Fuzzy logic and control theory provide many choices for the concrete definition of the fuzzy operators used. Our choice here is only motivated by simplicity, see also the Remark 7.3 at the end of this section. For fuzzy-set operations, we use the Zadeh operators. In this case the *AND* operation (intersection) is given as the minimum function

$$\mu_{A\wedge B}(x, y) = \min(\mu_A(x), \mu_B(y)), \quad x \in X, y \in Y, \tag{15}$$

where $A$ and $B$ are fuzzy sets and $X$ and $Y$ are the support spaces of these sets.[15] For the fuzzy *OR* operation, we use the maximum function

$$\mu_{A\vee B}(x, y) = \max(\mu_A(x), \mu_B(y)), \quad x \in X, y \in Y. \tag{16}$$

Fuzzy rules have the form

$$\text{IF}\,(x \in A) \quad \text{THEN} \quad (z \in B), \tag{17}$$

where $x$ and $z$ are the input and output variables of the rule and $A$ and $B$ are fuzzy sets defined on the support spaces $X$ and $Z$ of these variables. We evaluate the rules with the minimum operation as *implication function*, also known as Mamdani-type implication. In particular, the grade of membership to the output set (*rule consequence*) is capped at the grade of membership to the condition (*rule strength*)

$$\mu_{\text{out}}(z) = \min(\mu_A(x), \mu_B(z)). \tag{18}$$

The results of multiple rules are combined with the *OR* operation as aggregation function. To be precise, the highest grade of membership among the rule outputs determines the grade of membership of the final output set.

To be able to formulate the fuzzy rules for Relevance and Valence, we define for each feature $k \in F$, each associated action $i \in A$ and each goal $j \in G$ an indicative and a counter-indicative fuzzy set for the beliefs $ag_{ij}^{(k)}$ and the goal ambitions $a_j$ of the agent

12. This can be a many-to-many relation.

13. Like features, actions and goals are a consequence of the model the belief system has about the situation and can be complex in nature. For example, an action could represent a whole script of individual tasks or even be abstract like *"write a book"*.

14. For an overview on fuzzy control see, for example, [36].

15. For example:

$$\mu_{facilitates\wedge desired}(ag_{ij}^{(k)}, a_j^{(k)}) = min(\mu_{facilitates}(ag_{ij}^{(k)}), \mu_{desired}(a_j^{(k)})),$$

where $ag_{ij}^{(k)} \in [-1, 1], a_j^{(k)} \in [-1, 1]$, see below.

TABLE 1
Rules for Valence of a Specific Feature and Goal (see also [35])

| Judgment | Affordance/Ethics (perceived weight) | Effect (Moderation) | Goal state (Ambition) | Agreement | Weighing | Valence direction |
|---|---|---|---|---|---|---|
| ① | aid/good $^{p}$ (discretion) | facilitates $^{p}$ | desired goal $^{p}$ (self-disclosure) | Agree $^{p}$ | $p \sim p \cdot p \cdot p$ | p |
| | | | | Disagree$^{n}$ | $p \nsim p \cdot p \cdot n$ | n |
| ② | aid/good $^{p}$ (discretion) | inhibits $^{n}$ | desired goal $^{p}$ (self-disclosure) | Agree $^{p}$ | $p \nsim n \cdot p \cdot p$ | n |
| | | | | Disagree$^{n}$ | $p \sim n \cdot p \cdot n$ | p |
| ③ | aid/good $^{p}$ (discretion) | facilitates $^{p}$ | undesired goal $^{n}$ (non-disclosure) | Agree $^{p}$ | $p \nsim p \cdot n \cdot p$ | n |
| | | | | Disagree$^{n}$ | $p \sim p \cdot n \cdot n$ | p |
| ④ | aid/good $^{p}$ (discretion) | inhibits $^{n}$ | undesired goal $^{n}$ (non-disclosure) | Agree $^{p}$ | $p \sim n \cdot n \cdot p$ | p |
| | | | | Disagree$^{n}$ | $p \nsim n \cdot n \cdot n$ | n |
| ⑤ | obstacle/bad $^{n}$ (indiscretion) | facilitates $^{p}$ | desired goal $^{p}$ (self-disclosure) | Agree $^{p}$ | $n \nsim p \cdot p \cdot p$ | p |
| | | | | Disagree$^{n}$ | $n \sim p \cdot p \cdot n$ | n |
| ⑥ | obstacle/bad $^{n}$ (indiscretion) | inhibits $^{n}$ | desired goal $^{p}$ (self-disclosure) | Agree $^{p}$ | $n \sim n \cdot p \cdot p$ | n |
| | | | | Disagree$^{n}$ | $n \nsim n \cdot p \cdot n$ | p |
| ⑦ | obstacle/bad $^{n}$ (indiscretion) | facilitates $^{p}$ | undesired goal $^{n}$ (non-disclosure) | Agree $^{p}$ | $n \sim p \cdot n \cdot p$ | n |
| | | | | Disagree$^{n}$ | $n \nsim p \cdot n \cdot n$ | p |
| ⑧ | obstacle/bad $^{n}$ (indiscretion) | inhibits $^{n}$ | undesired goal $^{n}$ (non-disclosure) | Agree $^{p}$ | $n \nsim n \cdot n \cdot p$ | p |
| | | | | Disagree$^{n}$ | $n \sim n \cdot n \cdot n$ | n |

$$\begin{aligned} \text{Moderation (from } ag_{ij}^{(k)}\text{)}: &\quad \mathit{facilitates},\ \mathit{inhibits} \\ \text{Ambition (from } a_j\text{)}: &\quad \mathit{desired},\ \mathit{undesired} \end{aligned} \tag{19}$$

The rule conditions formulated for Relevance and Valence in the next sections consist of an expression corresponding to an ontological statement about features, actions, goals and their relations (see Table 1), as well as a measure of agreement by the agent to that statement, represented by the fuzzy set

$$\text{Agreement}: \quad \mathit{agree},\ \mathit{disagree}. \tag{20}$$

The Agreement value formed by the agent is complementary to the goal-directed beliefs described above. If someone understands that 'Being shrewd helps to become rich,' one does not necessarily has to agree and execute action upon the by itself 'true' statement. A person may know that 'Eating sugar is bad for my health' yet does it anyways. Thus, Agreement in Table 1 may be based on a sense of realism that the agent assigns to its mental world (cf. [30]), on general knowledge of the agent that contradicts a statement, or that perhaps is rejected via cultural biases. In our model, the Agreement value is an input value obtained from the belief system without being concerned at the moment how it is generated. In our examples, we assume that it matches the goal-driven beliefs to keep things simple.

Each of the above quantities has a range $[-1, 1]$ and we determine the grade of membership to the indicative or counter-indicative set by the sign and absolute value of the quantity as shown in Fig. 4. For instance

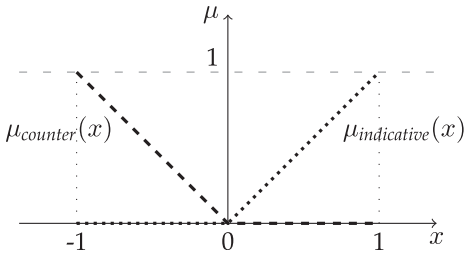


Fig. 4. Fuzzification of parameters with values in $[-1, 1]$.

$$\mu_{desired}(a_j) = \begin{cases} a_j & a_j \geq 0 \\ 0 & a_j < 0 \end{cases}$$

and

$$\mu_{undesired}(a_j) = \begin{cases} 0 & a_j \geq 0 \\ |a_j| & a_j < 0 \end{cases}. \tag{21}$$

### 7.1 Relevance

To determine the Relevance of a feature, we look at whether the agent expects an action associated with the feature to affect an important goal; particularly, whether there is an action that facilitates or inhibits a desired or undesired goal (compare (24)). In other words, Relevance describes how strong the goals of the agent are affected but not whether they are affected in a desired or undesired way.

We express the Relevance of a feature by an indicative ('relevant') and a counter-indicative value ('irrelevant'), which correspond to the grades of membership to the fuzzy (singleton) sets

$$\text{Relevance}: \mathit{relevant},\ \mathit{irrelevant}. \tag{22}$$

We use these two singleton sets also as the outputs of our fuzzy rules, with a prior weight function

$$\mu_S(x) = \begin{cases} 1 & \text{if} \quad x \in S \\ 0 & \text{else} \end{cases}, \tag{23}$$

where $S$ is one of the singleton sets $\mathit{relevant}$ or $\mathit{irrelevant}$. With the help of the unions

$$\begin{aligned} \mathit{affects} &= \mathit{facilitates}\ \textsc{or}\ \mathit{inhibits} \\ \mathit{important} &= \mathit{desired}\ \textsc{or}\ \mathit{undesired}, \end{aligned} \tag{24}$$

of the quantities in (19) we can now formulate a set of fuzzy rules for the Relevance of each feature $k \in F$.

First, let us look at rules based on the conditions that goal $j \in G$ is affected by action $i \in A^{(k)}$ associated with feature $k$:[16]

$$\begin{aligned} \text{IF} \quad & action\, i\, for\, feature\, k\, affects\ goal\, j \\ & \text{AND}\ j\, is\, important \\ & \text{AND}\ agree \\ \text{THEN} \quad & k\, is\, relevant. \end{aligned} \tag{25}$$

We can combine the rules for each goal by combining the rule conditions for the different actions with the fuzzy OR operator, such that we can write a single rule for each goal $j \in G$ affected by feature $k \in F$

$$\begin{aligned} \text{IF} \quad & (\text{ANY}\ action\, i \in A^{(k)}\ for\ feature\, k\, affects\, j) \\ & \text{AND}\ j\, is\, important \\ & \text{AND}\ agree \\ \text{THEN} \quad & k\, is\, relevant, \end{aligned} \tag{26}$$

where we informally expressed the OR operation over the possible actions $i \in A^{(k)}$ by the operator *ANY*.

In complementing every rule (25) for $k \in F$, $i \in A^{(k)}$, and $j \in G$, we also need to consider the $(2^3 - 1)$ rules that treat the negations of the individual terms in the conditions of (25), for instance, the rules

$$\begin{aligned} \text{IF} \quad & \text{NOT}\ (i\ affects\ j) \\ & \text{AND}\ j\, is\, important \\ & \text{AND}\ agree \\ \text{THEN} \quad & k\, is\, irrelevant, \\ \text{IF} \quad & i\ affects\, j \\ & \text{AND NOT}\ (j\, is\, important) \\ & \text{AND}\ agree \\ \text{THEN} \quad & k\, is\, irrelevant, \\ \text{etc.} & \end{aligned} \tag{27}$$

However, only the rules in (26) contribute to being relevant, all other rules result into being irrelevant.

In addition to the rules based on the goals of the agent we also take the appraisal weights of the feature into account. Therefore, we add the simple rules

$$\begin{gathered} \text{IF}\ feature\, k\, is\, (good\ \text{OR}\ bad)\text{THEN}\ k\ is\, relevant, \\ \text{IF}\ feature\, k\, is\, \text{NOT}\ (good\ \text{OR}\ bad)\text{THEN}\ k\ is\, irrelevant, \\ \text{IF}\ feature\, k\, is\, (beautiful\ \text{OR}\ ugly)\text{THEN}\ k\ is\, relevant \\ \text{etc.} \end{gathered} \tag{28}$$

to the rule set for feature $k \in F$.

After evaluating all the rules for feature $k \in F$, the consequences of the rules are combined into an output weight function $\mu^{(k)}$ for the Relevance of feature $k$ by applying the fuzzy OR operation, taking the highest grade of membership among all rule consequences $\mu_r^{(k)}$

$$\mu^{(k)}(x) = \max_{allrulesr} (\mu_r^{(k)}(x)), \tag{29}$$

where $x \in \{relevant, irrelevant\}$.

16. Remember that the quantities in (19) were obtained from the ambition and belief values $ag_{ij}^{(k)}$, $a_j$, and the Agreement values described in (20) are specific to each statement, i.e., depend on $i, j$ and $k$. In total, we obtain here $|G| \cdot |A^{(k)}|$ rules for each feature $k$.

## 7.2 Valence

The fuzzy rules for the calculation of the Valence of each feature $k \in F$ are based on the set of statements listed in Table 1, which is introduced in [35]. Columns 2-4 in the table represent an ontological statement about the feature in terms of Ethics and Affordances, the associated actions and affected goals, while column 5 represents whether this statement is in agreement with the agent's belief system. Columns 3-5 determine the expected direction of Valence (column 7) based on the beliefs and goals of the agent. If, for instance, the agent agrees that an action associated with the feature facilitates a desired goal, it will result in positive valence, or if the goal is undesired in negative valence. The gray cells in Table 1 describe situations where the statement combined with the agents Agreement does not match the encoding of the feature, for instance, if the feature is encoded as an aid, but the associated action inhibits a desired goal. The agent may experience this an internal state of mixed emotions, see [35].

The rules have conditions based on the action $i \in A^{(k)}$ and the goals $j \in G$ associated with $i$ and result in positive or negative Valence. The input quantities are the fuzzy sets obtained in (19) and the fuzzy sets on the values

$$\begin{aligned} \text{Affordance} &: aid,\ obstacle, \\ \text{Ethics} &: good,\ bad, \end{aligned}$$

obtained from the encoding in the Affordance and Ethics domain. The outputs of the rules are again fuzzy (singleton) sets

$$\text{Valence}: \quad valence_{(+)},\ valence_{(-)}. \tag{30}$$

with prior weight function

$$\mu_S(x) = \begin{cases} 1 & \text{if} \quad x \in S \\ 0 & \text{else}, \end{cases} \tag{31}$$

where $S$ is one of the singleton sets $valence_{(+)}$ or $valence_{(-)}$.

In terms of these quantities, the rules in Table 1 for feature $k \in F$, the action $i \in A^{(k)}$, and an affected goal $j \in G$ take the form

$$\begin{aligned} \text{IF} \quad & feature\, k\, is\, aid \\ & \text{AND}\ action\, i\, for\, k\, facilitates\, goal\, j \\ & \text{AND}\ j\, is\, desired \\ & \text{AND}\ agree \\ \text{THEN} \quad & k\, has\, valence_{(+)}, \end{aligned} \tag{32}$$

where this example corresponds to the first rule of Table 1.

The rule consequences of the rules in Table 1 can be determined by the signs of the belief, ambition, and agreement values

$$rule\, consequence = \begin{cases} valence_{(+)} & \text{if}\ \operatorname{sgn}(ag_{ij}^{(k)}\, a_j^{(k)}\, \beta_{ij}^{(k)}) = 1 \\ valence_{(-)} & \text{if}\ \operatorname{sgn}(ag_{ij}^{(k)}\, a_j^{(k)}\, \beta_{ij}^{(k)}) = -1, \end{cases} \tag{33}$$

where $\beta_{ij}^{(k)} = 1$ if the agent *agrees* with the statement indexed by $i, j$ and $k$, and $\beta_{ij}^{(k)} = -1$ if it *disagrees*. With the above

definitions, only the weight functions of the indicative and counter-indicative fuzzy sets representing Affordances and Ethics can be both non-null at the same time, which limits the number of effective rules to two per affected goal, feature and appraisal domain.

Finally, the results of the rules are again combined to an output weight function $\mu^{(k)}$ for the Valence of a feature by the fuzzy OR operation

$$\mu^{(k)}(x) = \max_{allrulesr}(\mu_r^{(k)}(x)), \tag{34}$$

where $x \in \{valence_{(+)}, valence_{(-)}\}$.

Note 5-8 mirrors that of 1-4. However, the source of the same final judgment is different. If an agency would like to discuss the motivation of a decision, the conversation about 4 is about the discretion an agency provides during conversation and the consequences thereof, whereas in 8, it is about the consequences of indiscretion. Also, column 2 still influences the rule strength, i.e., the results of the rules for judgment 5-8 can be different from those of judgment 1-4.

### 7.3 Remarks on Goals and Actions

- The choice of representing goal ambitions $a_j$ and the beliefs $ag_{ij}^{(k)}$ by values from the interval $[-1, 1]$ could be replaced by directly representing these quantities as fuzzy values. This choice would make the fuzzification process described in Equation (21) obsolete.
- As stated already at the beginning of this section, the fuzzy operators AND/OR and the implication and aggregation functions used in the rules can be implemented in many ways (see e.g., [36]). The reason for the choices made here is mere simplicity.

## 8 Responding: Use Intentions

The intentions of the agent to make use of a certain feature of the other agency are on the one hand a consequence of the Utility of the feature, where Utility is an intermediate variable that represents how useful a feature is for the agent with respect to its goals. It can be seen as a combination of Relevance and Valence: A highly relevant feature with high positive Valence will be highly useful for the agent and vice versa. On the other hand, Use Intention is influenced by the perceived values of the agent in the Aesthetics domain as well as by the Similarity of the agent to another agent that it judges, as depicted in Fig. 2.

### 8.1 Utilities

The agent can take action to achieve desired goal states or to avoid undesired ones. To determine the expected Utility of each action, we first look at the goals that are affected by these actions. Following [27] and [35],[17] we determined Utility by the action that is expected to achieve in total the most valuable goals for that feature. First, we define the expected Utility for action $i \in A^{(k)}$ with respect to goal $j \in G$ by the product of the belief that the goal will be achieved as a consequence of that action and the agent's ambition towards that goal

$$u_{ij}^{(k)} = ag_{ij}^{(k)}\, a_j, \tag{35}$$

where $i \in A^{(k)}$, $j \in G$, $k \in F$ and $u_{ij}^{(k)} \in [-1, 1]$.

The general expected Utility of an action can then be obtained by combining the Utility values of all goals affected by that action. Combination is done by taking the average with respect to a (generated[18]) weight distribution $\beta_{gu}$ over the expected utilities

$$\bar{u}_i^{(k)} = {}_{\beta_{gu}}\langle u_{ij}^{(k)}\rangle_{j\in G} \qquad i \in A^{(k)}, k \in F. \tag{36}$$

We also calculate the indicative (+) and counter-indicative (-) Utility values of feature $k \in F$. These are determined by the Utility values of the actions with the highest and lowest Utility values

$$u_{(+)}^{(k)} = \max_{i\in A}(\bar{u}_i^{(k)}), \qquad u_{(-)}^{(k)} = \max_{i\in A}(-\bar{u}_i^{(k)}). \tag{37}$$

If $u_{(+)}^{(k)}$ or $u_{(-)}^{(k)}$ would become negative by this definition, it is set to zero.

### 8.2 Remarks on Utility

- In [27], the algorithm described in Section 6.2 in [27] is used to average over values. It corresponds to assigning weights $\sim \frac{2^{k-1}}{2^{N_{i/c}}}$ to the ordered indicative/counter-indicative values (where $k$ is the index in the sorted indicative/counter-indicative list of values, $N_{i/c}$ the respectively largest index).

### 8.3 Calculating Use Intentions

As shown in Fig. 2, Use Intentions for a feature are, besides Utilities, influenced by Aesthetics and Similarity. Curved arrows in Fig. 2 symbolize that the variable from which they originate only has an indirect influence (interaction) on the destination variable via the relation they point to (cf. [4]). For instance, Aesthetics on its own is not enough to create Use Intentions for a feature, but only influences the effect of Utilities on Use Intentions. Again following [27], we model this interaction effect by using a linear model with interaction terms between Utilities and Aesthetics respectively Similarity:[19]

The indicative and counter-indicative Utility values of a feature $k \in F$ provide the first two components of the input vector to this linear model

$$x_1^{(k)} = u_{(+)}^{(k)} \qquad x_2^{(k)} = u_{(-)}^{(k)}. \tag{38}$$

The interaction terms are given by the following components of the input vector:

17. Here we derive Utility directly from beliefs about goals and actions. To find a formulation that derives Utility directly from Relevance and Valence or makes the use of this intermediate variable obsolete is the goal of future work.

18. [27] uses the algorithm described in Section 6.2 in [27] to calculate this average, see Remark 8.2.

19. This model can be seen as the lowest order approximation of the function that determines Use Intentions from Utility, Aesthetics and Similarity.

$$
\begin{aligned}
x_3^{(k)} &= sim^{(k)}\, u_{(+)}^{(k)} & x_4^{(k)} &= dis^{(k)}\, u_{(+)}^{(k)} \\
x_5^{(k)} &= sim^{(k)}\, u_{(-)}^{(k)} & x_6^{(k)} &= dis^{(k)}\, u_{(-)}^{(k)} \\
x_7^{(k)} &= p_{beautiful}^{(k)}\, u_{(+)}^{(k)} & x_8^{(k)} &= p_{ugly}^{(k)}\, u_{(+)}^{(k)} \\
x_9^{(k)} &= p_{beautiful}^{(k)}\, u_{(-)}^{(k)} & x_{10}^{(k)} &= p_{ugly}^{(k)}\, u_{(-)}^{(k)},
\end{aligned}
$$

and the result is obtained by multiplication with a weight matrix $\mathbf{B}_{ui}$

$$\vec{y}^{(k)} = \vec{x}^{(k)} \cdot \mathbf{B}_{ui} \qquad k \in F, \tag{39}$$

where the dimensions of the weight matrix $\mathbf{B}_{ui}$ is $2 \times 10$ and the result $\vec{y}^{(k)} = (ui_{(+)}^{(k)}, ui_{(-)}^{(k)})$ contains the positive and negative Use Intentions.

The weight matrix $\mathbf{B}_{ui}$ is a parameter to our model and determines the influence the components of $\vec{x}^{(k)}$ have on the resulting Use Intentions.

## 9 Responding: Involvement Parallel to Distance

As illustrated in Fig. 2, Involvement and Distance result from the appraisal domains Affordances, Aesthetics, and Epistemics, from the calculated Relevance and Valence and are influenced by interaction terms between Relevance/Valence and Aesthetics/Similarity, corresponding to the curved arrows in Fig. 2. As for Use Intentions, we use a linear model to compute Involvement and Distance where the components of the input vector $\vec{x}^{(k)}$ are given by the following terms (see also [27], Section 6.5, Table 1 of that paper):

$$
\begin{aligned}
x_1^{(k)} &= aest_{(+)}^{(k)} & x_2^{(k)} &= aest_{(-)}^{(k)} \\
x_3^{(k)} &= ep_{(+)}^{(k)} & x_4^{(k)} &= ep_{(-)}^{(k)} \\
x_5^{(k)} &= aff_{(+)}^{(k)} & x_6^{(k)} &= aff_{(-)}^{(k)} \\
x_7^{(k)} &= rel^{(k)} & x_8^{(k)} &= irr^{(k)} \\
x_9^{(k)} &= val_{(+)}^{(k)} & x_{10}^{(k)} &= val_{(-)}^{(k)} \\
x_{11}^{(k)} &= rel^{(k)}\, aest_{(+)}^{(k)} & x_{12}^{(k)} &= rel^{(k)}\, aest_{(-)}^{(k)} \\
x_{13}^{(k)} &= irr^{(k)}\, aest_{(+)}^{(k)} & x_{14}^{(k)} &= irr^{(k)}\, aest_{(-)}^{(k)} \\
x_{15}^{(k)} &= rel^{(k)}\, sim^{(k)} & x_{16}^{(k)} &= rel^{(k)}\, dis^{(k)} \\
x_{17}^{(k)} &= irr^{(k)}\, sim^{(k)} & x_{18}^{(k)} &= irr^{(k)}\, dis^{(k)} \\
x_{19}^{(k)} &= val_{(+)}^{(k)}\, aest_{(+)}^{(k)} & x_{20}^{(k)} &= val_{(+)}^{(k)}\, aest_{(-)}^{(k)} \\
x_{21}^{(k)} &= val_{(-)}^{(k)}\, aest_{(+)}^{(k)} & x_{22}^{(k)} &= val_{(-)}^{(k)}\, aest_{(-)}^{(k)} \\
x_{23}^{(k)} &= val_{(+)}^{(k)}\, sim^{(k)} & x_{24}^{(k)} &= val_{(+)}^{(k)}\, dis^{(k)} \\
x_{25}^{(k)} &= val_{(-)}^{(k)}\, sim^{(k)} & x_{26}^{(k)} &= val_{(-)}^{(k)}\, dis^{(k)}.
\end{aligned}
$$

Again, the result is obtained by multiplication with a weight matrix parameter $\mathbf{B}_{id}$

$$\vec{y}^{(k)} = \vec{x}^{(k)} \cdot \mathbf{B}_{id} \qquad k \in F, \tag{40}$$

where the dimension of $\mathbf{B}_{id}$ is $2 \times 26$ and the result $\vec{y}^{(k)} = (inv^{(k)}, dist^{(k)})$ contains the Involvement and Distance results.

Involvement and Distance are combined into the Involvement-Distance Trade-off (IDT) using a *fuzzy-or* operator (see e.g., [38]), as motivated in [39] and [40]. This operator allows for compensation between the two quantities, i.e., Distance may be partly compensated by Involvement and vice versa, and reads as

$$idt^{(k)} = \beta_{idt} \max(inv^{(k)}, dist^{(k)}) + (1 - \beta_{idt}) \frac{inv^{(k)} + dist^{(k)}}{2}, \tag{41}$$

with $k \in F$ and a weight parameter $\beta_{idt}$ from the agent's belief system.

## 10 Satisfaction

By the end of the affective decision-making process, the agent takes the most valuable feature and chooses the best action related to this feature. This selection is based on the Satisfaction expected from the features, which is the output in Fig. 2. The indicative and counter-indicative weights for Use Intentions are reduced to a single value by taking their average $ui^{(k)} = 0.5\,(ui_{(+)}^{(k)} + ui_{(-)}^{(k)})$. Then the Satisfaction value $s^{(k)}$ is calculated as the weighted mean of IDT and Use Intentions $ui^{(k)}$ for every feature, with weights $\beta_{idt}^{(s)}$ and $\beta_{ui}^{(s)}$, as described in [35]

$$s^{(k)} = \beta_{s,idt}\, idt^{(k)} + \beta_{s,ui}\, ui^{(k)} \qquad k \in F. \tag{42}$$

The final action $i_{\max}$ will be selected from the actions $i \in A^{(k)}$ associated with the chosen feature $k_{\max}$. To perform this selection, we calculate Satisfaction values $s_i^{(k)}$ for each action, which depend on the action tendency $t(i) \in T$ associated with action $i$ (see (13)), as described in [35]

$$
s_i^{(k)} = \begin{cases}
\beta_{s,i,p} inv^{(k)} + \beta_{s,d,p}(1 - dist^{(k)}) + \beta_{s,u,p}\bar{u}_i^{(k)}, & t(i) = \text{p.}, \\
\beta_{s,i,n}(1 - inv^{(k)}) + \beta_{s,d,n} dist^{(k)} + \beta_{s,u,n}\bar{u}_i^{(k)}, & t(i) = \text{n.}, \\
\beta_{s,i,c} inv^{(k)} + \beta_{s,d,c} dist^{(k)} + \beta_{s,u,c}\bar{u}_i^{(k)}, & t(i) = \text{c.}, \\
\beta_{s,i,a}(1 - inv^{(k)}) + \beta_{s,d,a} dist^{(k)} + \beta_{s,u,a}\bar{u}_i^{(k)}, & t(i) = \text{a.},
\end{cases} \tag{43}
$$

where the used weight parameters $\beta_{s,*,*}$ are provided by the belief system. Satisfaction values are based on Involvement and Distance for the feature $k$ as well as the utilities $\bar{u}_i^{(k)}$ for the individual actions and utilize the connection between the action tendency of an action and the Engagement of the agent. If the Involvement of the agent is high, it tends towards acting positively, if the agent feels Distant, it will tend towards acting negatively. The tendency towards an action that leads to a change is high if both Involvement and Distance are high, and the choice between acting in a negative way compared to an avoiding action depends on the character of the agent, i.e., the weights $\beta_{s,*,d}$ compared to $\beta_{s,*,a}$. The final action chosen will be the one with the highest Satisfaction value

$$i_{\max} = \underset{i \in A^{(k)}}{\arg\max}(s_i^{(k)}). \tag{44}$$

## 11 Example

Let us see, then, how the model behaves when it is in operation. In the following demonstration, Silicon Coppélia

impersonates a plastic surgeon (Alice) who feels lonely and faces an other agency (Bob), who may have a medical condition that falls within her area of expertise. In this limited context, Coppélia has but two goals in life: To cure people through medical treatment and to find an attractive dating partner. For reference we provide the full numerical details for each step in this example, these can be skimmed on a first reading.

### 11.1 Alice Encodes Features of Bob

In this example, we will model the involved agencies by their looks and the eventual condition they suffer from. In particular, we will define the following sets:

- *gender*: For gender we restrict ourselves to
$$\Omega_{gender} = \{male, female\}.$$
- *age*: For age we use $\Omega_{age} = \mathbb{N}$ and define 5 overlapping categories, represented by the intervals
$$\begin{aligned} child: &\quad 0-14,\\ adolescent: &\quad 12-21,\\ young: &\quad 18-33,\\ mature: &\quad 28-55,\\ old: &\quad 50-99. \end{aligned}$$
- *attractiveness*: We consider 5 levels of attractiveness, which we represent by a sequence of inclusive sets to encode the ordering of these levels
$$ugly \subset unsightly \subset average \subset attractive \subset beautiful,$$
which makes the set of possible attractiveness features simply the largest element in the above sequence, i.e., $\Omega_{attr} = beautiful$.
- *condition*: We will consider only a single medical condition (i.e., a mole perhaps caused by basal cell carcinoma)
$$\Omega_{condition} = \{mole\}.$$

Using these definitions, the space of possible features for each agent becomes

$$\Omega \subset \Omega_{gender} \cup \Omega_{age} \cup \Omega_{attr} \cup \Omega_{condition}, \tag{45}$$

and let Alice (A) and Bob (B) be represented by the following features:

$$\begin{aligned} F_A &= \{female\} \cup mature \cup attractive = f_1^A,\\ F_B &= (\{male\} \cup young \cup beautiful) \cup \{mole\} = f_1^B \cup f_2^B. \end{aligned} \tag{46}$$

In Equation (46), we assume that the belief system evaluates all features related to the looks of the agent identically and makes use of the remark at the end of Section 4.1, combining all those features into a single feature $f_1$. For consistency, we also denote the feature representing the medical condition of an agent with $f_2$.

Let us now determine the encoding Alice assigns to Bob's features. From a moral perspective, Alice thinks that Bob is a proper dating partner, but, though she is looking for a young beautiful man, she is a bit concerned about the age difference as he looks pretty young. Bob is surely a good dating partner, but he is so highly attractive that Alice is afraid that keeping him satisfied will be an obstacle, though not a big one. He is clearly beautiful, which also gives her at the same time the feeling of being in a fairy tale. She thinks he is in need of help with his mole; for this reason she encodes the mole mostly as harmless, though, as it could be pathological, it has a touch of being bad. The mole is something she can treat, but what seems to be a slight obstacle to her is that by her professionalism she would not date a patient. She is tempted to look at Bob's mole as a beauty spot, but for that it is just a bit too suspect. Then Alice's encoding of the features of Bob may look as in Table 2. Let us also assume that Alice is at work, thus she is focused more on the medical condition than on the looks of the other. We define the feature weights $w^{(f_1)} = 0.5$ and $w^{(f_2)} = 1.0$. The resulting affordance and perceived weights are given together with the affordance weights Alice has about herself in Table 2.

Note that the information contained in the encoding in Table 2 again is a result of the world model of Alice's belief system. As such, the description in the paragraph above, which tries to give an example of how the belief system might derive these values, contains many assumptions that could result from past experience or be just plain imagination of Alice.

### 11.2 Similarity Between Alice and Bob

To evaluate the Similarity between the two agencies, let us assume that Alice encodes her own features, which is only $f_1^A$ in this case, with the appraisal weights given in Table 2. Then the features augmented with their encoding (see (6)) will take the form[20]

$$\begin{aligned} \hat{f}_1^B &= f_1^B \times \{(eth_{(+)}, [0.0, 0.8]),\ (eth_{(-)}, [0.0, 0.4]),\\ &\qquad (aff_{(+)}, [0.0, 0.8]),\ (aff_{(-)}, [0.0, 0.4]),\ \ldots\},\\ \hat{f}_2^B &= f_2^B \times \{(eth_{(+)}, [0.0, 0.8]),\ (eth_{(-)}, [0.0, 0.2]),\\ &\qquad (aff_{(+)}, [0.0, 0.8]),\ (aff_{(-)}, [0.0, 0.2]),\ \ldots\},\\ \hat{f}_1^A &= f_1^A \times \{(eth_{(+)}, [0.0, 0.8]),\ (eth_{(-)}, [0.0, 0.0]),\\ &\qquad (aff_{(+)}, [0.0, 0.8]),\ (aff_{(-)}, [0.0, 0.2]),\ \ldots\}, \end{aligned} \tag{47}$$

where for convenience, we denote with $A \times \{(b, c) \mid c \in C\}$ the set[21] $\{(a, b, c) \mid a \in A,\ c \in C\}$, and

$$\widehat{F}^B = \hat{f}_1^B \cup \hat{f}_2^B, \qquad \widehat{F}^A = \hat{f}_1^A. \tag{48}$$

For the evaluation of Similarity and Dissimilarity, we will treat the distinct features of Alice and Bob symmetrically and set the parameters $\alpha$ and $\beta$ in (9) to 1.0 such that (5) becomes

20. Note that we use the unweighted appraisal weights $b_i^{(k)}$ for similarity calculation.

21. In strict notation $A \times \{(b, c) \mid C \in C\}$ denotes the set $\{(a, (b, c)) \mid a \in A,\ c \in C\}$, i.e., here we identify the tuples $(a, (b, c)) \equiv (a, b, c)$.

TABLE 2
Perceived Feature Encoding

| | Bob | | Alice |
|---|---|---|---|
| | young beautiful man | mole | mature attractive woman |
| | $w^{(k)} \cdot d_i^{(k)} = p_i^{(k)}$ | $d_i^{(k)} = p_i^{(k)}$ | $d_i^{(k)}$ |
| $eth_{(+)}$ | $0.5 \cdot 0.8 = 0.4$ | 0.8 | 0.8 |
| $eth_{(-)}$ | $0.5 \cdot 0.4 = 0.2$ | 0.2 | 0.0 |
| $aff_{(+)}$ | $0.5 \cdot 0.8 = 0.4$ | 0.8 | 0.8 |
| $aff_{(-)}$ | $0.5 \cdot 0.4 = 0.2$ | 0.2 | 0.2 |
| $aest_{(+)}$ | $0.5 \cdot 1.0 = 0.5$ | 0.2 | 0.8 |
| $aest_{(-)}$ | $0.5 \cdot 0.0 = 0.0$ | 0.8 | 0.0 |
| $ep_{(+)}$ | $0.5 \cdot 0.8 = 0.4$ | 1.0 | 1.0 |
| $ep_{(-)}$ | $0.5 \cdot 0.2 = 0.1$ | 0.0 | 0.0 |

$$\begin{aligned} sim &= 1 - dis \\ &= \frac{\mu(\widehat{F}^A \cap \widehat{F}^B)}{\mu(\widehat{F}^A \cap \widehat{F}^B) + \mu(\widehat{F}^A \backslash \widehat{F}^B) + \mu(\widehat{F}^B \backslash \widehat{F}^A)}. \end{aligned} \tag{49}$$

The feature set of Alice has nothing that is related to a medical condition, so $F^A$ and $F^B$ only intersect on $f_1^A$ and $f_1^B$. With

$$f_1^{A \cap B} := f_1^A \cap f_1^B = [28, 33] \cup attractive, \tag{50}$$

we get[22]

$$\begin{aligned} \widehat{F}^A \cap \widehat{F}^B &= \hat{f}_1^A \cap \hat{f}_1^B \\ &= f_1^{A \cap B} \times \big\{(eth_{(+)}, [0.0, 0.8]),\ (eth_{(-)}, [0.0, 0.0]), \\ &\qquad (aff_{(+)}, [0.0, 0.8]),\ (aff_{(-)}, [0.0, 0.2]),\ \ldots\big\}. \end{aligned} \tag{51}$$

The feature sets differ both in the looks of the agencies as well as in the medical condition that Bob has but Alice does not. Let us now examine the set differences involved in the calculation of Similarity: The distinctive features of Bob compared to Alice are given as

$$\begin{aligned} F^B \backslash F^A &= f_1^B \backslash f_1^A \cup f_2^B = f_1^{B \backslash A} \cup f_2^B \\ &= (\{male\} \cup [18, 27] \cup (beautiful \setminus attractive)) \cup \{mole\}, \end{aligned}$$

where we again use the notation $f_1^{B \backslash A} := f_1^B \backslash f_1^A$, and we get[23]

$$\begin{aligned} \widehat{F}^B \backslash \widehat{F}^A &= \hat{f}_1^B \backslash \hat{f}_1^A \cup \hat{f}_2^B \\ &= f_1^{B \backslash A} \times \big\{(eth_{(+)}, [0.0, 0.0]),\ (eth_{(-)}, [0.0, 0.4]), \\ &\qquad (aff_{(+)}, [0.0, 0.0]),\ (aff_{(-)}, [0.2, 0.4]),\ \ldots\big\} \\ &\quad \cup f_2^B \times \big\{(eth_{(+)}, [0.0, 0.8]),\ (eth_{(-)}, [0.0, 0.2]), \\ &\qquad (aff_{(+)}, [0.0, 0.8]),\ (aff_{(-)}, [0.0, 0.2]),\ \ldots\big\}. \end{aligned} \tag{52}$$

On the contrary, for the features exclusive to Alice we get

$$F^A \backslash F^B = f_1^A \backslash f_1^B = f_1^{A \backslash B} = \{female\} \cup [34, 55], \tag{53}$$

and[24]

$$\begin{aligned} \widehat{F}^A \backslash \widehat{F}^B &= \hat{f}_1^A \backslash \hat{f}_1^B \\ &= f_1^{A \backslash B} \times \big\{(eth_{(+)}, [0.0, 0.0]),\ (eth_{(-)}, [0.0, 0.0]),\ \ldots \\ &\qquad (ep_{(+)}, [0.8, 1.0]),\ (ep_{(-)}, [0.0, 0.0])\big\}. \end{aligned} \tag{54}$$

We choose the weight measures $\mu$ of the features in a way that results in balanced values between the different subsets that can appear in the terms of (49).[25] In addition, let us assume that Alice also does not give the medical condition of the agent much consideration when comparing for similarity. This leads to the following choices for the weight measures:

$$\begin{aligned} \mu(male) = \mu(female) = 1.0, \\ \ldots = \mu([18, 21]) = \mu([22, 27]) = \mu([28, 33]) \\ = \mu([34, 47]) = \mu([34, 49]) = \mu([50, 55]) = \ldots = {}^1\!/_3, \\ \mu(ugly) = {}^1\!/_5,\ \mu(unsightly) = {}^2\!/_5,\ \ldots,\ \mu(beautiful) = {}^5\!/_5, \\ \mu(mole) = {}^1\!/_5. \end{aligned} \tag{55}$$

Using equal weights for the appraisal variables, i.e., $\mu(i) = 1.0$, $i \in D$, and the weights from (55), we can finally evaluate the weights for the sets in (51), (52), (53), and (54), according to (8) as

$$\begin{aligned} \mu(\widehat{F}^A \cap \widehat{F}^B) &= ({}^1\!/_3 + {}^4\!/_5) \cdot 3.4 = {}^{289}\!/_{75}, \\ \mu(\widehat{F}^B \backslash \widehat{F}^A) &= (1 + {}^1\!/_3 + {}^1\!/_5) \cdot 1.0 + {}^1\!/_5 \cdot 4.0 = {}^{150}\!/_{75} \\ \mu(\widehat{F}^A \backslash \widehat{F}^B) &= (1 + {}^1\!/_3) \cdot 0.2 = {}^{20}\!/_{75}. \end{aligned} \tag{56}$$

Putting these results and (55) together gives us that Alice assigns a value of $sim = 1 - dis = {}^{289}\!/_{459} \approx {}^2\!/_3$ for her Similarity (resp. Dissimilarity) to Bob.

### 11.3 Alice's Goals and Actions

Next, let us assume that the goals of Alice are to find an attractive dating partner and to cure people with a medical condition

$$G = \{date, cure\}, \tag{57}$$

where the labels represent the following specific goals:

1) *date*: Date *young beautiful men*,
2) *cure*: Cure *men* with a medical condition.

Further, we set the ambition for each goal to

$$a_{date} = 0.5 \qquad \text{and} \qquad a_{cure} = 0.75. \tag{58}$$

22. Again we use the simplified notation from footnote 21.
23. Again we use the simplified notation from footnote 21.
24. Again we use the simplified notation from footnote 21.
25. In particular, we set the weights of the largest subsets that can result from the set intersections or differences for each of the feature domains *gender*, *age*, *attractiveness* and *condition* to 1.0. E.g., for *age* the only subsets that can appear in (49) are the intervals defined by the age categories listed in Section 11.1, the intervals on which these intersect, the set differences between two adjacent categories and the empty set. For *age* we balance the weights by assigning a weight of $^1\!/_3$ to each intersection interval and $^1\!/_3$ to each interval that is exclusive for an age category (cf. (55)). This gives a maximum weight of 1.0 e.g., if both agencies fall into the same age category. Also note that the weight measure $\mu$ does not need to be normalized as (49) already delivers a relative value.

TABLE 3
Action Tendencies and Beliefs $ag_{ij}^{(k)}$

| feature | action | tendency | beliefs $ag_{ij}$ | |
|---|---|---|---|---|
| | | | date | cure |
| young beautiful man ($f_1^A$) | refuse | negative | $-1$ | $1/2$ |
| | invite | positive | $3/4$ | $-1/2$ |
| | skip | avoid | 0 | 0 |
| mole ($f_2^A$) | treat | positive | $-1/2$ | 1 |
| | reject | avoid | $1/2$ | $-1/4$ |

Alice can now pick from a repertoire of actions

$$\begin{aligned} A^{(f_1)} &= \{invite, refuse, skip\}, \\ A^{(f_2)} &= \{treat, reject\}, \end{aligned} \tag{59}$$

where *invite* stands for making a dating appointment, *refuse* for disapproval and *skip* for postponing the decision, i.e., acting neutral. These actions are categorized as action tendencies, according to (13), which is shown in Table 3. Moreover, Alice assigns to each action a belief $ag_{ij}^{(k)}$ that it will facilitate the respective goal.

For the last two columns of Table 3, Alice believes that *refuse* clearly inhibits her date with Bob, while it allows her to treat his mole (remember that she does not date her patients). On the other hand, *invite* clearly facilitates her goal to date Bob (he still has to agree too), but then she will not be able to take him as a patient anymore. Analogous, to *treat* his condition will cure him, but inhibit to date him, and, on the opposite, to *reject* treatment will allow her to date Bob, but not cure him (though he can still receive treatment from someone else). Note that again, these values are based on Alice's belief system and contain knowledge and assumptions about the situation. For instance, her belief that *invite* facilitates her goal to *date* encodes Alice's preference for *young beautiful men*. If she prefers *mature attractive men*, this value should be less.

### 11.4 Relevance of Bob to Alice

To determine the Relevance of the features of Bob, Alice first calculates the importance of her goals, simply by taking the absolute value of the corresponding goal ambitions

$$important_{date} = 0.5 \quad \text{and} \quad important_{cure} = 0.75. \tag{60}$$

To calculate whether an action affects a goal, Alice determines in Table 3 whether an action inhibits or facilitates a goal:

To evaluate the conditions given by rule (26) for the relations between Bob's features and Alice's goals, the following statements apply:

1) any action for *young beautiful man* affects *date* and *date* is *important* and *agree*
2) any action for *young beautiful men* affects *cure* and *cure* is *important* and *agree*
3) any action for *mole* affects *date* and *date* is *important* and *agree*
4) any action for *mole* affects *cure* and *cure* is *important* and *agree*

For simplicity's sake, Alice agrees with all of these statements, that is, all *agree* values equal 1 and *disagree* values equal 0. Then the rule strengths equal the grade of membership to Relevance and are given by

$$\begin{aligned} r_1 &= \min(1.0, 0.5, 1.0) = 0.5, \\ r_2 &= \min(0.5, 0.75, 1.0) = 0.5, \\ r_3 &= \min(0.5, 0.5, 1.0) = 0.5, \\ r_4 &= \min(1.0, 0.75, 1.0) = 0.75. \end{aligned} \tag{61}$$

In taking the maximum over the features, we get as final result for the Relevance of Bob's features to Alice

$$relevant^{(f_1)} = 0.5 \text{ and } relevant^{(f_2)} = 0.75. \tag{62}$$

Irrelevance of Bob's features is in turn determined by rules with conditions that negate[26] the 'will affect' or 'is important' statement, and the rules with the maximum rule consequence are in this case:

1) for *young beautiful man*, *skip* does not affect *date* and *cure* is *important* and *agree*,
2) for *mole*, *reject* does not affect *cure* and *cure* is *important* and *agree*

which in turn result in the values for being irrelevant

$$\begin{aligned} irrelevant^{(f_1)} &= \min(1.0, 0.5, 1.0) = 0.5, \\ irrelevant^{(f_2)} &= \min(0.75, 0.75, 1.0) = 0.75. \end{aligned} \tag{63}$$

### 11.5 Alice Estimates Valence

To calculate Valence, let us look at Table 1. In our case, the encoded weights for Affordance and Ethics are equal, thus we can restrict ourselves to evaluate the Affordance values.[27] Further, for both features the indicative affordance weight (*aid*) is higher than the counter-indicative (*obstacle*), which means that judgments 1-4 will dominate the results. Further, the agent has no undesired goals, which leaves us with judgments 1 and 2. For the sake of simplicity, Alice fully agrees to all statements, and restricting ourselves to the actions with non-zero Moderation values, we are left with the following eight statements to evaluate:

1) agree that *young beautiful man* is an aid and *invite* facilitates desired goal *date*
2) agree that *young beautiful men* is an aid and *refuse* inhibits desired goal *date*
3) agree that *young beautiful man* is an aid and *invite* inhibits desired goal *cure*
4) agree that *young beautiful man* is an aid and *refuse* facilitates desired goal *cure*
5) agree that *mole* is an aid and *treat* inhibits desired goal *date*
6) agree that *mole* is an aid and *reject* facilitates desired goal *date*
7) agree that *mole* is an aid and *treat* facilitates desired goal *cure*

26. For the negation of a fuzzy set, the grade of membership is given by $\mu(\bar{A}) = 1 - \mu(A)$.
27. For Ethics, we will simply obtain identical values.

TABLE 4
Grades of Membership by Feature, Action, and Goal

Goal: date

| feature | action (tendency) | facil | inhib | affects |
|---|---|---|---|---|
| young beautiful man ($f_1^A$) | refuse (neg) | 0 | 1 | 1 |
| | invite (pos) | $3/4$ | 0 | $3/4$ |
| | skip (av) | 0 | 0 | 0 |
| mole ($f_2^A$) | treat (pos) | 0 | $1/2$ | $1/2$ |
| | reject (av) | $1/2$ | 0 | $1/2$ |

Goal: cure

| feature | action (tendency) | facil | inhib | affects |
|---|---|---|---|---|
| young beautiful man ($f_1^A$) | refuse (neg) | $1/2$ | 0 | $1/2$ |
| | invite (pos) | 0 | $1/2$ | $1/2$ |
| | skip (av) | 0 | 0 | 0 |
| mole ($f_2^A$) | treat (pos) | 1 | 0 | 1 |
| | reject (av) | 0 | $1/4$ | $1/4$ |

8) agree that *mole* is an aid
and *reject* inhibits desired goal *cure*

Table 4, rule 1 - 4 pertain to feature *young beautiful man* and rule 5 - 8 to feature *mole*. The rule strengths for these rules and their output sets are given by (33)

$$\begin{aligned}
r_1&=\min(1.0,0.4,0.75,0.5)=0.4, \quad \text{out}_1=valence_{(+)}\\
r_2&=\min(1.0,0.4,1.0,0.5)=0.4, \quad \text{out}_2=valence_{(-)}\\
r_3&=\min(1.0,0.4,0.5,0.75)=0.4, \quad \text{out}_3=valence_{(-)}\\
r_4&=\min(1.0,0.4,0.5,0.75)=0.4, \quad \text{out}_4=valence_{(+)}\\
r_5&=\min(1.0,0.8,0.5,0.5)=0.5, \quad \text{out}_5=valence_{(-)}\\
r_6&=\min(1.0,0.8,0.5,0.5)=0.5, \quad \text{out}_6=valence_{(+)}\\
r_7&=\min(1.0,0.8,1.0,0.75)=0.75, \quad \text{out}_7=valence_{(+)}\\
r_8&=\min(1.0,0.8,0.25,0.75)=0.25, \quad \text{out}_8=valence_{(-)}.
\end{aligned}$$

In taking the maximum for each feature and for each of the indicative and counter-indicative Valence values, Alice comes up with the end result that

$$\begin{aligned}
valence_{(+)}^{(f_1)}&=0.4, \quad valence_{(-)}^{(f_1)}=0.4,\\
valence_{(+)}^{(f_2)}&=0.75, \quad valence_{(-)}^{(f_2)}=0.5.
\end{aligned} \tag{64}$$

### 11.6 Alice's Use Intentions for Bob

The general utilities $\bar{u}^{(k)}$ that Alice expects from her actions as related to Bob's features are calculated as the average over the expected utilities for the affected goals, which produces Table 5:

In Table 5, to keep things simple, Alice again takes the average for the general utilities (cf. (36)) with equal weights for all entries in a row. The actions that provide Alice maximum Utility are *treat* for Bob's feature of the *mole* and *refuse* for his looks. Alice then takes the maximum and minimum over the general utilities for each of Bob's features as described in (37) to obtain

$$u_{(+)}^{(f_1^B)}=0, u_{(-)}^{(f_!^B)}={}^1\!/_{16}, \quad u_{(+)}^{(f_2^B)}={}^1\!/_4, u_{(-)}^{(f_2^B)}=0. \tag{65}$$

Now we need to define the regression matrix that determines how Utilities, Similarity and Aesthetics are combined into Use Intentions in (39). For once, we let Alice ignore Bob's Aesthetics when determining her Use Intentions for the features, and define[28]

TABLE 5
Utilities of Actions by Features and Goals

| feature | action (tendency) | utilities | | |
|---|---|---|---|---|
| | | date | cure | general |
| young beautiful men ($f_1^A$) | refuse (neg) | $-1/2$ | $3/8$ | $-1/16$ |
| | invite (pos) | $3/8$ | $-3/8$ | 0 |
| | skip (avoid) | 0 | 0 | 0 |
| mole ($f_2^A$) | treat (pos) | $-1/4$ | $3/4$ | $1/4$ |
| | reject (avoid) | $1/4$ | $-3/16$ | $1/32$ |

$$\mathbf{B}_{ui}=\begin{array}{l} \\ \bar{u}_{(+)} \\ \bar{u}_{(-)} \\ sim\cdot\bar{u}_{(+)} \\ dis\cdot\bar{u}_{(+)} \\ sim\cdot\bar{u}_{(-)} \\ dis\cdot\bar{u}_{(-)} \\ beautiful\cdot\bar{u}_{(+)} \\ ugly\cdot\bar{u}_{(+)} \\ beautiful\cdot\bar{u}_{(-)} \\ ugly\cdot\bar{u}_{(-)} \end{array}\begin{array}{c} \begin{array}{cc} ui_{(+)} & ui_{(-)} \end{array} \\ \begin{pmatrix} 1/3 & 0 \\ 0 & 1/3 \\ 1/3 & 0 \\ 0 & 1/3 \\ 0 & 1/3 \\ 1/3 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \end{array}. \tag{66}$$

Using the results from (65) and the matrix $\mathbf{B}_{ui}$, we can calculate the indicative and counter-indicative Use Intentions for the fused feature *young beautiful men* as

$$\begin{aligned}
\vec{ui}^{(f_1)}&\approx(0.0,{}^1\!/_{16},0.0,0.0,{}^1\!/_{24},{}^1\!/_{48},\ldots)\cdot\mathbf{B}_{ui}\\
&=({}^1\!/_{144},{}^5\!/_{144}),
\end{aligned} \tag{67}$$

as well as for feature *mole* as

$$\begin{aligned}
\vec{ui}^{(f_2)}&\approx({}^1\!/_4,0.0,{}^1\!/_6,{}^1\!/_{12},0.0,0.0,\ldots)\cdot\mathbf{B}_{ui}\\
&=({}^5\!/_{36},{}^1\!/_{36}).
\end{aligned} \tag{68}$$

These values reflect that Alice is focused on work through the feature weight $w^{(f_1)}$ and that dating Bob works against her preferred goal to treat his *mole*.

### 11.7 Alice Trades Distance for Involvement

To determine Involvement and Distance, we need to define the regression matrix $\mathbf{B}_{id}$ that is used in (40) to calculate these values from the results of the previous sections. In this example we let indicative values feed into Involvement and counter-indicative values into Distance. To simplify the example of Alice and Bob, let us set all non-linear terms to 0. Leaving out some of the rows, then, $\mathbf{B}_{id}$ looks like this:[29]

28. We labeled rows with the names of the corresponding input variables and columns with the names of the related output variables for a better overview.

29. See footnote 28.

$$\mathbf{B}_{id} = \begin{array}{c} \\ beautiful \\ ugly \\ realistic \\ unrealistic \\ aid \\ obstacle \\ rel \\ irr \\ val_{(+)} \\ val_{(-)} \\ rel \cdot sim \\ irr \cdot sim \\ \vdots \end{array} \begin{array}{c} \begin{array}{cc} inv & dist \end{array} \\ \begin{pmatrix} 1/5 & 0 \\ 0 & 1/5 \\ 1/5 & 0 \\ 0 & 1/5 \\ 1/5 & 0 \\ 0 & 1/5 \\ 1/5 & 0 \\ 0 & 1/5 \\ 1/5 & 0 \\ 0 & 0 \\ 0 & 0 \\ \vdots & \vdots \end{pmatrix} \end{array}. \tag{69}$$

With regard to *young beautiful men*, this gives Alice

$$\begin{aligned} \vec{id}^{(f_1)} &= (1/2, 0.0, 2/5, 1/10, 2/5, 1/5, 1/2, 1/2, 2/5, 2/5, \ldots) \cdot \mathbf{B}_{id} \\ &= (44/100, 24/100), \end{aligned} \tag{70}$$

and for *mole*, it gives

$$\begin{aligned} \vec{id}^{(f_2)} &= (1/5, 4/5, 1.0, 0.0, 4/5, 1/5, 3/4, 3/4, 3/4, 1/2, \ldots) \cdot \mathbf{B}_{id} \\ &= (70/100, 45/100). \end{aligned} \tag{71}$$

In choosing $\beta_{idt} = 0.5$, the Involvement-Distance Trade-offs that play out in Alice's mind for the two features are

$$idt^{(f_1)} \approx 0.39 \qquad \text{and} \qquad idt^{(f_2)} \approx 0.64. \tag{72}$$

### 11.8 Why Bob Makes Alice Satisfied

To calculate the Satisfaction value, Alice takes the average Use Intentions for Bob's features. Choosing again equal weights everywhere, the Satisfaction values for Alice's features are approximately

$$s^{(f_1)} \approx 0.21 \quad \text{and} \quad s^{(f_2)} \approx 0.36. \tag{73}$$

Therefore, the focus of Alice becomes feature *mole*, which has the following Satisfaction values of the related individual actions, indexed by their action tendencies as defined in (43):

$$s^{(mole)}_{positive} \approx 0.25 \cdot 0.7 + 0.25 \cdot 0.55 + 0.5 \cdot 0.08 = 0.3525, \tag{74}$$

$$s^{(mole)}_{avoid} \approx 0.25 \cdot 0.3 + 0.25 \cdot 0.45 + 0.5 \cdot 0.08 = 0.2275, \tag{75}$$

where we set all the weights $\beta_{s,u,*}$ related to Utilities to 0.5 and all weights $\beta_{s,i,*}$ and $\beta_{s,d,*}$ related to Involvement and Distance equally to 0.25. Because Alice's Involvement outweighs Distance and Utilities, she chooses to *positively approach* Bob, which is to *treat* his *mole* and *cure* him. That gives her most Satisfaction, although dating him promised her to live in a fairy tale. Given Alice's focus on her work and the higher ambition to *cure*, this is quite the expected result. In Section 12, we will show that the contribution of affect to the decision process can lead to the choice of an action that by rational beliefs would not be the most optimal to facilitate the agent's goals.

## 12 Simulations

To further examine the capabilities of our system, we performed a couple of experiments with different configuration settings in a simplified setup.

In our simulations, Alice is a social robot programmed as a physical-exercise coach for adults. Bob is an adult who should exercise to correct his curved spine. To help Bob with his task, Alice plays an exergame, i.e., a game that is also a form of exercise, with the purpose to stimulate him to do his gymnastics. We consider this situation to be observed as a single feature only, representing the game and 'opponent' (Bob) all in one. The possible goal states in this setup are either *success* or *failure* of the game. At each step in the affective process, Alice can select from, here, two actions: A move that makes Bob perform the exercise, which she considers an act of *negative-approach* to Bob, who is reluctant, and which facilitates *success* of the game, or, alternatively, a move that will not motivate Bob to perform the exercise, which she considers an act of *positive-approach*, as Bob will be pleased to get away without having to exert himself, but which will lead towards *failure* of the game.

In selecting different values for parameters and input values of the model (the agent's 'character'), we designed four variants of Silicon Coppélia: The first variant, 'purely rational decisions', favors goal-directed factors over affect related ones and is closest to a 'rational' agent that operates by logical reasoning. The second variant, 'purely emotional decisions', does the opposite, focusing on affect related influences and ignoring goal-driven ones, with the effect that for some parameter ranges Alice chooses the move that leads to undesired *failure* because she is too involved with Bob. The third version, 'balanced decisions', represents a real world scenario that balances between the two previous scenarios, making a decision that is influenced by both goal-driven and affect related elements. The fourth variant, 'fuzzy encoding', resembles the purely emotional variant from the second scenario, but encounters ambiguity in the Aesthetics value of the feature encoding, perceiving Bob as partly beautiful and partly ugly at the same time, to examine the effect of fuzziness in the model.

Except for Aesthetics, we kept the encoding of features constant for all domains: $aff_{(+)} = 0.6$, $ep_{(+)} = 0.6$, and 0.0 for all other values.

### 12.1 Simulation 1: Purely Rational Decisions

In the first scenario, all affective influences on the decisions were set to zero, i.e., only the Use Intentions and Utility contributed to the final decision. As expected, as long as *success* was the desired goal, the *negative-approach* move towards the Bob was chosen and in the situation when, for demonstration purposes, *failure* was desired, the decision was made in favor of the *positive-approach* move. Fig. 5 shows the obtained Satisfaction values when varying the goal ambitions from $-1.0$ to 1.0 for *success* and in parallel from 1.0 to $-1.0$ for *failure*. The action chosen is the one with the higher Satisfaction value and the decision changes when the lines of $s_{neg}$ and $s_{pos}$ cross. The encoding value in the Aesthetics

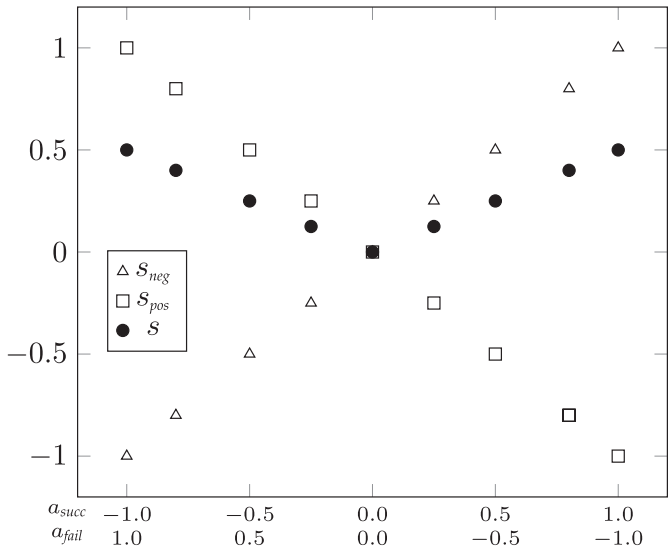


Fig. 5. Satisfaction values for goal ambitions.

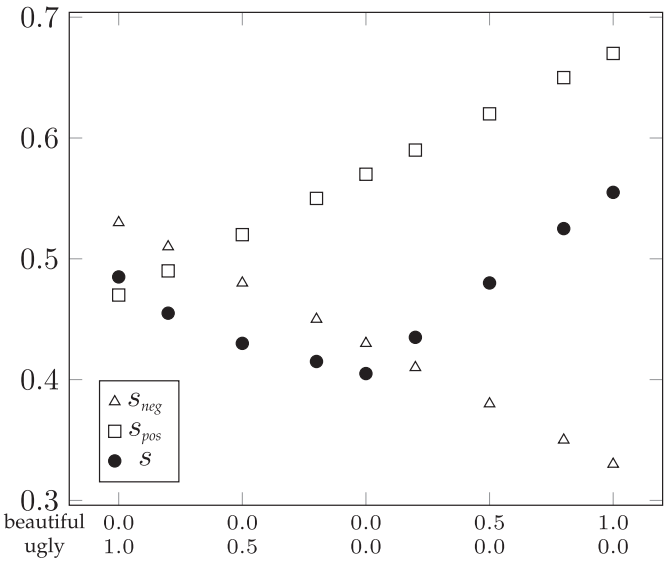


Fig. 6. Satisfaction values for aesthetics.

domain was kept constant at absolutely beautiful (1.0) in this scenario.

### 12.2 Simulation 2: Purely Emotional Decisions

In the second scenario, the same goal states were present as in the previous one, but goal ambitions were set to medium ambitions for *success* (+0.5). Now merely affective influences contributed to the decision making, i.e., Use Intentions were weighted with zero and only the Involvement-Distance Trade-off contributed to the final decision. On the input side, we modified the encoding values of the Aesthetics domain as one of the driving factors of decision making. As shown in Fig. 6, as long as Bob was assessed as attractive enough (i.e., with a straight spine), Alice in this experiment chose the *positive-approach* move. This changed once Bob exercised with a hunched back.[30] Noteworthy about this decision-making process is that it was not fed by the desire for *success* or *failure* but rather the consequence of the action tendencies associated with the actions, leading to the result. To please Bob was not in any way part of the goals and ambitions of Alice, which were only *success* in the game. This also means that in this scenario, the final decision countered the rational goals of the agent. In a sense, we saw Silicon Coppélia execute not so much goal-directed but rather bias-driven behaviors. Translated into everyday talk, Alice thought: "I don't mind *failing* because Bob's back sure looks good!"

### 12.3 Simulation 3: Balanced Decisions

In the third setting, both the affective as well as the rational contributions were taken into account. The setup was identical to the previous scenario, except that in the calculation of the Satisfaction value all weights were set non-zero. Additionally, the Aesthetics value was fed also to the Use Intentions. As shown in Fig. 7, with our parameter choices for Alice's character, Alice always chose the *negative-approach* move. However, with other choices for the model's parameters or 'character settings', either the affective or the rational part may dominate the final result.

30. The bias towards the *positive* action tendency comes from the fact that in addition to the Aesthetics value also the Epistemics and Affordances domain contributed with only indicative input values.

### 12.4 Simulation 4: Fuzzy Encoding

In the last simulation, we explored the effect of fuzzy values in the encoding by changing the Aesthetics values from absolutely beautiful to zero and then to increasingly beautiful and ugly at the same time with settings otherwise identical to the purely emotional scenario in Section 12.2.

We can see from Fig. 8 that when Bob's back is perceived as equally curved (ugly) as straight (beautiful) the same decision is obtained as with the Aesthetics values encoded to zero. However, the fuzziness of encoded factors influences the overall Satisfaction value $S$ of the feature, increasing its dominance in settings where multiple features are present. We would like to note, however, that there are many more parameters than presented in this example that may influence the effect of fuzziness in the model.

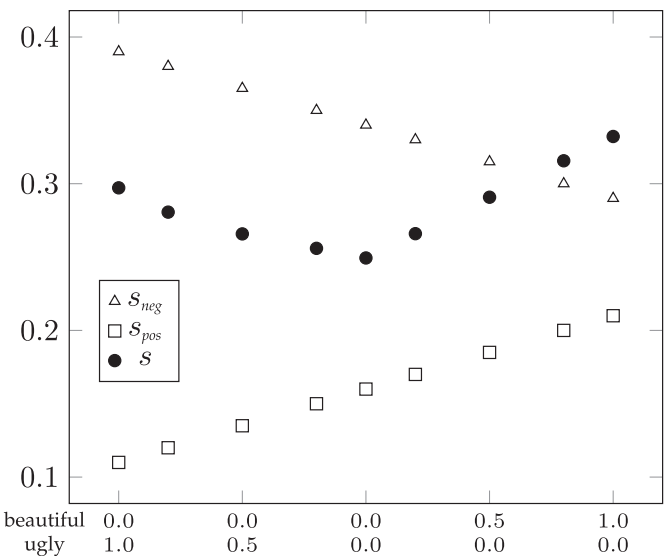


Fig. 7. Satisfaction values for aesthetics.

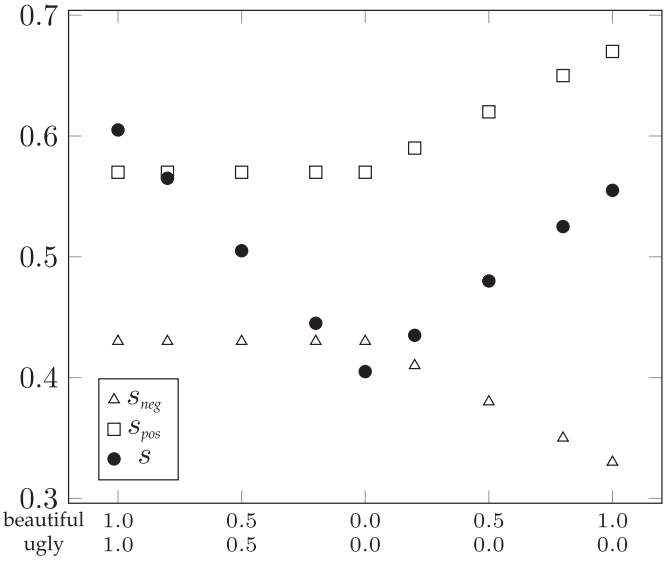


Fig. 8. Satisfaction values for fuzzy aesthetics.

The simulations only touched upon a small portion of the configuration and setup possibilities of the system. Additionally, we are not concerned here how to find the right values for all the parameters of the system in a practical application. We merely wanted to demonstrate the capabilities of the Silicon Coppélia system in simulating affective behaviors.

## 13 Connecting Silicon Coppélia With Computer Vision

In the following demonstration, we return to the setup of Section 11. To give Silicon Coppélia eyesight, we connected face recognition, a gender-and-age classifier, and a face-attractiveness ranker to the Silicon Coppélia software. For age and gender recognition, we employed Pressel's dlib's built-in face recognition tool [41], which detects human faces in an image and, based on facial landmarks such as eyes, nose, lips, and chin, estimates gender and age. Classifications ran on a pre-trained model. To obtain a score for facial attraction, we trained the face-attractiveness ranker, a machine-learning model written in Python by Leung [42], using as benchmark the SCUT-FBP set of 5500 images (jpeg, png) [43] of Asians and Caucasians labeled for attractiveness from 1 to 5. SCUT-FBP5500 has an inbuilt face-recognition system to detect the facial landmarks of an input face. Obviously, how attractive a face is and why is disputable but for the sake of argument, we assumed that this robot's aesthetic judgment exactly followed the benchmark.

For technical reasons, we modified the setup in the example in Section 11 slightly: We redefined Coppélia's goals and switched gender, now making *young beautiful women* the preferred dating partners of Coppélia:

1) *date*: Date *young beautiful women*,
2) *cure*: Cure *women* with a medical condition.

In this setting, we also did not have a recognition system for medical conditions at hand, instead we assumed that we can correlate the medical condition to the attractiveness of the other (remember, Coppélia impersonates a plastic surgeon). For simplicity we chose *skip* if the other agent was detected to be *male*, as in this case none of the available actions either facilitate or inhibit any of the robot's goals. Thus we restricted ourselves to model the agent only by *age* and *attractiveness* and defined the following two features:

$$f_1\text{: } agerange,$$
$$f_2\text{: } attractiveness(SCUT\text{-}FBP).$$

TABLE 6
Encoding of Age and Attractiveness

| age | $aest_{(+)}$ | $aest_{(-)}$ | $aff_{(+)}$ | $aff_{(-)}$ | SCUT | $aest_{(+)}$ | $aest_{(-)}$ | $aff_{(+)}$ | $aff_{(-)}$ |
|---|---|---|---|---|---|---|---|---|---|
| 0-2 | 0.4 | 0.0 | 0.4 | 0.0 | 1.0 | 0.0 | 1.0 | 1.0 | 1.0 |
| 4-6 | 0.5 | 0.0 | 0.5 | 0.0 | 1.5 | 0.0 | 0.7 | 0.7 | 0.7 |
| 8-12 | 0.7 | 0.0 | 0.7 | 0.0 | 2.0 | 0.0 | 0.6 | 0.6 | 0.6 |
| 15-20 | 1.0 | 0.0 | 1.0 | 0.0 | 2.5 | 0.0 | 0.2 | 0.3 | 0.2 |
| 25-32 | 0.8 | 0.0 | 0.8 | 0.0 | 3.0 | 0.2 | 0.0 | 0.2 | 0.0 |
| 38-43 | 0.1 | 0.0 | 0.1 | 0.0 | 3.5 | 0.5 | 0.0 | 0.5 | 0.0 |
| 48-53 | 0.0 | 0.5 | 0.0 | 0.5 | 4.0 | 0.7 | 0.0 | 0.7 | 0.0 |
| 60-100 | 0.0 | 1.0 | 0.0 | 1.0 | 4.5 | 0.8 | 0.0 | 0.8 | 0.0 |
| | | | | | 5.0 | 1.0 | 0.0 | 1.0 | 0.0 |

We used slightly changed non-overlapping age ranges as listed in Table 6, and represented attractiveness simply by a number between 1 and 5, rounding the SCUT-FBP score to the nearest half. Again for technical reasons, we did not evaluate Similarity from the features but set it to a constant value of $sim = 0.55$.[31]

We implemented the observation process by encoding features for the Aesthetics and Affordance domain only and deduced the appraisal weights from the outputs of the recognition systems as listed in Table 6.

There, the Aesthetic weights are a straightforward mapping from the age and the SCUT-FBP score, presuming that Coppélia's goal to date *beautiful women* in the *age range (15-20)* reflects her Aesthetic judgment. Affordance values reflect that good dating partners are an *aid* to Coppélia in terms of finding a date. Low attractiveness is an *obstacle* for this goal, but on the other hand, low attractiveness makes Coppélia believe that the other agent has a medical condition that Coppélia can treat with her professional skills, which makes it an *aid*. This resulted in an ambiguous view on the Affordance of the low attractiveness of the other.

The actions that were available to Coppélia are the same as listed in (59), but in this case, we related the actions for *date* to both features and the actions for *cure* to the attractiveness score, as we deduced the presence of a condition from this feature

$$\begin{aligned} A^{(f_1)} &= \{refuse, invite, skip\}, \\ A^{(f_2)} &= \{refuse, invite, skip, treat, reject\}. \end{aligned} \tag{76}$$

In addition to the feature encoding, we made Coppélia believe that the actions will affect Coppélia's goals by matching the goals against the detected age range and attractiveness score. With the same goals as in (57), Table 7 shows the chosen values for the feature set detected in our sample images in Fig. 9.

31. In this representation, we treated the features as plain labels, which would lead to positive similarity only for exact matches. However, this is only for technical reasons and the feature representation could be easily changed to a setup similar to Section 11.2.

TABLE 7
Beliefs $ag_{ij}$ for Observed Agencies

| action | {(25-32), 4} | | {(25-32), 3} | | {(15-20), 4} | | {(15-20), 1.5} | |
|---|---|---|---|---|---|---|---|---|
| | date | cure | date | cure | date | cure | date | cure |
| invite | 0.9 | 0.0 | 0.2 | 0.0 | 0.8 | 0.0 | -0.6 | 0.0 |
| refuse | -0.3 | 0.0 | 0.0 | 0.0 | -0.2 | 0.0 | 0.0 | 0.0 |
| skip | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| treat | 0.0 | -0.6 | 0.0 | -0.1 | 0.0 | -0.4 | 0.0 | 0.7 |
| reject | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | -0.3 |

Next we fed images of males and females, young and old, into the system, one at a time. To do so, we used the face database 'Labeled Faces in the Wild' by Huang *et al.* [44]. Fig. 9 shows the output for four example images. Note that in Fig. 9, the age of Character 4 was misjudged by our detection system: The woman in the picture was between 30 and 40 years of age, but for Coppélia she was an adolescent.

In the remainder, we offer two small experiments of Coppélia building up affect with different levels of ambition on the basis of age and attractiveness of people in a picture. We manipulated the level of ambition of Coppélia wanting to achieve both goals: In Experiment 1, wanting a date was stronger than to help; reversely so in Experiment 2. The research question in both experiments was who Coppélia would want to date and who was in need of help.

### 13.1 Looking for a Date

In Experiment 1, the ambition level to achieve the goals was set to

$$a_{date} = 0.75 \qquad \text{and} \qquad a_{cure} = 0.5. \tag{77}$$

Table 8 provides the results for all dimensions Coppélia checked.

Table 8 shows how the encoded Aesthetics and Affordance weights of Age and SCUT-FBP score shaped the development of Engagement (Involvement, Distance) and Satisfaction with the characters in view of the prioritized goal to *date* one of them. For dating purposes, the lonesome plastic surgeon Coppélia estimated that Character 3 and 4 were of the preferred age group (15-20) rather than Character 3 and 4 (25-32) (Fig. 9). Yet, Character 1 and 3 had a higher SCUT-FBP score ($> 3.9$) with number 1 in Coppélia's eyes being the prettiest of them all (4.12) (Fig. 9). Although (mistakenly) thought to be of the right age, Coppélia deemed Character 4 to be least pretty (SCUT-FBP = 1.51), causing a drop in Involvement and a rise in Distance. Most interestingly, however, Table 8 shows that Coppélia was not least Satisfied with Character 4 (Satisfaction $\approx$ (0.24, 0.28)) (Table 8). That 'honor' was bestowed upon Character 2 (Satisfaction $\approx$ (0.15, 0.14)), although Coppélia found 2 way prettier than 4. Coppélia was least satisfied with Character 2 because 2 was of the less preferred age group and within that group less pretty than 1. Additionally, Character 2 was not in need of help either, therefore immaterial to the lower-priority goal Coppélia had (i.e., to cure people with facial deficiencies). Character 4, however, did have some meaning to the lower-priority goal so that in spite of raising much Distance and little Involvement as a dating partner, Character 4 did exert higher Use Intentions than Character 2 because Coppélia guessed number 4 to a certain extent should be treated in her clinic.

The dating competition, then, was between Character 1 (the prettiest but of less preferred age) and Character 3 (also pretty but in the proper age range). The level of Satisfaction as calculated from Use Intentions and Involvement-Distance trade-off made Character 3 (Satisfaction $\approx$ (0.37, 0.31)) outdo Character 1 (Satisfaction $\approx$ (0.35, 0.32)): Coppélia gave in a little on beauty to date the younger candidate.

### 13.2 Trying to Help

For Experiment 2, the ambition levels were set to

$$a_{date} = 0.5 \qquad \text{and} \qquad a_{cure} = 0.75. \tag{78}$$

Table 9 offers the results of Coppélia's affective assessment based on these priorities.

To provide help, the lonesome plastic surgeon Coppélia judged, in comparison to Experiment 1 in Section 13.1, that her Involvement with Character 3 was still the highest (Involvement = (0.6, 0.48)) just as she was still most Distant (Distance = (0.3, 0.48)) towards Character 4 (Table 9). However, in the trade-off between Involvement and Distance, the gap between Character 4 (IDT $\approx$ (0.45, 0.47)) and the prettiest pair Character 1 (IDT $\approx$ (0.46, 0.43)) and 3 (IDT $\approx$ (0.51, 0.41)) decreased due to the lower ambition to date them.

Also in terms of Satisfaction, Character 4 (Satisfaction $\approx$ (0.23, 0.30)) could almost maintain the same level as in Experiment 1, while Character 3 (Satisfaction $\approx$ (0.31, 0.26)), who was preferred dating-wise (a lower-priority goal in Experiment 2), and Character 4 (Satisfaction $\approx$ (0.31, 0.26)) suffered the loss of Satisfaction. That Character 4 yet caught

(a) Character 1:
Gender: female
Age range: (25-32)
SCUT-FBP: 4.12

(b) Character 2:
Gender: female
Age range: (25-32)
SCUT-FBP: 2.79

(c) Character 3:
Gender: female
Age range: (15-20)
SCUT-FBP: 3.92

(d) Character 4:
Gender: female
Age range: (15-20)
SCUT-FBP: 1.51

Fig. 9. Sample pictures from labeled faces in the wild [44].

TABLE 8
Results for Simulation With Computer Vision and Goal Ambitions Favoring *date*

| | Character 1 | | Character 2 | | Character 3 | | Character 4 | |
|---|---|---|---|---|---|---|---|---|
| | Age range | SCUT-FBP | Age range | SCUT-FBP | Age range | SCUT-FBP | Age range | SCUT-FBP |
| $aest_{(+)}$ | 0.8 | 0.7 | 0.8 | 0.2 | 1.0 | 0.7 | 1.0 | 0.0 |
| $aest_{(-)}$ | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.7 |
| $aff_{(+)}$ | 0.8 | 0.7 | 0.2 | 0.2 | 1.0 | 0.7 | 1.0 | 0.7 |
| $aff_{(-)}$ | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.7 |
| Relevance | 0.75 | 0.75 | 0.2 | 0.2 | 0.75 | 0.75 | 0.6 | 0.6 |
| Irrelevance | 1.0 | 0.5 | 1.0 | 0.9 | 1.0 | 0.6 | 1.0 | 0.5 |
| $Valence_{(+)}$ | 0.75 | 0.7 | 0.2 | 0.2 | 0.75 | 0.7 | 0.0 | 0.5 |
| $Valence_{(-)}$ | 0.3 | 0.5 | 0.0 | 0.1 | 0.2 | 0.4 | 0.6 | 0.6 |
| Use Int. useful | 0.3375 | 0.3375 | 0.075 | 0.075 | 0.3 | 0.3 | 0.0 | 0.175 |
| Use Int. no use | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Involvement | 0.62 | 0.57 | 0.28 | 0.26 | 0.7 | 0.57 | 0.52 | 0.36 |
| Distance | 0.26 | 0.2 | 0.2 | 0.2 | 0.24 | 0.2 | 0.32 | 0.5 |
| IDT Trade-off | 0.53 | 0.4775 | 0.26 | 0.245 | 0.585 | 0.4775 | 0.47 | 0.465 |
| Satisfaction | 0.349375 | 0.323125 | 0.14875 | 0.14125 | 0.3675 | 0.31375 | 0.235 | 0.27625 |
| Action | invite | | invite | | invite | | treat | |

TABLE 9
Results for Simulation With Computer Vision and Goal Ambitions Favoring *Cure*

| | Character 1 | | Character 2 | | Character 3 | | Character 4 | |
|---|---|---|---|---|---|---|---|---|
| | Age range | SCUT-FBP | Age range | SCUT-FBP | Age range | SCUT-FBP | Age range | SCUT-FBP |
| $aest_{(+)}$ | 0.8 | 0.7 | 0.8 | 0.2 | 1.0 | 0.7 | 1.0 | 0.0 |
| $aest_{(-)}$ | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.7 |
| $aff_{(+)}$ | 0.8 | 0.7 | 0.2 | 0.2 | 1.0 | 0.7 | 1.0 | 0.7 |
| $aff_{(-)}$ | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.7 |
| Relevance | 0.5 | 0.6 | 0.2 | 0.2 | 0.5 | 0.5 | 0.5 | 0.7 |
| Irrelevance | 1.0 | 0.5 | 1.0 | 0.9 | 1.0 | 0.6 | 1.0 | 0.5 |
| $Valence_{(+)}$ | 0.5 | 0.5 | 0.2 | 0.2 | 0.5 | 0.5 | 0.0 | 0.7 |
| $Valence_{(-)}$ | 0.3 | 0.6 | 0.0 | 0.1 | 0.2 | 0.4 | 0.5 | 0.5 |
| Use Int. useful | 0.225 | 0.225 | 0.05 | 0.05 | 0.2 | 0.2 | 0.0 | 0.2625 |
| Use Int. no use | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Involvement | 0.52 | 0.5 | 0.28 | 0.26 | 0.6 | 0.48 | 0.5 | 0.42 |
| Distance | 0.26 | 0.22 | 0.2 | 0.2 | 0.24 | 0.2 | 0.3 | 0.48 |
| IDT Trade-off | 0.455 | 0.43 | 0.26 | 0.245 | 0.51 | 0.41 | 0.45 | 0.465 |
| Satisfaction | 0.28375 | 0.27125 | 0.1425 | 0.135 | 0.305 | 0.255 | 0.225 | 0.298125 |
| Action | invite | | invite | | invite | | treat | |

up was also because of the Use Intentions (0.2625) Coppélia had for her, which in view of her goal to cure were highest for Character 4. Compared to the results of Section 13.1, the two prettiest in the group were still *en vogue* due to the lower-priority goal of dating, but the runner up was number 4 with a strong indication to be helped. Although Coppélia still felt little friendship for her, from a functional or instrumental viewpoint, Character 4 was worth helping while Coppélia kept her (professional) distance.

## 14 Conclusion

The current paper delivers a full-fledged formal model of an agent/robot system that builds up affect with its user (or other agency). It is the implementation of a psychological model that has been tested for years with real users. Moreover, the system is completely implemented and made accessible for the community to evaluate (open source). A number of examples and simulation experiments showed the internal consistency of our work. The next step would be to confront our system with actual users and ask them how human-like they think Silicon Coppélia responds.

In relating back to Fig. 1 where Silicon Coppélia was placed in the field of affective computing, our system is not focused on emotions per se like, for instance, WASABI is (Affect Simulation Architecture for Believable Interactivity) [28]. Coppélia is not focused on coping strategies either like, for example, EMA is [45]. However, EMA could be said to be at the affective heart of Coppélia because she also does appraisals based on beliefs about the world that lead to actions that are believed to have utility to change a situation. Yet, Coppélia does more than that and for different purposes. First, Coppélia is focused on friendship formation; coping and emotions are not her main concern. It is the affective process that is leading in which emotions are a 'side effect.' It could well be that Coppélia builds up antipathy for her user and turns her back on the interaction. This is not coping with emotions but relationship building for better or for worse. Second, the dimensions Coppélia takes into account are restricted to those important for evaluation

of (virtual) characters (not personalities, like WASABI) as based on empirical research [4]: Other models do not encode Ethics, Aesthetics, Affordances, and Epistemics for their assessments nor produce values for Involvement, Distance, and Use Intentions and their trade-offs to select a response. Third, Coppélia's software architecture is such that algorithms and modules (e.g., moral reasoning) can be replaced and (simulation) experiments can be run to test different schools of thought. For a more complete overview of affective models, consult Kowalczuk and Czubenko (2016) [46].

In relating back to our simulation in Section 13, EMA would 'perceive' the photos of women but not for any specific dimensions unless defined in terms of Coppélia: Aesthetics and Affordances. WASABI would search for certain 'pulses' in the photos that may generate emotions, which Coppélia is not concerned about in the plastic-surgeon simulation. As a response, EMA would want to deal with the generated emotions so to serve regulation (e.g., of suicidal tendencies) [45]. In the plastic-surgeon simulation, Coppélia is not focused on regulating her emotions but on servicing her goals: Making friends and performing her job. It could be that Coppélia utterly hates someone and leaves it that way. Different from WASABI, Coppélia would not calculate an 'emotion vector,' generating secondary emotions. In the simulation of Section 13, WASABI and more so EMA would be capable of appraisal of the environment in view of a goal and calculate an action – Coppélia is way more particular in what dimensions specifically count for her purposes of continued interaction and building up rapport. In sum, WASABI would come up with a general emotional response to the photos, EMA would come up with a strategy to regulate evoked emotions but neither would build up friendship and help people with a medical condition.

Applied to 'Alice' and 'Bob', Alice made the 'wrong' move despite her desire to win, because Bob looked good with his straight back. Although the lonely surgeon was attracted to others, she helped the person in need while keeping her distance. This is the type of decision that makes an artificial agent more human-like: Coppélia performs an elaborate trade-off between the pros and cons of certain actions originating from the appraisal of another agency's features in view of the agent's own goals and concerns and 'deflected' by its biases (cf. [9]). One can imagine that if the artificial agent shares the goals of the user (e.g., Alice wants to care for Bob), it may serve as a personal assistant or coach with which the user may experiment and try out alternative behaviors (e.g., see [7], Section 5).

The Silicon Coppélia system is feature-based so that decisions may be fed by tiny details but also by judgments over larger, more general sets. The sets are basically fuzzy because pros and cons may be attached to an individual feature or features may participate in more appraisal domains (e.g., 'inner beauty' relates to Ethics as well as Aesthetics).

Apart from simply counting frequency numbers, we also modeled the (biased) perceptions of those numbers, using weights. What remains an external task is not only where the features come from (e.g., computer vision or Deep Learning), but also on what grounds they are weighted (e.g., frequency, amplitude or salience). To retrieve features, one could think of face analysis, speech, or physiology (e.g., see Section 13 or [7], Section 2).

With respect to the origin of perceived weights, Silicon Coppélia attempts to explain how perceptions determine the expected satisfaction of an agent for each of the actions it can choose from. Coppélia assumes a belief system that assigns values that indicate how much a feature, when observed, contributes to each appraisal dimension $d_i^{(k)}$. Additionally, a tracking system should monitor the changes in the (social) environment and the actions of the agent in that environment. The tracking system determines how relevant features are given the current situation. This is encoded in the bias weights $w^{(k)}$. For now, we assume that the results of the belief system and the tracking system can be linearly combined to obtain the perceived contributions of the features in all appraisal dimensions $p_i^{(k)}$. Empirical research is needed to build a probably more realistic model of how the perceived values, which form the inputs for Silicon Coppélia, should be derived from raw inputs.

Regarding the aggregation from features to concept, we attempted to quantify the relative merit of action tendencies while at several places applying 'parsimony' or 'simplicity' as our criterion for choosing mathematics to the psychology. We combined the perceived appraisals of features in fuzzy sets to escape the constraints of binary logic, while avoiding commitment to a fully probabilistic interpretation. For the combination of fuzzy sets, we used simple min/max rules, because of as yet there is no empirical evidence for continuity of derivatives or any other reason to consider more sophisticated operators. That of course may change in the future as (neuro)psychology is advancing. Additionally, we assume that the rules apply independently for each feature and for each appraisal dimension, which need not be the case but we have no grounds to decide otherwise. Further research should indicate where and how the model needs to be adapted for a better fit with actual human behavior.

Silicon Coppélia showed us how many algorithmic choices had to be made for simplicity's sake, because we did not have the empirical backing to make other choices. There appear to be many parameters that need to be set to generate a relatively simple interaction, because humans vary in the way they perform an interaction and in the reasons why they perform it (cf. [24]). In other words, currently Coppélia forces the researcher to explicate the theory that drives the interaction; put differently, the theory under study provides the basic guidelines for parameter settings. How weights are established and how high they should be, should be subjected to psychological experimentation; in that sense, building Silicon Coppélia uncovered many research gaps. An alternative direction of research could be to learn the parameters of the model from empirical training data using machine learning techniques.

If we want robots and artificial agents to act 'emotionally,' they should have goals and concerns of their own. Only if those goals are supported or hampered will there be reason for the agent to cheer at its user or to protest. Thus far, the goals of the robot or agent are inserted by the user; we do not have learning systems in place that may make the machine change its goals (acquire new, forget the old) but that is not infeasible in the long run. For example, an option could be to build a semantic perception model of an

environment after [47] and let the robot explore locations that have high semantic information content. The learning algorithm makes the robot take on tasks that are not pre-programmed such as inspection of a location and following the user. From the work of [48], developmental patterns may emerge where the robot discovers the affordances of objects and people and explores vocal interaction with its user.

Another thing, apart from the question where the input comes from, is that once the agent has made an affective decision, we do not have a large library of appropriate expressions (e.g., language, gestures) that we can use to communicate the 'affective' state of the machine to the human user (cf. [7], Section 3). That would be a great job for natural-language engineers and interaction designers. One strand of affect-expression design could be body language, providing a mapping of the various positions of the robot's joints and head to the expression of valence and arousal [49]. For other options of robotic expression of emotion, see [24].

This paper described the affective decision process used in the core of Sillicon Coppélia, a software framework for virtual agents and robots. The structure of the decision process is based on the Interactively Perceiving and Experiencing Fictional Characters (I-PEFiC) framework. The calculations use fuzzy logic to combine the sets of internal and perceived variables, resulting in an expected satisfaction value for each possible action, which is used to suggest which action to take. A synthetic example of a one-to-one as well as a simulation with computer vision was used to demonstrate the decision process depending on the state of each agent.

The current architecture of Silicon Coppélia is that of a feedforward system, controlled mathematically. That means that each path Coppélia travels can be checked, which improves our understanding of the basic system. The only 'feedback circuit' available at the moment is that an output action changes the situation Coppélia is in and so changes the input features that Coppélia processes in the next round. Adaptation of behavior of the system, then, is after completing a full loop.

To route back certain intermediate values as input to itself may be a next step in Coppélia's development (cf. WASABI [28] or FLAME [15]). Central would be the feedback to Relevance and Valence with expected utility of actions as its main concern. That would take local and global self-monitoring modules, analysis of intermediate states in view of Coppélia's goals, and self-adjustment procedures to 'maintain stability,' a kind of self-management. Although a necessary step perhaps, the complexity of the architecture would multiply and Coppélia's behaviors and affective decision making will become harder to follow.

For others to explore our system and review its strengths and weaknesses, we implemented our Silicon Coppélia system in Ptolemy II, an open-source framework for actor-based, visual software design (see [50]). Its architectural design is integrative: It is part of a larger system, the Robot Brain Server [51], which runs more types of artificial intelligence as interconnected services, deployed in the Cloud or, to enhance data privacy, as an enclosed local system. The source code of the implementation is available at [8] and we invite the community to experiment with our version, download a local copy and change whatever they feel should be changed. This way, we make robots more lovable together.

## Acknowledgments

This study was performed as part of the employment of the first author at The Hong Kong Polytechnic University in the PAL project of the Artificial Intelligence in Design Laboratory under Grant AiDLab RP2P3. The part at Vrije Universiteit Amsterdam (first and second author) was supported by Communicating with and Relating to Social Robots: Alice Meets Leolani, NWO Open Competition - Digitalisation (SSH) under Grant: 406.DI.19.005 and by (first author) VUvereniging under Grant: AB/rk/2019/100. Pui Yi Leung is kindly acknowledged for connecting computer vision to our affective processing system and running the associated simulations. The authors would like to thank Luminis Technologies B.V. for supporting this work, in particular Marcel Offermans for commenting on an earlier draft of this paper.

## References

[1] J. F. Hoorn and E. A. Konijn, "Perceiving and experiencing fictional characters: An integrative account," *Japanese Psychol. Res.*, vol. 45, no. 4, pp. 250–268, 2003.

[2] E. A. Konijn and J. F. Hoorn, "Some like it bad: Testing a model for perceiving and experiencing fictional characters," *Media Psychol.*, vol. 7, no. 2, pp. 107–144, 2005.

[3] E. A. Konijn and B. J. Bushman, "World leaders as movie characters? Perceptions of George W. Bush, Tony Blair, Osama bin Laden, and Saddam Hussein," *Media Psychol.*, vol. 9, no. 1, pp. 157–177, 2007.

[4] H. C. van Vugt, J. F. Hoorn, and E. A. Konijn, "Interactive engagement with embodied agents: An empirically validated framework," *Comput. Animation Virt. Worlds*, vol. 20, no. 2/3, pp. 195–204, 2009.

[5] J. F. Hoorn, E. A. Konijn, and M. A. Pontier, "Dating a synthetic character is like dating a man," *Int. J. Soc. Robot.*, vol. 11, no. 2, pp. 235–253, Apr. 2019. [Online]. Available: https://doi.org/10.1007/s12369-018-0496-1

[6] M. Pontier and J. F. Hoorn, "How women think robots perceive them - as if robots were men," in *Proc. 5th Int. Conf. Agents Artif. Intell.*, 2013, pp. 496–504.

[7] R. A. Calvo, S. D'Mello, J. Gratch, and A. Kappas, Eds., *The Oxford Handbook of Affective Computing*. Oxford, U.K.; New York, NY, USA: Oxford Univ. Press, 2015.

[8] Silicon Coppélia - An implementation in Ptolemy, 2016. [Online]. Available: https://bitbucket.org/robopop/silicon-coppelia

[9] M. Pudane, E. Lavendelis, and M. A. Radin, "Human emotional behavior simulation in intelligent agents: Processes and architecture," *Procedia Comput. Sci.*, vol. 104, pp. 517–524, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050917301680

[10] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.

[11] J. D. Velásquez, "Modeling emotions and other motivations in synthetic agents," in *Proc. 14th Nat. Conf. Artif. Intell. and 9th Conf. Innovative Appl. Artif. Intell.*, 1997, pp. 10–15.

[12] L.-F. Rodríguez and F. Ramos, "Development of computational models of emotions for autonomous agents: A review," *Cogn. Comput.*, vol. 6, no. 3, pp. 351–375, 2014.

[13] E. Hudlicka, "This time with feeling: Integrated model of trait and state effects on cognition and behavior," *Appl. Artif. Intell.*, vol. 16, no. 7/8, pp. 611–641, 2002.

[14] S. C. Marsella and J. Gratch, "EMA: A process model of appraisal dynamics," *J. Cogn. Syst. Res.*, vol. 10, no. 1, pp. 70–90, Mar. 2009.

[15] M. S. El-Nasr, J. Yen, and T. R. Ioerger, "Flame—Fuzzy logic adaptive model of emotions," *Auton. Agents Multi-Agent Syst.*, vol. 3, no. 3, pp. 219–257, Sep. 2000.

[16] A. Ortony, G. L. Clore, and A. Collins, *The Cognitive Structure of Emotions*. Cambridge, U.K.: Cambridge Univ. Press, 1990.

[17] N. H. Frijda, "The laws of emotion," *Amer. Psychol.*, vol. 43, no. 5, 1988, Art. no. 349.

[18] C. A. Smith and R. S. Lazarus, "Emotion and adaptation," *Handbook of Personality: Theory and Research*, Hoboken, NJ, USA: Wiley, 1990, pp. 609–637.

[19] J. J. Gross, "Emotion regulation in adulthood: Timing is everything," *Curr. Directions Psychol. Sci.*, vol. 10, no. 6, pp. 214–219, 2001.

[20] J. F. Hoorn, *Psychological Aspects of Technology Interacting With Humans*. Hoboken, NJ, USA: Wiley-Blackwell, 2015, pp. 176–201.
[21] J. Gratch and S. C. Marsella, "Appraisal models," in *The Oxford Handbook of Affective Computing*, R. A. Calvo, S. D'Mello, J. Gratch, and A. Kappas, Eds. Oxford, U.K.; New York, NY, USA: Oxford Univ. Press, 2015, pp. 54–67.
[22] A. Newell, *Unified Theories of Cognition*. Cambridge, MA, USA: Harvard Univ. Press, 1990.
[23] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, 2017.
[24] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit., 1st Int. Workshop Emotion Synthesis rePresentation Anal. Continuous space*, 2011, pp. 827–834.
[25] M. Pontier, G. Siddiqui, and J. F. Hoorn, "Speed dating with an affective virtual agent - developing a testbed for emotion models," in *Intelligent Virtual Agents*, J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, and A. Safonova, Eds. Berlin, Germany: Springer, 2010, pp. 91–103.
[26] M. Pontier and G. F. Siddiqui, "An affective agent playing tic-tac-toe as part of a healing environment," in *Proc. Int. Conf. Princ. Pract. Multi-Agent Syst.*, 2009, pp. 33–47.
[27] J. F. Hoorn, M. Pontier, and G. F. Siddiqui, "Coppélius' concoction: Similarity and complementarity among three affect-related agent models," *Cogn. Syst. Res.*, vol. 15, pp. 33–49, 2012.
[28] C. W. Becker-Asano and I. Wachsmuth, "Affective computing with primary and secondary emotions in a virtual human," *Auton. Agents Multi-Agent Syst.*, vol. 20, no. 1, pp. 32–49, 2010.
[29] P. Gebhard, "ALMA: A layered model of affect," in *Proc. 4th Int. Joint Conf. Auton. Agents Multiagent Syst.*, 2005, pp. 29–36.
[30] J. F. Hoorn, *Epistemics of the Virtual*. Amsterdam, Netherlands: John Benjamins Publishing, 2012.
[31] J. Doolaard (Producer) and S. Burger (Director), "Ik ben alice / alice cares [documentary film]," 24th International Film Festival Rotterdam, NCRV, Amsterdam, 2015. [Online]. Available: https://vimeo.com/ondemand/alicecares
[32] J. de Jager and A. Grijzenhout, In Zora's gezelschap [In Zora's company] [TV Report]. Dit is de Dag. EO: NPO 2. 2014. [Online]. Available: http://www.npo.nl/artikelen/dit-is-de-dag-geeft-bejaarden-zorgrobot
[33] A. Tversky, "Features of similarity," *Psychol. Rev.*, vol. 84, no. 4, pp. 327–352, 1977.
[34] J. F. Hoorn, E. A. Konijn, D. Germans, S. Burger, and A. Munneke, "The in-between machine - the unique value proposition of a robot or why we are modelling the wrong things," in *Proc. Int. Conf. Agents Artif. Intell.*, 2015, pp. 464–469.
[35] J. F. Hoorn, M. Pontier, and G. F. Siddiqui, "When the user is instrumental to robot goals: First try-agent uses agent," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, 2008, pp. 296–301.
[36] J. J. Buckley and E. Eslami, *An Introduction to Fuzzy Logic and Fuzzy Sets*, vol. 13. Berlin, Germany: Springer Science & Business Media, 2002.
[37] Y. A. Tolias, S. M. Panas, and L. H. Tsoukalas, "Generalized fuzzy indices for similarity matching," *Fuzzy Sets Syst.*, vol. 120, no. 2, pp. 255–270, 2001.
[38] H.-J. Zimmermann, *Fuzzy Set Theory and its Applications*, 2nd ed. Berlin, Germany: Springer, 1991.
[39] J. F. Hoorn, "A robot's experience of its user: Theory," in *Proc. 30th Annu. Conf. Cogn. Soc.*, 2008, pp. 2504–2509.
[40] T. Bosse, J. F. Hoorn, M. A. Pontier, and G. F. Siddiqui, "A robot's experience of another robot: Simulation," in *Proc. 30th Int. Annu. Conf. Cogn. Sci. Soc.*, 2008, pp. 2498–2503.
[41] D. Pressel, "Rude carnie: Age and gender deep learning with TensorFlow," 2002. [Online]. Available: https://github.com/dpressel/rude-carnie
[42] J. Leung, "Just a simple face attractiveness ranker," 2018. [Online]. Available: https://github.com/joshualeung/mini-face-rank
[43] L. Liang, L. Lin, L. Jin, D. Xie, and M. Li, "SCUT-FBP5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction," in *Proc. 24th Int. Conf. Pattern Recognit.*, 2018, pp. 1598–1603.
[44] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Amherst, MA, Tech. Rep. 07–49, Oct. 2007. [Online]. Available: http://vis-www.cs.umass.edu/lfw
[45] I. Gratch *et al.*, "Detecting suicidal thoughts: The power of ecological momentary assessment," *Depression Anxiety*, vol. 38, pp. 8–16, 2021. [Online]. Available: https://doi.org/10.1002/da.23043
[46] Z. Kowalczuk and M. Czubenko, "Computational approaches to modeling artificial emotion — An overview of the proposed solutions," *Front. Robot. AI*, vol. 3, 2016, Art. no. 21.
[47] Y. Girdhar and G. Dudek, "Modeling curiosity in a mobile robot for long-term autonomous exploration and monitoring," *Auton. Robots*, vol. 40, no. 7, pp. 1267–1278, Sep. 2015.
[48] P.-Y. Oudeyer and L. B. Smith, "How evolution may work through curiosity-driven developmental process," *Topics Cogn. Sci.*, vol. 8, no. 2, pp. 492–502, 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/tops.12196
[49] A. Beck, L. Cañamero, and K. A. Bard, "Towards an affect space for robots to display emotional body language," in *Proc. 19th Int. Symp. Robot Hum. Interactive Commun.*, 2010, pp. 464–469.
[50] The Ptolemy project, 2016. [Online]. Available: http://ptolemy.eecs.berkeley.edu
[51] J. F. Hoorn, "The robot brain server: Design of a human-artificial systems partnership," in *Proc. 1st Int. Conf. Intell. Hum. Syst. Integr.*, 2018, pp. 531–536.

**Johan F. Hoorn** received the PhD degree in computer science from Vrije University Amsterdam, Netherlands as well as a PhD degree in general and comparative literature from Vrije University Amsterdam, Netherlands. He is currently a full professor of social robotics with The Hong Kong Polytechnic University, Department of Computing and School of Design, Hong Kong. His current research interests include theory of emotion, creativity, reality perception, and moral reasoning, implemented in artificial agents and social robots.

**Thomas Baier** received the graduated degree in physics from the Technical University of Munich, Germany, and the PhD degree in mathematics and its applications from the Central European University, Budapest, Hungary. He is currently a postdoc with VU Amsterdam, Faculty of Social Sciences, Department of Communication Sciences, Netherlands. He is currently involved in the Alani project, a joint project with the Computational Lexicology and Terminology Lab at the VU Amsterdam, Netherlands.

**Jeroen van Maanen** received the MSc degree in mathematics from Radboud University Nijmegen, Netherlands. He is currently working toward the PhD degree in learning systems, using probability theory and Minimum Description Length Principle (MDL) at Vrije University Amsterdam, Netherlands in collaboration with The Hong Kong Polytechnic University, Hong Kong. He is a senior software architect and developer at The Future Group. His research interests include model theory, axiomatic set theory, computability theory, and constructivist mathematics.

**Jeroen Wester** is currently a designer and IT architect with Wieswies. He develops web applications for a wide range of companies, and he provides training and consulting services. His current focus is on the application of blockchain technology. Previously, he worked for a Semantic Web startup company named Aduna. He has a background in AI and Semantic Web technologies.